Development of an In Silico Platform for the Prediction of Off-Target Binding

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Manuel Sellner

2024

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Erstbetreuer: Prof. Dr. Martin Smieško

Zweitbetreuer: Prof. Dr. Daniel Ricklin

Externer Experte: Dr. Christian Kramer

Basel, den 19.12.2023

Dekan
Prof. Dr. Marcel Mayor

Thesis advisor: Professor Martin Smieško

Manuel S. Sellner

Development of an In Silico Platform for the Prediction of Off-Target Binding

Abstract

Human beings are constantly being exposed to a diverse array of chemical compounds, both intentionally and unintentionally. The rigorous assessment of chemical toxicity is therefore paramount for human health. Traditionally, evaluating small molecule-induced toxicity involves costly and time-consuming in vitro and in vivo tests. In many cases, toxic effects begin with off-target binding, an undesired interaction between a small molecule and a protein. In the pharmaceutical industry, off-binding assessment is often performed in late pre-clinical stages of drug development. However, neglecting off-target toxicity can lead to drug failure, resulting in significant financial loss and years of development time.

This thesis presents innovative computational tools designed for the early assessment of off-target liabilities. These tools offer a cost-effective and rapid alternative to traditional methods, enabling their application in the early stages of pharmaceutical development to guide the design of safe drugs. The thesis begins by examining the impact of dataset quality on deep learning-based predictions of drug-target interactions. It shows that correct data handling is critical and demonstrates how a detailed characterization of intermolecular interactions improves predictions. Leveraging this understanding, we introduce PanScreen, an online platform automating the prediction of off-target binding. PanScreen encompasses a portfolio of pharmaceutically and toxicologically relevant off-targets and provides qualitative and quantitative predictions, including binding poses and estimated affinities. To complement this structure-based approach, we developed a deep learning model that preserves molecular similarities in the form of Euclidean distances in latent space. This model enhances protein-structure-free screening of ultra-large databases, accelerating similarity-based searches by orders of magnitude. We show that it can be applied to different similarity metrics, including alignment-based 3D shape similarities.

These in silico tools hold promise for predicting off-target interactions in diverse applications, offering an inexpensive and fast option to complement traditional methods. Specifically, PanScreen represents a significant step in this direction. As the development of these tools is an ongoing process, we offer a roadmap that outlines avenues for further improvement, aiming to enhance their robustness and accuracy. Ultimately, we envision a future where chemical safety assessment is rapid, cost-effective, and does not involve animal testing.

Contents

Acknowledgments

First and foremost, I would like to thank Prof. Martin Smieško. As my first supervisor, he guided me through my four years of Ph.D. studies, always assisting with his expertise and constructive criticism. His guidance was not only limited to scientific matters, he also extended my horizon in gardening and playing table tennis using office utensils. I always enjoyed spending time with him, whether it was at work, in his garden, hiking, or traveling through South Korea.

Next, I want to thank Prof. Markus Lill, the head of the Computational Pharmacy group. Although not an official supervisor in the scope of this thesis, he provided great assistance in my deep learning endeavors. Our discussions on the various projects I was involved in greatly improved my understanding of the matter and the quality of my work. I also highly appreciate his suggestions of eminent novelists, whose works I still enjoy to this day.

I also want to thank Prof. Daniel Ricklin, my second supervisor, and Dr. Christian Kramer, the external expert of this thesis. With his great attitude, Daniel turned our yearly status review meetings into a pleasant experience (which is rare for meetings!). I also deeply appreciate the time Christian is taking to assess my combined works of the past four years, knowing that his schedule is packed to the brim with responsibilities and commitments that demand his attention.

This thesis would definitely not have been possible without the great financial contribution of the Biographics Laboratory 3R foundation and the Animalfree Research foundation. I am hugely grateful for their trust in me and this project.

Even though my Ph.D. studies started with a nation-wide lockdown due to a global pandemic, I had the chance to get to know many amazing people. Over the past four years, I was happy to see the Computational Pharmacy group grow with researchers and students from diverse backgrounds. I am very grateful for all the interactions I had with the postdocs, Dr. Peter Rüthemann, Dr. Amr Abdallah, Dr. Jerôme Eberhardt, and the past and current Ph.D. students, Dr. Jacek Kędzierski, Florian Hinz, Justin Diamond, Roman Aschwanden, Matthew Masters, Dr. Soo Jung Lee, and Dr. André Fischer. I thorougly enjoyed the time we spent together, working, playing table tennis, or having long and intense discussions about various (and sometimes completely absurd) topics! Moreover, I am very glad to have had the opportunity to supervise five Master's students and interns, and I am thankful for their contributions to my projects. I wish Santhosh, Livius, Floriane, Cédric, and Valentin all the best in their future scientific and non-scientific endeavors.

Last but not least, I want to extend my gratitude to my family. I received huge support from my parents, Theres and Thomas, without whom my scientific career would not have been possible. They never grew tired of asking when I will finish my thesis, which is, of course, a Ph.D. student's favorite question. My brother Beni followed the scientific path two years before me and was never short of answers to any question I might have had. With her lighthearted nature, my sister Pamela helped me free my mind when I could not stop pondering a problem I was stuck with (sometimes by "convincing" me to join her workouts). I cannot express how incredibly grateful I am for the love and support I received from my girlfriend Rijana, not only during my Ph.D. studies but during the past 14 years. She was always happy to hear me talk about my day, even if it involved seemingly confused blabbering about something called "latent space" or "Transformers" (which is a movie, right?). I also apologize for the countless times I was completely absorbed in my work, my mind disentangled from reality, and my responses as vague as Ikea furniture instructions.

What is there that is not poison? All things are poison and nothing is without poison. Solely the dose determines that a thing is not a poison.

Theophrastus Bombast von Hohenheim

1

Introduction

Everything in nature is made up of atoms, which, in turn, combine to form molecules. These molecules are essential for life as we know it. As humans we are not just made up of molecules, we are also constantly exposed to a plethora of chemical compounds, some of which are known and well characterized, others are still unknown. Plants used as a food source contain a large amount of different chemical compounds known as secondary metabolites.[1] Although many of these compounds will not have observable effects on humans, some can have beneficial properties such as antioxidant,[2] antimicrobial,[3,4] or anti-inflammatory[5–7] effects, or unwanted properties such as cytotoxicity[8,9] or carcinogenicity.[10,11] However, the chemicals we ingest do not necessarily originate from the plant itself. With global industrialization, the use of pesticides and fertilizers has skyrocketed.[12,13] Often, residues of these chemicals can be found in the food we consume or the water we drink.[14–17]

However, food is not the only way through which humans are exposed to chemicals. Cosmetic products, shampoos, and perfumes contain various fragrances that we are exposing ourselves to. These chemical compounds can be inhaled or absorbed by the skin. Even clothing can be a source of chemical contamination.[18] One major source of chemical exposure is the intake of pharmaceutical drugs. Even though pharmaceutical products are usually very well characterized, adverse drug reactions remain common.[19] Given that safety concerns are the second most common reason for failed drug development projects,[20] rigorous assessment of potential toxic effects of drugs is essential for the pharmaceutical industry.

Although human health is generally of primary concern, chemical contamination often affects many different species. Volatile chemical compounds from combustion engines and industrial waste water are only a few sources of how humans cause increasing environmental pollution.[21,22] A common side effect of such anthropogenic environmental pollution is endocrine disruption, the interference of xenobiotic chemicals with the hormone system.[23–27] Since endocrine disruption can affect animal reproduction, it can have severe adverse effects for humans and wildlife.[28]

In a study in 2020, Wang et al. found that in 19 analyzed countries, more than 350,000 chemicals and mixtures were registered for use and production.[29] Many of them were not known to the public for various reasons. Thus, it is of great importance to characterize these chemicals to assess their effects on humans and wildlife. The field of toxicology assesses such effects through various in vivo, in vitro, and in silico methods. As Theophrastus Bombast von Hohenheim (better known as Paracelsus) recognized already 500 years ago, everything can be toxic depending on the dose. Therefore, a toxicologist usually describes the risk of a chemical compound as the product of hazard and exposure.[30] Hazard can be seen as the inherent ability of a chemical to harm the environment, humans, or wildlife. Exposure, on the other hand, is a measure of the probability and intensity of contact with a chemical. It includes e.g. the dose of a chemical to which a living being is exposed and the duration of the exposure. Examples of common hazardous chemicals are myristicin contained in nutmeg[31] or cyanide contained in almonds.[32] Although they are hazardous, humans are generally exposed to those compounds sufficiently low to pose no significant risk to human health. A chemical that has high exposure but is not hazardous is water. Humans consist mostly of water and can safely consume large amounts of water everyday due to its low hazard. However, even with a low hazard, too high exposure can be toxic or even lethal.[33]

1.1 Adverse Outcome Pathways

Adverse outcome pathways (AOPs) are a means of characterizing the complete mechanism of toxicity for a given adverse outcome. AOPs describe how a toxicant (or more generally, a stressor) can cause one or more key events by triggering a molecular initiating event, which ultimately leads to an adverse outcome (see Figure 1.1). They therefore comprise causal links between all steps involved in the formation of a chemical-induced adverse effect.[34] Connections between key events are called key event relationships, and it is further possible to make connections between different AOPs based on shared key events. AOPs include events at the (macro)molecular, cellular, organ, and individual levels.

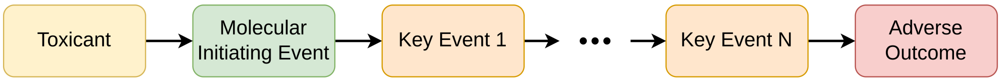


Figure 1.1: Elements of an AOP. There may be one or more key events that lead to an adverse outcome

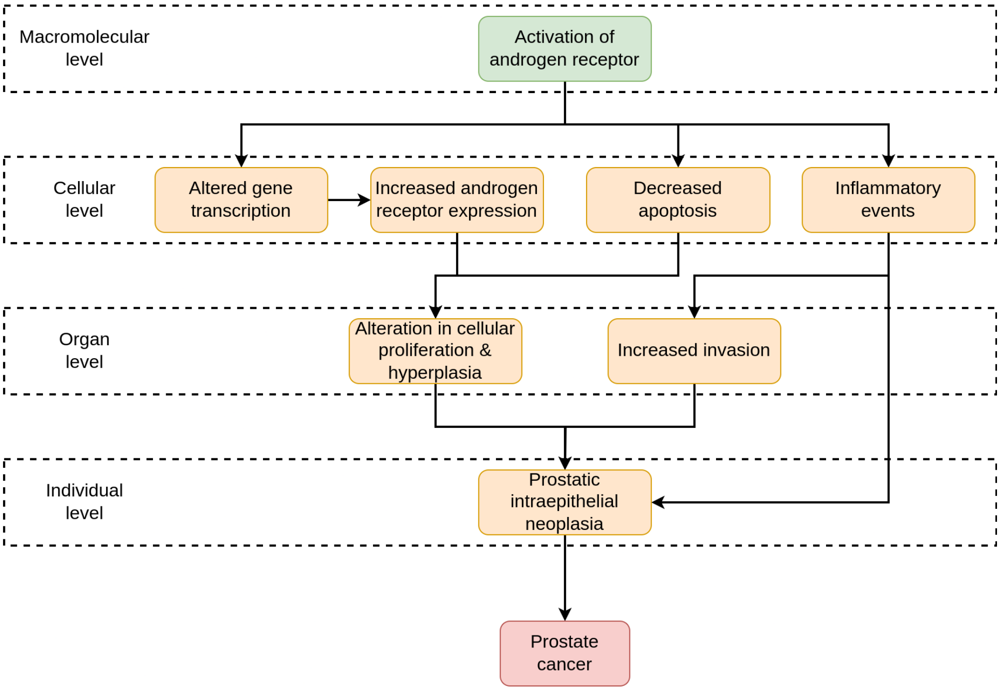

Macromolecular level
Activation of androgen receptor
Cellular level
Altered gene transcription
Increased androgen receptor expression
Decreased apoptosis
Inflammatory events
Organ level
Alteration in cellular proliferation & hyperplasia
Increased invasion
Individual level
Prostatic intraepithelial neoplasia
Prostate cancer

Figure 1.2: Complete AOP for the formation of prostate cancer based on androgen receptor activation. Figure adopted from AOP 495 in the AOP Wiki. The molecular initiating event, key events, and adverse outcome are shown with green, orange, and red color, respectively.

The Organization for Economic Co-operation and Development (OECD) developed the AOP knowledge base[35] which serves as a resource for AOP development and sharing. The AOP Wiki[36] is part of the AOP knowledge base and can be used to search for known AOPs. An example of a complete AOP is the formation of prostate cancer based on androgen receptor activation (AOP 495 in the AOP Wiki; Figure 1.2). This AOP begins with the activation of the androgen receptor as the molecular initiating event. On a cellular level, this leads to altered gene transcription, increased expression of androgen receptors, decreased apoptosis in epithelial cells, and inflammatory events in light-exposed tissues. This, in turn, causes increased invasion and alterations in cell proliferation leading to hyperplasia. Taken together, these effects can cause prostatic intraepithelial neoplasia leading to the formation of prostate cancer.

Due to the complexity of some AOPs, their development and review process can take several years. This contributes to the slow growth of the number of known and well-described AOPs.[37] Despite these challenges, AOPs are still widely used in current research.[38–42]

1.2 New Approach Methodologies

Already more than 60 years ago, Russell and Burch introduced the concept of the 3R, the replacement, reduction, and refinement of animal experimentation.[43,44] Animal welfare may be the most obvious motivation to replace animal models. However, there is also a long-lasting discussion about the value of animal models in chemical risk assessment and whether the findings of these experiments can be translated to humans.[45–51] Finally, conducting animal tests also comes with a much higher financial burden compared to non-animal experiments.[52]

Since the need for alternatives to animal testing has been recognized in many industries and regulatory bodies, the concept of new approach methodologies (NAMs) has been introduced.[53] NAMs include any methods such as in vitro, in silico, in chemico, or ex vivo that can be used for chemical hazard and risk assessment without the use of animals.[54–56] This thesis introduces an in silico NAM which we hope can one day contribute to the replacement, reduction, or refinement of animal testing.

1.3 Types of Toxicity

In order to assess the toxicity of a chemical compound, one first needs to understand exactly how molecules can exhibit toxic effects. Although not an exhaustive overview, this section will cover some of the most important types of toxicity.

1.3.1 Mutagenicity and Genotoxicity

The European Food Safety Authority defines mutagenicity as "the capacity to cause permanent, typically negative, changes to an organism and any offspring by altering the structure of its DNA".[57] This alteration can usually occur at the level of single genes, blocks of genes, or chromosomes.[58,59] A mutagenic substance is called a mutagen and can affect DNA in germ cells and somatic cells.[60,61] Mutagens can lead to different outcomes, such as cell death or cell growth, which eventually leads to tumors.[62–64] This means that not all mutagens are carcinogenic. Sodium azide is an example of a substance that is mutagenic but not carcinogenic.[65,66]

Since mutagenicity can have extremely severe consequences, it is imperative to develop a test system for it. A routinely used mutagenicity test is the so-called Ames test.[67,68] This test is a bacterial reverse mutation test. It uses a mutated bacterial strain that lost the ability to synthesize a certain amino acid (usually histidine or tryptophane). This bacterial strain is then grown on an agar plate lacking the specific amino acid. Because bacteria cannot synthesize the missing amino acid themselves, they will not be able to grow. However, upon the addition of a mutagen to the plate, bacteria sometimes mutate and revert to the wild type that is capable of synthesizing the amino acid and, therefore, growing in the medium.[69] The standard test protocol uses several different bacterial strains with different mutations.[70] Since a reverse mutation is not guaranteed, a positive Ames test is usually predictive of the mutagenicity of a compound, while a negative test is not conclusive.[71,72]

One mechanism of mutagenicity is the alkylation of DNA by electrophilic compounds. An example of a DNA alkylating substance is mustard gas which has been used for chemical warfare in the past.[74] However, DNA alkylating properties can also be used pharmaceutically in form of alkylating antineoplastic agents. These compounds alkylate DNA in tumor cells, leading to an alteration of the biological function of the cell.[75] Examples of therapeutically

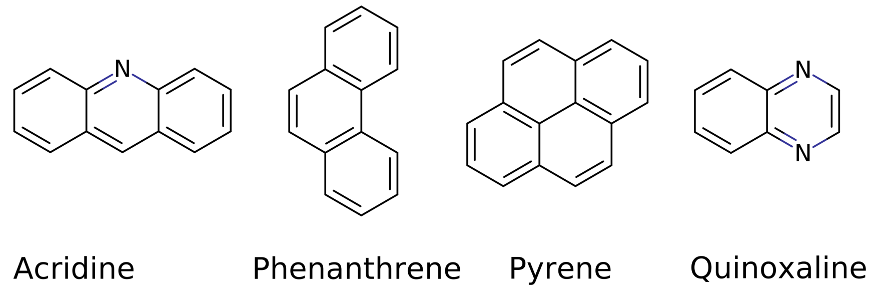


Figure 1.3: Molecular scaffolds of known DNA intercalating mutagens.[73]

used DNA alkylating agents include carmustine,[76,77] temozolomide,[78] busulfan,[79] and melphalan.[80] Another very common mechanism of mutagenicity is DNA intercalation in which a typically planar aromatic molecule binds between a base pair, thus altering the structure and function of the DNA.[73,81] Although DNA intercalating compounds bind non-covalently (and thus reversibly) to DNA, they can cause frameshift mutations.[82] Many molecular scaffolds with DNA intercalating properties have been identified.[83] Some examples include acridine,[84,85] phenanthrene,[86] pyrene,[87,88] and quinoxaline[89] (Figure 1.3).

Although all mutagens are genotoxic because they can permanently alter DNA, not all genotoxic substances are mutagenic. In fact, some DNA intercalating molecules such as quinoline are genotoxic without having mutagenic effects.[82,90]

1.3.2 Carcinogenicity

Cancer formation is extremely complex and involves many different cellular processes. Chemical-induced carcinogenesis is thus only one of many ways that lead to cancer formation.[91] Although many carcinogenic substances such as benzo[a]pyrene and aflatoxins act through genotoxicity, there are various other mechanisms of carcinogenicity that do not involve DNA alteration.[91,92] Peroxisome proliferator-activated receptors (PPARs) are a type of nuclear receptor involved in glucose and fatty acid metabolism and energy homeostasis.[93] It has been shown that activation of PPARγ can inhibit glioma growth in vivo.[94] Similarly, activation of PPARβ/δ can regulate lung cancer growth.[95] However, the role of these receptors is much more complex. Contrary to the previous finding, Han et al. suggested that activation of PPARβ/δ can also promote lung cancer growth.[96] Furthermore, chronic activation of PPARα has been shown to induce the development of liver cancer.[97] Thus, activation or inhibition of these receptors is closely intertwined with the formation of cancers in humans.

Adjacent cells are able to communicate with each other through gap junctions, which are formed by close membrane contacts permeated with numerous channels. These channels allow the exchange of ions and small molecules, which is a means of cellular communication.[98,99] This communication is essential for cell homeostasis and its disruption can activate a cascade of processes, ultimately leading to the formation of cancer.[100,101] The insecticides chlordane and dichlorodiphenyltrichloroethane (DDT) have been shown to down-regulate the gap junctional intercellular communication, leading to an increased probability of cancer formation.[102–105]

1.3.3 Endocrine Disruption

The endocrine system (i.e. hormonal system) is a fundamental messenger system in the human body and regulates various processes such as metabolism, sleep, growth, stress, reproduction, and development. In this messenger system, hormones usually bind to specific receptors, which in turn leads to a change in cellular function. Different hormones thereby bind to a wide array of receptors, either on the surface or inside of cells.[106] Therefore, it is not surprising that the disruption of this intricate system can lead to various adverse outcomes. These include, but are not limited to, alterations in sperm quality, fertility, nervous system and immune function, or malformations of the sex organs, endometriosis, and cancers.[107,108] Chemical compounds that interfere with the hormonal system are called endocrine disruptors and can be found in many foods and everyday products.

Perhaps one of the best known endocrine disruptors is bishenol A, a compound that is often used in plastics and epoxy resins. Bishpenol A was shown to exhibit estrogenic effects (among others), causing kidney damage, obesity, and decreased female reproductive health.[109,110] Phytoestrogens are substances contained in plants and foods such as soy, oats, and coffee and have estrogenic or antiestrogenic effects.[111,112] A third type of estrogenic (and androgenic) chemicals are phthalates, which are commonly used as plasticizers. Like other endocrine disruptors, they can cause a reduction in reproductive health in both men and women.[113–115]

A metabolite of the insecticide DDT, which has already been introduced in Section 1.3.2 due to its carcinogenicity, also acts as an androgen antagonist.[116] After a spill of a mixture of pesticides including DDT on Lake Apopka in 1980, the fertility of alligators living in this lake has decreased significantly.[117,118] This effect has later been linked to DDT and its metabolites.[119,120]

Polychlorinated biphenyls used, e.g., as coolants and flame retardants, as well as perfluorooctanoic acid, widely used in the past, e.g. in the production of non-stick cookware, both cause thyroid disruption. These compounds are also classified as persistent organic pollutants because they cannot be easily degraded by chemical or biological processes.[121–123]

1.3.4 Drug-Induced Liver Injury

Most people have taken paracetamol to treat occasional headaches at least once in their life. However, this seemingly harmless drug is responsible for approximately 50% of acute liver failure cases in some western countries.[124,125] This type of toxicity falls under the umbrella term drug-induced liver injury (DILI). DILI describes damage to the liver at any cellular level induced by a drug or its metabolites.[126] A South Korean study reported an extrapolated 12 yearly cases of DILI per 100,000 people. According to their findings, herbal medications were one of the main causes of DILI.[127]

It is assumed that a main reason for the liver's susceptibility to drug-induced damage is the fact that most drugs are taken up into hepatocytes where they are metabolized.[126] In some cases, such as with paracetamol, reactive metabolites are formed which can covalently bind to macromolecules and disrupt normal cell function.[128] Diclofenac, a commonly used non-steroidal anti-inflammatory drug, is metabolized by cytochrome P450 (CYP) 2C8 and UDP-Glucuronosyltransferase 2B7 into reactive quinone imine and acyl glucuronide species which can form covalent adducts with proteins, eventually leading to hepatotoxicity.[129,130]

In the late 1990s, troglitazone, a PPARγ agonist used for the treatment of type 2 diabetes, was approved as a drug. Only three years later, it has already been withdrawn from the market due to numerous reports of hepatotoxicity. Although the initial mechanism of toxicity was thought to be due to the formation of reactive metabolites, this has not been proven conclusively.[131,132] Another possible explanation could be the activation of downstream processes invoked by PPAR binding that ultimately leads to apoptosis.[133,134] Inhibition of bile salt export pumps and subsequent accumulation of toxic bile salts in hepatocytes could be another explanation for troglitazone hepatotoxicity.[135,136] In fact, this mechanism is well known in the formation of DILI and is the driving factor of the toxic effects of e.g. bosentan, a drug used to treat pulmonary hypertension.[137,138]

DILI is the cause of approximately 20% of drug development failures in clinical phases and 30% of market withdrawals.[139] Because it encompasses very diverse mechanisms of toxicity, its prediction is extremely challenging.[126] Nevertheless, DILI plays a major role in drug development.[140]

1.3.5 Off-Target Activity

A pharmaceutical target that is intended to be modulated by a drug is called an on-target. In certain cases, binding of a drug to its intended target can lead to toxicity, referred to as on-target toxicity. For example, statins are designed to inhibit the 3-hydroxy-3-methylglutaryl coenzyme A reductase in the liver, leading to lower cholesterol levels. However, binding of statins to the same target in different tissues, such as muscles, can cause adverse effects.[141,142] Conversely, any protein that is not the intended target of a drug is considered an off-target. Thus, this definition of on- and off-targets usually applies only to pharmaceutical, cosmetic, or sometimes agricultural settings because other industrial chemicals, used e.g. in the production of plastics or fuels, do not have an intended biological target.

Binding of a chemical to off-targets is generally undesired, as this may lead to adverse effects. Unspecific kinase inhibitors binding to kinases other than their primary target (i.e., off-targets) can, for example, lead to cardiotoxicity.[143] Since kinases are very flexible and often share similar binding site topologies, their inhibitors often bind to dozens of different kinases.[144] This can be problematic because in these cases, the true mechanism of action of a drug may be promiscuous and unknown.

However, in certain cases, off-target activity can be desired. Spironolactone is a drug that was originally developed as a potassium-sparing diuretic due to its ability to inhibit the mineralcorticoid receptor.[145] It was later found out that the same compound also inhibits the androgen receptor (that is, an off-target).[146,147] This effect was therapeutically exploited to treat acne and hair loss.[148] After that, it was discovered that spironolactone can also induce degradation of xeroderma pigmentosum group B protein which is involved in DNA repair. Therefore, it has the potential to be used in cancer therapy.[149] This shows that whether a biological target is considered an on-target or an off-target depends only on its intended use, and the off-target of one scientist may be the on-target of another.

Lin et al. have shown that many cancer therapeutics are still effective even after CRISPR-based knock-out of their intended target.[150] They suggest that these drugs actually work through off-target effects. In addition, they discuss that this lack of knowledge about the true mode of action of cancer drugs may be one of the reasons why oncology projects have the highest failure rate in the pharmaceutical industry.[20]

Off-target toxicity does not have a specific toxicity endpoint. Many of the previously described toxic effects such as the induction of liver cancer due to PPARα activation (Section 1.3.2) or the various forms of endocrine disruption (Section 1.3.3) can be effects of off-target activity. In these cases, binding of a molecule to an off-target can be seen as a molecular initiating event (cf. Section 1.1).

Toxicologically relevant effects, as seen in the case of kinases, and therapeutic success stories (such as spironolactone) highlight the value of investigating off-target activities. The promiscuous mode of action exhibited by anti-cancer drugs further underscores the importance of a detailed examination of these effects.

1.4 Computational Toxicology Methods

This section covers some of the most widely used methods in computational toxicology. In many cases, the same techniques can also be utilized in drug development projects that are not related to toxicology.

1.4.1 Dose-Response Modeling

Fritz Haber was a German chemist who received the Nobel prize in chemistry in 1918 for the invention of the Haber-Bosch process. However, he has also become known as the “father of chemical warfare" because of his experiments with toxic gases during the first world war. During this time, he stated that the concentration of a gas (C) multiplied with the time of exposure (t) can be used to determine the toxicity of a gas (K; Equation 1.1).[151]

$$C \cdot t = K \tag{1.1}$$

Although this equation is only applicable in some special cases, it has been widely adopted in dose-response modeling and is known as "Haber's law".[152]

In general, dose-response modeling aims to find a relationship between the administered dose of a substance and the occurrence of a biological effect such as mortality. Miller et al. addressed the limitations of Haber's law and showed that it is a special case of Equation 1.2, where C is the administered concentration of a substance, C_0 is a threshold concentration below which no biological effects can be observed, K is a constant biological response (e.g. mortality), t is the time at which response K can be observed, and α and β are parameters to control the relative importance of C and t.[153]

$$(C - C_0)^{\alpha} \cdot t^{\beta} = K \tag{1.2}$$

They showed that Haber's law corresponds to a special situation with no threshold concentration ($C_0 = 0$) and $\alpha = \beta = 1$. In modern dose-response modeling, determining C_0 and fitting α and β to existing data allows to assess the risk of a chemical substance.[154] However, since K does not necessarily need to be defined as a toxicological response, the same approach can also be used in pharmaceutical development.

1.4.2 Physiologically-Based Pharmacokinetics

The purpose of physiologically-based pharmacokinetics (PBPK) includes the prediction of drug-time profiles, first-in-human dose, drug-drug interactions, and pharmacokinetics across age and race.[155] In PBPK modeling, different organs are considered in the form of individual compartments connected by the blood system. Each of these compartments is described by tissue volume, arterial and venous blood flow rate, a tissue-partition coefficient, and permeability. Organs commonly included are the brain, thymus, lung, heart, stomach, pancreas, spleen, intestinal tract, liver, kidneys, adipose tissue, muscle, and bone.[156]

PBPK models combine physiological information on the included compartments with drug properties such as molecular weight, solubility, basicity or acidity, plasma protein binding, kinetics (usually Michaelis-Menten kinetics), and metabolic information. A large set of (differential) equations is used to calculate and model pharmacokinetic aspects such as clearance, mean residence time, or blood-plasma concentration ratio.[157] Due to the high complexity of these models, commercial PBPK platforms have been developed that allow easy generation and use of such models.[158,159]

For the pharmaceutical industry, PBPK modeling is fundamental for understanding drug-time profiles and selecting, e.g., first-in-human doses. Since PBPK models provide information on exposure to chemical substances, they are also valuable for toxicologists and regulatory bodies (where the method is sometimes referred to as physiologically-based toxicokinetics).[160,161]

1.4.3 Rule-Based Models

Rule-based systems, also called (human) expert systems, are models that leverage existing knowledge about the toxicity of specific structural fragments. These fragments, or "structural alerts", have been experimentally connected to certain toxic endpoints. Generally, such rule-based systems follow a simple if-then premise in which the probability of toxicity is determined based on the presence of structural alerts. These systems can be further divided into human and statistical expert systems. Human expert systems rely on clearly defined knowledge obtained by human experts, while statistical systems try to correlate structural information with outcomes using (mostly) regression models.[162,163] Human expert systems are generally more accepted because of their interpretability, but are also prone to a higher rate of false negatives due to the still limited human understanding of toxicity mechanisms.[163] Furthermore, the absence of a structural alert does not necessarily mean that the compound is non-toxic.

This technology is routinely used in the prediction of mutagenicity and carcinogenicity where many structural alerts have been identified.[73,164–166] Expert systems are included in several software packages, some of the most prominent being ToxTree[167], Lazar[168], and Derek.[169,170]

1.4.4 Read-Across

In the 1860s, Alexander Butlerov recognized that chemicals with similar structures have similar properties.[171,172] This assumption is still used today in a method called read-across. In read-across, properties of a query molecule are predicted based on the properties of one or more similar "analog" molecules.[173] These properties are usually toxicological (or pharmaceutical) endpoints, but can, in theory, also be of physicochemical nature.

Chemical similarity can be described in several different ways. Usually, chemical features are encoded in (binary) feature vectors, called fingerprints, which can then be compared to each other.[174] To compare fingerprints, metrics such as the Tanimoto or Dice coefficient or the Tversky index can be used.[175–177] Chemical features can be extracted from 2D and 3D structures based on connectivity properties and chemical substructures.[174,178] Commonly used fingerprints include the Morgan fingerprint,[179] Daylight fingerprint,[180] and extended connectivity fingerprints.[181] Such molecular fingerprints are still often applied in research and new fingerprints are regularly developed.[182–185]

1.4.5 Quantitative Structure-Activity Relationship

More than 60 years ago, Corwin Hansch and his colleagues published an article in Nature, in which they described what is often considered the first quantitative structure-activity relationship (QSAR) model.[186] They found a way of correlating the octanol-water partition coefficient (logP) with the Hammett substituent constant to predict the concentration of auxins (plant growth regulators) that induces a 10% growth in their test system. Their model is described in Equation 1.3, where C is the auxin concentration, π is an approximation of the octanol-water partition coefficient, σ is the Hammett substituent constant, and k, k', k'', and ρ are programmable parameters.

$$\log(\frac{1}{C}) = k\pi + k'\pi^2 + \rho\sigma + k'' \tag{1.3}$$

In their study, they found $k = 4.08$, $k' = 2.14$, $\rho = 2.78$, and $k'' = 3.36$ to give quite accurate predictions.

More generally, a QSAR model is any kind of model that predicts a biological or toxicological outcome based on (physico)chemical properties. This is formalized in Equation 1.4, where P_c is the predicted outcome of compound c, $f(\cdot)$ is the QSAR model, and θ is a feature vector containing chemical properties of compound c.

$$P_c = f(\theta_c) \tag{1.4}$$

A model that is trained on congeneric compounds is called a local QSAR model while a global QSAR model is trained on diverse chemical substances.[154] Local QSAR models thereby have a more limited applicability domain but are often more accurate than global QSAR models.

There are various options to choose as models. They can be either linear, such as linear[187] or multiple linear regression[188] or partial least squares[189], or non-linear like support vector machines[190], random forest[191], K-nearest neighbor[192], or artificial neural networks[190]. Simpler models usually have better interpretability, while more complex models are better able to capture complex relationships in the provided data.

Much consideration should be put into the selection of the chemical features used to train a QSAR model. Ideally, the chosen features should be relevant for the prediction of the endpoint, reduce the chance of overfitting, and be physically or chemically meaningful, thus providing good interpretability.[154] Several methods such as forward selection, backward elimination, or genetic algorithms can be used to filter the often hundreds or even thousands of calculated molecular descriptors.[193–195]

1.5 Drug-Target Interaction Prediction

The main part of this thesis focuses on the prediction of off-target interactions. Therefore, it is essential to have an overview of existing methods used to predict drug-target interactions (DTIs). In this section, we will cover classical, machine learning-based, and deep learning-based techniques for predicting DTIs. These DTIs can be of qualitative (e.g., the generation of protein-ligand binding modes) or quantitative (e.g., the prediction of binding affinities) nature.

Often, the terms artificial intelligence (AI), machine learning, and deep learning are used interchangeably. However, the terms refer to different computational areas and should there-

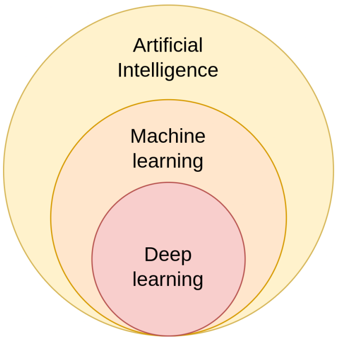


Figure 1.4: Relationship between AI, machine learning, and deep learning.

fore be distinguished (cf. Figure 1.4). AI covers all computational algorithms that resemble human intelligence. Thus, simple QSAR models such as described in Section 1.4.5 can already be seen as AI systems. Machine learning is a subcategory of AI that includes algorithms that can learn from data without explicit human instructions and make predictions based on the learned parameters. Deep learning is itself a subcategory of machine learning which involves the use of deep artificial neural networks.[196] Hence, although deep learning and machine learning are both AI systems, it is important to distinguish between these techniques.

1.5.1 Classical Methods

Molecular Docking

Molecular docking is a technique to generate protein-ligand binding complexes and approximate their free energy of binding.[197] Although molecular docking is most often applied to small molecule-protein complexes, it can also be employed on protein-protein systems.[198–202] However, in this overview, only docking of small molecules to proteins will be covered. Many different docking programs have been developed over the past decades, but their principles are the same. Every docking program consists of at least two steps, the pose generation using a search algorithm and the pose scoring using a scoring function.

Search algorithms allow placing ligands within the binding site of a protein to generate their predominant binding modes. They are usually classified as either systematic, deterministic, or stochastic.[203] Systematic search methods most commonly use well-defined steps to adapt the translation, rotation, and torsions of the small molecule to find optimal poses.[204] Due to this combinatorial approach, they are generally good at sampling the conformational space, but are also computationally expensive (depending on the defined granularity). Deterministic search algorithms generate conformations based on previously generated poses by following certain rules, e.g. to find energy minima. Stochastic methods, on the other hand, include certain randomness. This method comprises commonly applied techniques such as genetic algorithms,[205,206] monte carlo,[207,208] and simulated annealing.[209,210] It should be noted that the computational cost of a search algorithm usually heavily depends on the number of rotatable bonds of the small molecule. This happens because the search space increases exponentially with the number of torsional degrees of freedom.

Each of the generated poses needs to be scored using a scoring function. Scoring functions are usually either physics-based, knowledge-based, empirical, machine learning-based, or a combination thereof.[211] Physics-based scoring functions include the use of a force field that describes the different interactions between the protein and the ligand. Examples for docking programs with physics-based scoring functions are DOCK[212] and LigandFit.[213] Knowledge-based scoring functions are derived from statistical observations in large libraries of experimentally determined protein-ligand complexes and are included e.g. in PoseScore[214] and MotifScore.[215] Programs such as smina[208] and Glide[216] include empirical scoring functions, which include scaling factors for the individual interaction terms which can be tuned to fit experimentally observed data. More recently, machine learning-based scoring functions have emerged. While they are usually not directly included in a docking program, they are used for re-scoring generated poses. The implemented scoring function is used to assess the quality of the generated poses and rank them accordingly. Gnina is a recent example of a docking program that includes deep learning-based scoring functions.[217]

The quality of a ligand pose is usually validated on the basis of the root mean squared deviation (RMSD) to the crystal structure. An arbitrarily chosen RMSD threshold of 2 Å is often applied to determine if a pose is considered "native" or not. Thus, poses with an RMSD of less than 2 Å to the experimentally determined pose are usually considered high quality.[218,219] The use of RMSD, however, has some pitfalls. For example, it does not consider interaction patterns. Thus, a ligand can have the same interactions as in the crystal structure while having a large RMSD (e.g. in near-symmetrical ligands). Conversely, ligands can have a small RMSD due to, e.g. a small rotation of a small ligand but lose important interactions with the protein. In this case the pose would not be considered correct although the RMSD may be small. Additionally, ligands can have flexible, solvent-exposed tails. In these cases, the docking pose can often deviate from the crystal structure in these flexible parts of the ligand, while the rest has good overlap. This would lead to an increased RMSD although the pose would be considered correct.[220] For this reason, alternatives to the use of the RMSD such as the relative displacement error[221] or the interactions-based accuracy classification[222] have been developed. While they address some of the shortcomings of the RMSD, they also have their limitations and so far, no perfect pose-validation method is available.

Molecular docking can therefore be used for qualitative (using the docking pose) and quantitative (using the score) DTI prediction. However, it has some well-known limitations. Most of the time, ligands are docked to rigid protein structures, which does not allow simulation of induced fit effects. Multiple tools allow flexible docking, in which the side chains (and sometimes also the backbone) of the protein are treated flexibly and can therefore adapt to the ligand.[223–227] While these methods can give more accurate results, they are also much more computationally expensive due to the explosion of degrees of freedom. Especially when using rigid-body docking, sampling of high-quality poses can become a limitation of the performance.[228] Even in the case where good poses can be sampled, the scoring functions are often not very accurate, leading to a wrong ranking of poses and to wrong estimations of the binding free energy.[229,230] Often, docking tools work well in certain protein groups and worse on others.[230] Thus, the use of a consensus approach in which several different programs are applied is often preferred.[231,232]

Molecular Mechanics - Generalized Born Surface Area

While molecular docking is generally fast, its predicted binding affinities are also inaccurate. Molecular mechanics/generalized born surface area (MM-GBSA) is a method that usually provides better predictions but comes at a higher computational cost.[233,234] It calculates the binding free energy by comparing the energy of the protein-ligand bound state with the unbound state. This is shown in Equation 1.5 where G_{P-L}, G_P, and G_L are the energy of the protein-ligand complex, unbound receptor, and unbound ligand, respectively.

$$\Delta G_{bind} = G_{P-L} - (G_P + G_L) \tag{1.5}$$

Each energy term is thereby defined as a combination of a molecular mechanics force field contribution, a solvation free energy contribution, and a term for the conformational entropy. Equation 1.6 shows this relationship.

$$G_X = E_X^{MM} + G_X^{solv} - TS_X \tag{1.6}$$

Where E_X^{MM} is the molecular mechanics contribution, usually calculated based on a force field, G_X^{solv} is the solvation free energy contribution, and TS_X is the conformational entropy term. The solvation free energy is defined as $G_X^{solv} = G_X^{GB} + G_X^{SA}$, where G_X^{GB} represents the generalized Born solvation term and G_X^{SA} is a nonpolar solvation term, usually calculated using the solvent-accessible surface area. The generalized Born solvation term is an approximation of the generally more accurate but also more computationally expensive Poisson-Boltzmann electrostatic contribution and is used to calculate the solvation free energy in implicit solvent.[235,236] Directly using the Poisson-Boltzmann equation to calculate the solvation free energy would turn this method into molecular mechanics / Poisson-Boltzmann surface area (MM-PBSA).

MM-GBSA and MM-PBSA can be used for single-point free energy calculations of protein-ligand complexes generated by molecular docking. However, it is often applied to a set of snapshots from molecular dynamics simulations. This allows to account for flexible adaptation of the binding site residues as well as averaging over different conformations. In this approach, when combining Equations 1.5 and 1.6, the binding free energy is defined according to Equation 1.7.

$$\Delta G_{bind} = \langle \Delta E^{MM} \rangle + \langle \Delta G^{solv} \rangle - \langle \Delta TS \rangle \tag{1.7}$$

In this equation, $\langle \cdot \rangle$ represents the average over all states generated by molecular dynamics simulation.

Free Energy Perturbation

Free energy perturbation (FEP) is a method based on statistical mechanics to accurately calculate absolute or relative binding free energies. Since FEP methods require running many molecular dynamics simulations, they have a high computational cost. To understand FEP, we first have to define how the energy in a closed system is described. Equation 1.8 shows the definition of the Helmholtz free energy F as the difference between the internal energy U and the product of temperature T and entropy S.

$$F = U - TS \tag{1.8}$$

The internal energy can also be described as the sum of the energies E of all microstates in a system multiplied with the probability p_i of a system being in microstate i (Equation 1.9).

$$U = \sum_i E_i p_i \tag{1.9}$$

Further, the entropy S can also be defined in terms of the probabilities of microstates as shown in Equation 1.10.

$$S = -k_B \sum_i p_i \ln p_i \tag{1.10}$$

Where k_B is the Boltzmann constant. Therefore, the Helmholtz free energy can also be defined as shown in Equation 1.11.

$$F = \sum_i E_i p_i + k_B T \sum_i p_i \ln p_i \tag{1.11}$$

We can further simplify this definition using the partition function Z defined in Equation 1.12.

$$Z = \sum_i \exp\left(-\frac{E_i}{k_B T}\right) \tag{1.12}$$

By rearranging the partition function to isolate E_i, we can simplify Equation 1.11 to get Equation 1.13.

$$F = -k_B T \ln\left(\sum_i \exp\left(-\frac{E_i}{k_B T}\right)\right) \tag{1.13}$$

This function can be used to calculate the absolute or relative free energy of protein-ligand systems.

There are many different ways to calculate absolute binding free energies using FEP.[237–239] Generally, a system is simulated in an unbound state A and the sampled ligand configurations are mapped into a protein binding site to obtain the bound state B. The binding free energy can then be calculated based on the difference between the bound and unbound states. This is shown in Equation 1.14 where $\langle \cdot \rangle_A$ represents the ensemble average over the configurations sampled from state A.

$$\Delta F = -k_B T \ln \left\langle \exp\left(-\frac{E_B - E_A}{k_B T} \right) \right\rangle_A \tag{1.14}$$

Calculating the absolute binding free energy this way is often very difficult because simulating the ligand in the unbound state usually does not efficiently sample the conformations the ligand would assume in a bound state. Therefore, absolute binding free energy calculation with FEP often suffers from a sampling problem.

For this reason, the relative binding free energy is usually preferred, as it is easier to calculate. This method calculates the relative difference in binding free energy of two structurally similar molecules. This can be achieved by using the thermodynamic cycle such as in Figure 1.5. In this cycle, ΔF_1 and ΔF_2 represent the binding free energies of ligands A and B, respectively, ΔF_3 represents the alchemical transformation of unbound ligand A to unbound ligand B and ΔF_4 represents the alchemical transformation of bound ligand A to bound ligand B. In this case, $\Delta F_1 + \Delta F_4 - \Delta F_2 - \Delta F_3 = 0$ must be true by the definition of the thermodynamic cycle. Therefore, we can state that $\Delta F_2 - \Delta F_1 = \Delta F_4 - \Delta F_3$. This means that instead of directly calculating and comparing the binding free energies of ligands A and B (ΔF_1 and ΔF_2), one can also calculate the alchemical transformation of ligand A to ligand B

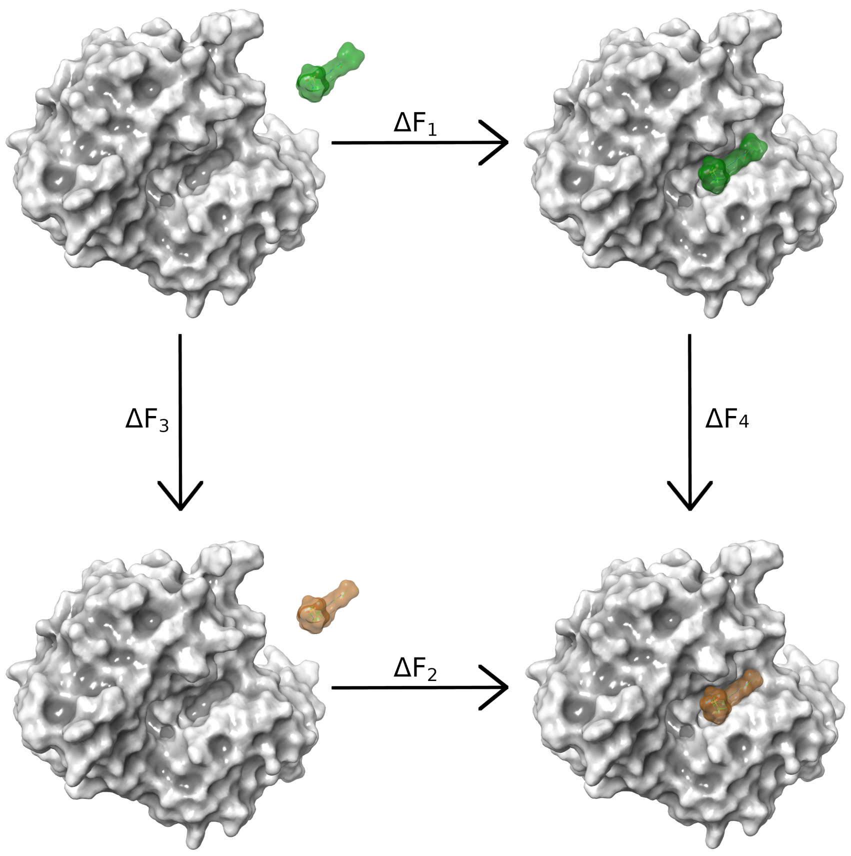


Figure 1.5: Thermodynamic cycle used in relative binding free energy calculation using FEP. Although not shown in the figure, the protein and ligand are assumed to be solvated in water. The ligands represented by green and orange surfaces are structurally highly similar.

in the unbound and bound state (ΔF_3 and ΔF_4). Since in this approach, it is not necessary to change from unbound to bound state or vice versa, the sampling of low-energy conformations is much easier. The transformation from one ligand to another is thereby simulated in several intermediate steps (called Lambda windows). This necessity for many intermediate simulations is the main reason why FEP calculations are so computationally expensive.

Due to their rigorous sampling and solid foundation in physical principles, FEP methods are currently considered the gold standard for free energy calculation, apart from quantum mechanical methods.[240–242]

1.5.2 Machine Learning-based Methods

Machine learning-based techniques for DTI prediction can be roughly divided into two categories. The first category contains similarity-based methods and the second category contains feature vector-based methods. Although there are other types of methods, such as matrix factorization or network-based approaches, these will not be covered here. For more detailed information about all of these methods, several reviews can be consulted.[243–247]

Generally, feature vector-based methods require the extraction of sets of features from drug molecules and target proteins. Intuitively, the selection of these features can greatly influence the performance of a model.[248] Features that do not adequately capture the information contained in the underlying data make it difficult for a model to learn complex relationships. On the other hand, features that are too specific can increase the chance of overfitting. Different types of molecular fingerprints (see Section 1.4.4) and physicochemical or structure-derived descriptors are used to represent small molecules.[249] For the presentation of proteins, several methods have been developed over the past decades. These include the position-specific scoring matrix,[250] pseudo amino acid composition,[251] dipeptide composition,[252] Composition, Transition and Distribution,[253] enhanced amino acid composition,[254] and dipeptide deviation from expected mean.[255] There are also tools that automatically create vector representations of proteins.[256]

The extracted drug and protein feature vectors are used as inputs to various machine learning models that allow classification or regression, including support vector machines, K-nearest neighbor, logistic regression, decision trees, and random forest models.[257–259]

Selecting the best features for feature vector-based methods can be difficult. Similarity-based methods do not have this problem. For these methods, a similarity matrix is needed for both the drugs and the targets. A very commonly used method in this field is the nearest neighbor method. In this method, the most similar drug or target (i.e., the nearest neighbor) is used to predict interaction pairs of new drugs or targets. For example, Equation 1.15 shows how the interaction profile of a new drug ($p_{d_{new}}$) could be predicted based on the similarity $s_d(\cdot,\cdot)$ between the new drug d_{new} and its nearest neighbor $d_{nearest}$, and the interaction profile of the nearest neighbor ($p_{d_{nearest}}$).

$$p_{d_{new}} = s_d(d_{new}, d_{nearest})p_{d_{nearest}} \tag{1.15}$$

Bipartite local models are another similarity-based method. In this technique, bipartite graphs of drugs and targets are constructed in which the edges between drugs and targets symbolize the existence of an interaction between them. This method is especially useful if information about interactions between drugs and targets in the graph is missing. For example, consider the case where the information on the interaction between drug d_i and target t_j is missing, but d_i has known interactions with other targets in $T = \{t_1, t_2, ..., t_n\}$ and target t_j has known interactions with other drugs in $D = \{d_1, d_2, ..., d_n\}$. In this case, a vector v_{d_i} can be constructed by checking for edges between d_i and all other targets in T. If there is a known edge to t_1, element 1 in v_{d_i} would be set to 1, if there is no edge, it would be set to -1. This process is repeated for all targets in T (excluding t_j). The same approach can be used to create a vector v_{t_j} by checking the edges between the target t_j and the drugs in D (excluding d_i). These vectors can then be used to predict the presence or absence of an edge between d_i and t_j by using a support vector machine.[247,260,261]

Another similarity-based technique that uses support vector machines is the pairwise kernel method. This approach uses the similarity between drugs (kernel function $s_d(\cdot,\cdot)$) and targets (kernel function $s_t(\cdot,\cdot)$) to compute the similarity between drug-target pairs (pairwise kernel $s_{dt}(\cdot, \cdot)$) as shown in Equation 1.16.

$$s_{dt}((d_i, t_i), (d_j, t_j)) = s_d(d_i, d_j) \cdot s_t(t_i, t_j) \tag{1.16}$$

A drug-target pair similarity matrix can then be used to train a support vector machine which predicts new drug-target interactions. Compared to bipartite local models, this approach is more efficient because it uses a single model that generalizes to all drug-target pairs.[247,262]

Unlike the classical methods introduced in Section 1.5.1, these machine learning-based methods do not require 3D structures of proteins or ligands as input. Due to the general sparsity of such 3D data, this allows these methods to be used on much larger existing datasets. However, for the same reason, they are often not as interpretable as 3D-based methods such as molecular docking.

1.5.3 Deep Learning-based Methods

In the field of AI, deep learning has made huge progress in recent years. It revolutionized multiple fields such as speech recognition,[263] image generation,[264] real-time object detection,[265] robotics,[266] and protein structure prediction.[267] Most recently, ChatGPT, a large language model developed by OpenAI has disrupted society beyond the scientific community.[268–270] In this section, we will cover some deep learning-based approaches to tackle the problem of DTI prediction.

Convolutional Neural Networks

Convolutional neural networks (CNNs) usually combine convolution operations and pooling layers followed by a fully connected layer. In a convolution operation, a kernel is applied to the input tensor. A kernel is a tensor that contains learnable parameters and is usually

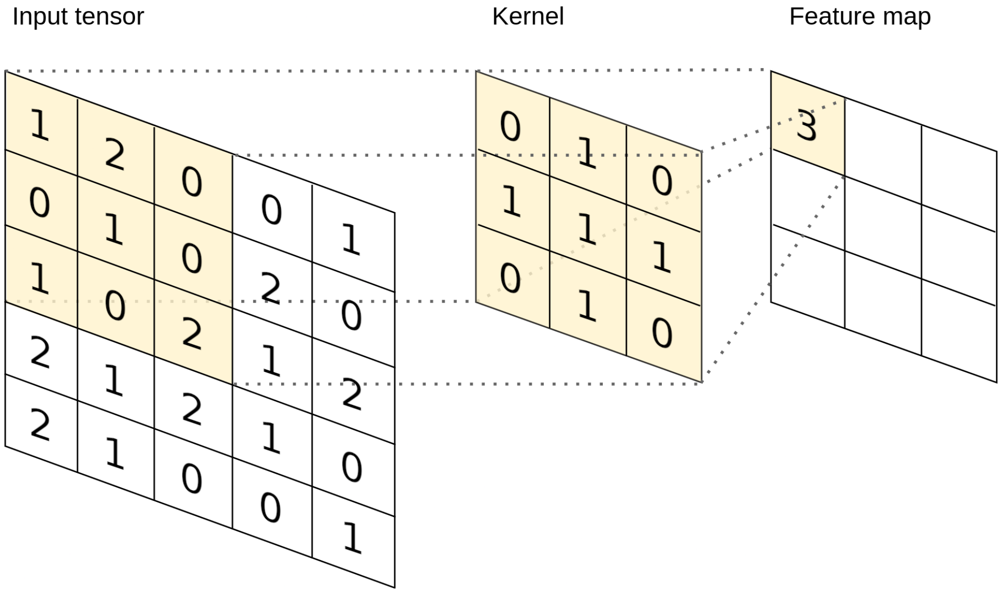


Figure 1.6: Convolution operation in a CNN. The pair-wise product of the kernel and the current window in the input tensor is calculated and the result is summed to form the feature map. The kernel is shifted across the input tensor and this operation is repeated for every step. Figure reproduced from Yamashita et al.[271], licensed under the Creative Commons Attribution 4.0 International License,[272] with minor changes.

smaller than the input tensor. The input tensor can be 1-, 2-, or 3-dimensional and the kernel must match its dimensions. During convolution operations, the kernel is shifted across the input tensor and for every position, the element-wise product between the current window of the input tensor and the kernel is calculated and summed (see Figure 1.6 for a visualized convolution operation).[271] Normally, a pooling layer follows after a convolution layer to reduce the size of the tensor. This combination of convolutional layers and pooling layers allows CNNs to capture shift-invariant features of the input data and detect patterns. Most often, the final layer is flattened and passed through a fully connected layer to make a prediction for a given task. CNNs are mostly applied in computer vision, but they can also be used for DTI prediction.

CNNs can be used to create fairly simple models such as the one proposed by Hasan Mahmud et al.[273] They applied a 1D CNN to feature vectors (similar to those used in the machine learning methods introduced in Section 1.5.2) to predict DTIs. In their paper, they show that this simple method outperforms classical machine learning models such as K-nearest neighbors or extreme gradient boosting.

Monteiro et al. used a 1D CNN to learn DTIs from a combination of protein amino acid sequences and ligand SMILES strings.[274] They also combined the CNN with an autoencoder trained on protein and ligand features. Autoencoders are a specific type of encoder-decoder network that can be used for unsupervised learning. The encoder maps the input data into so-called latent space. This latent space usually has a lower dimensionality than the input and acts as an information bottleneck. The decoder then tries to recreate the input from the latent space encoding. Thus, to successfully reconstruct the input, the model must learn to encode as much relevant information in the latent space as possible. This method can also be seen as a form of compression where the latent space encoding acts as the compressed data. In the study by Monteiro et al., the combination of a CNN with an autoencoder did not significantly improve performance over a CNN alone.[274]

Mahmoud et al. used a 3D CNN to predict protein-ligand binding affinities based on 3D protein-ligand complexes.[275] They showed that this approach significantly outperformed the AutoDock Vina docking program in finding native ligand poses.[276] Furthermore, they demonstrated how the inclusion of explicit water molecules in their predictions further improved performance and they were able to correctly identify a native pose under the first-ranked pose in almost 90% of the cases.[275]

Graph Neural Networks

A molecule can also be seen as a graph $G = (V, E)$ consisting of vertices (atoms) V and edges (bonds) E. This fact is leveraged in graph convolutional networks (GCNs). The following introduction to GCNs is based on the original publication by Kipf and Welling, the inventors of GCNs.[277] In a GCN, each node x_i is described in a $N \times D$ feature matrix X where N is the number of nodes and D is the number of features describing each node. The graph structure is usually provided in form of an adjacency matrix A. The goal of a GCN is to update the node features X to use them to predict a downstream task. The updated node features Z can be written as $Z = f(X, A)$ where $f(\cdot, \cdot)$ is a non-linear function. Generally, a GCN consists of L layers and the node features are iteratively updated layer by layer. The update of the node features can be described by Equation 1.17.

$$H^{l+1} = f\big(H^{(l)}, A\big) \tag{1.17}$$

Here, H represent the hidden (intermediate) representation of the node features with $H^{(0)} = X$ and $H^{(L)} = Z$. In its simplest form, the non-linear function $f(\cdot, \cdot)$ can be described according to Equation 1.18 where $\sigma(\cdot)$ is a non-linear activation function and $W^{(l)}$ is a weight matrix of layer l.

$$f(H^{(l)}, A) = \sigma\big(AH^{(l)}W^{(l)}\big) \tag{1.18}$$

In their publication, Kipf and Welling provide an example for a GCN consisting of two layers. This is described in Equation 1.19.

$$Z = f(X, A) = \operatorname{softmax}\big(\hat{A}\operatorname{ReLU}\big(\hat{A}XW^{(0)}\big)W^{(1)}\big) \tag{1.19}$$

It should be noted that here, A is defined as $A + I$ where I is the identity matrix. This is done to add self-connections to the nodes (i.e., to allow them to be updated by themselves). They further defined $\hat{A} = \hat{D}^{-\frac{1}{2}}A\hat{D}^{-\frac{1}{2}}$ where $\hat{D}$ is the diagonal node degree matrix, which represents a symmetric normalization introduced to retain the scale of the node features after the graph convolution operations.

Lim et al. used a GCN on 3D protein-ligand binding modes to predict DTIs.[278] They showed that this method outperformed CNNs in most tasks. They also created a model with an additional gated attention mechanism which further improved the performance. While CNNs are translation-invariant, they are not invariant to rotations and scaling. GCNs on the other hand, are invariant to translation and rotation (scaling is usually not an issue with molecular graphs) as long as no explicit atom coordinates are provided. This may be one reason for the better performance of GCNs over CNNs reported by Lim et al.

Deep learning is often seen as a black box and is considered to lack interpretability.[279] This could be improved by integrating more physics into otherwise uninterpretable models. Moon and colleagues created PIGNet, a "physics-informed deep learning model toward generalized drug–target interaction predictions".[280] They used a gated graph attention network (a special form of GCN) to predict node features in 3D protein-ligand complexes. These node features were then used as parameters for a physics-based scoring function that is directly integrated into the model. Their scoring function contained terms for hydrogen bonding, van der Waals interactions, hydrophobic interactions, and interactions with metals. These terms were used to calculate atom-atom pairwise binding energies. The partial binding energies of all atom pairs were then summed up to obtain a final score. They recently published an improved version of PIGNet which includes an additional entropy regularization based on the number of rotatable bonds in a molecule.[281] They show that their model is on par or better than many deep learning-based and classical DTI prediction methods. By calculating interaction energies based on atom-atom pairs, Moon et al. effectively remove all explicit information on the ligand and protein atomic environment. This may benefit the generalizability of the model. In this regard, Volkov and colleagues reported that the use of explicit non-covalent protein-ligand interactions does not improve a model's performance if explicit structural information about the protein and ligand is provided.[282] They suggested that in these cases, the model primarily learns to memorize structures. This means that such a model will perform well on input that is similar to the training data, but may perform badly for structurally different data. Therefore, such a model will not be much better than any similarity-based method and will share its limitations. Thus, trying to remove what Volkov et al. call "hidden biases" may be essential for the development of generalizable models.

Transformer Models

In 2017, Vaswani et al. published an article in which they introduced a novel model architecture for natural language processing called the Transformer model.[283] Since then, this model architecture (and adaptations thereof) has been used in some of the most powerful deep learning models known so far. This includes AlphaFold2 for the prediction of protein structures[267] and ChatGPT, the groundbreaking large language model by OpenAI.[270] Transformers are encoder-decoder networks that act on sequences. There are two parts that contribute greatly to their success. These are the positional encoding and the scaled dot-product attention mechanism. Positional encoding is required to provide a sense of the order of elements in a sequence. This step is required because the Transformer model does not make use of convolutions or recurrence. In the original article, they used a combination of sine and cosine functions with different frequencies to encode a sequence. The attention mechanism uses queries, keys, and value matrices. This is described in Equation 1.20, where Q, K, and V are tensors containing the queries, keys, and values, and d_k is the dimensionality of the keys.

$$\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1.20}$$

The scaling by $\frac{1}{\sqrt{d_k}}$ is added for increased stability by preventing vanishing gradients in cases where d_k is very large. In self-attention, the same tensor is passed through three different linear layers to obtain Q, K, and V. This is shown in Equations 1.21 where X is the input tensor and W_Q, W_K, and W_V are the learnable weigths to calculate the queries, keys, and values, respectively.

$$\begin{aligned} Q &= XW_Q \\ K &= XW_K \\ V &= XW_V \end{aligned} \tag{1.21}$$

In cross attention, the keys and values are generated by the encoder and the queries are provided by the first multi-head attention block in the decoder. Further details about the architecture of Transformers can be found in the original publication.[283] Classical Transformers that are used, e.g., for language translation, predict sequences element by element where each prediction depends on the previously predicted elements. Such models are called autoregressive.[284–286]

Since their first introduction, Transformers have not only been used on sequence-based data such as text and amino acid sequences. With minor adaptations, Transformers can also work on molecular graphs and have been often used to predict DTIs.[287–292] In the works presented in Chapters 2-5, we also employed Transformer models for different tasks.

In recent years, diffusion models have gained increased attention. These models work by gradually adding noise to data and training a model to denoise the data again. This allows the trained model to generate new data from noise with possible conditioning to guide the generation. An overview of diffusion models is provided by Croitoru et al.[293] Diffusion models have also been used to tackle the DTI problem. DiffDock used reverse diffusion of random starting poses of small molecules to denoise translational, rotational, and torsional degrees of freedom to generate valid protein-ligand binding poses.[294]

RoseTTAFold All-Atom was developed in the lab of David Baker and is another generative diffusion model used in DTI prediction. Extending the functionality of AlphaFold2, this model is able to generate not only 3D structures of proteins, but also complete protein-ligand, protein-nucleic acid, protein-metal, and covalently modified protein structures.[295]

Finally, Bryant et al. from Frank Noé's group have developed Umol, a universal molecular network.[296] This model does not have a diffusion-based architecture, but follows the evoformer approach that was already used in AlphaFold.[297,298] Similar to RoseTTAFold All-Atom, Umol is also capable of predicting full-atomistic protein-ligand complexes. According to their article, it performs better than DiffDock and RoseTTAFold All-Atom in predicting correct binding poses.

1.5.4 Datasets

In a newspaper article published in 1957 in The Times, William D. Mellin, a US Army Specialist, was quoted to say "If the problem has been sloppily programmed, the answer will be just as incorrect. If the programmer made mistakes, the machine will make mistakes. It can't correct them because it can't do one thing. It cannot think for itself".[299] In this sense, the term "garbage in, garbage out" was coined. It refers to the assumption that flawed or erroneous input data will lead to equally flawed output data. In deep learning, where data are at the core, this concept is fundamental. Therefore, researchers need to train their models with the right data of the right quality.

For the problem of DTI prediction, several databases are available. One of the most widely used is the PDBbind dataset.[300,301] In its latest 2020 release, this database contains the 3D crystal structures of more than 19,000 protein-ligand complexes with accompanying binding affinity data. The PDBbind set is divided into three distinct parts: the general, the refined, and the core set. The general set is the largest of the three and contains structures of sometimes questionable quality. This is because there are no quality criteria for the structures in the general set and they can therefore suffer, for example, from poor resolution, artifacts introduced by interactions with crystal mates, or multiple co-crystallized ligands in a single binding site. The refined set is formally a subset of the general set, but with the growing size of the PDBbind database, it has been distributed separately. It contains more than 5,000 protein-ligand complexes that meet certain quality criteria. These criteria concern the included 3D structures (e.g., structures should not have a low resolution > 2.5 Å, covalently bound ligands, or obvious steric clashes between the protein and the ligand) and the associated binding data (e.g., the affinity data should be K_i/K_d rather than IC50, the reported affinity should be an exact value between 10 mM and 1 pM, and the protein used in the assay should match the crystal structure). These are arguably critical quality checks, and omitting them could be detrimental to a deep learning model trained on these data. Finally, the core set is a small subset of the refined set with high diversity that is often used for benchmarking. Except for the core set, the PDBbind database receives annual updates. Chapter 2 of this thesis investigates the impact of different parts and splits of the PDBbind dataset on the performance of a deep learning model for DTI prediction.

Recently, Siebenmorgen et al. addressed some of the issues of the PDBbind dataset.[302] They used semi-empirical quantum mechanical methods to refine the almost 20,000 structures in this dataset. In addition, they ran 10 ns long molecular dynamics simulations of almost 17,000 complexes in explicit solvent. They suggest that their dataset is more suitable for AI-based methods due to its superior quality. The dataset is distributed under the name MISATO.[303]

The Schrödinger FEP benchmark set contains 103 protein structures from 14 subsets.[240,304] Each protein has an associated set of ligands with experimental binding free energies. This dataset has been used to benchmark Schrödinger's FEP+ tool and to analyze the maximum achievable accuracy of relative binding free energy calculation.

The Binding MOAD ("Mother Of All Databases") currently contains over 40,000 protein-ligand complexes obtained from the PDB.[305–307] For more than 15,000 complexes, additional binding data are provided. This database only includes complexes where the ligand is either a small molecule, a co-factor, a peptide of not more than 10 amino acids, or an oligonucleotide with less than 5 nucleotides. Only crystal structures with a resolution of at least 2.5 Å are considered.

The Davis and KiBA sets are two commonly used benchmark datasets containing binding affinity information for kinase inhibitors.[308,309] The Davis set contains 72 inhibitors that were tested against 442 kinases, while the KiBA dataset includes 467 kinases and more than 50,000 compounds, although not all of these were tested against all kinases.

While the PDB is by far the most popular database for experimentally determined protein structures (e.g. from X-ray, cryo EM, or NMR),[307] several different databases provide information on the binding affinities of ligands. The most notable are PubChem,[310], ChEMBL,[311], and BindingDB.[312] According to their statistics, PubChem alone contains more than 100 million unique compounds and almost 300 million reported bioactivities.

References

[1] Eng Soon Teoh. Secondary Metabolites of Plants. In *Medicinal Orchids of Asia*, pages 59–73. Springer International Publishing, Cham, 2016.

[2] Sofia C. Lourenço, Margarida Moldão-Martins, and Vítor D. Alves. Antioxidants of Natural Plant Origins: From Sources to Food Industry Applications. *Molecules*, 24(22):4132, 11 2019.

[3] Marjorie Murphy Cowan. Plant Products as Antimicrobial Agents. *Clinical Microbiology Reviews*, 12(4):564–582, 10 1999.

[4] Natalia Vaou, Elisavet Stavropoulou, Chrysa Voidarou, Christina Tsigalou, and Eugenia Bezirtzoglou. Towards Advances in Medicinal Plant Antimicrobial Activity: A Review Study on Challenges and Future Perspectives. *Microorganisms*, 9(10):2041, 9 2021.

[5] Katarina Radovanović, Neda Gavarić, and Milica Aćimović. Anti-Inflammatory Properties of Plants from Serbian Traditional Medicine. *Life*, 13(4):874, 3 2023.

[6] Yashika Gandhi, Ravi Kumar, Jyotika Grewal, Hemant Rawat, Sujeet K. Mishra, Vijay Kumar, Santosh K. Shakya, Vipin Jain, Gajji Babu, Preeti Sharma, Arjun Singh, Ravindra Singh, and Rabinarayan Acharya. Advances in anti-inflammatory medicinal plants and phytochemicals in the management of arthritis: A comprehensive review. *Food Chemistry Advances*, 1:100085, 10 2022.

[7] Lina Karrat, Mohammad Yaser Abajy, and Ream Nayal. Investigating the anti-inflammatory and analgesic properties of leaves ethanolic extracts of Cedrus libani and Pinus brutia. *Heliyon*, 8(4):e09254, 4 2022.

[8] Godwin Upoki Anywar, Esezah Kakudidi, Hannington Oryem-Origa, Andreas Schubert, and Christian Jassoy. Cytotoxicity of Medicinal Plant Species Used by Traditional Healers in Treating People Suffering From HIV/AIDS in Uganda. *Frontiers in Toxicology*, 4:832780, 5 2022.

[9] Merajuddin Khan, Mujeeb Khan, Syed F. Adil, and Hamad Z. Alkhathlan. Screening of potential cytotoxic activities of some medicinal plants of Saudi Arabia. *Saudi Journal of Biological Sciences*, 29(3):1801–1807, 3 2022.

[10] W.M.F. Jongen and F.O. Dorgelo. Naturally occurring carcinogens and modulating factors in food of plant origin. *Netherlands Journal of Agricultural Science*, 34(3):395–404, 8 1986.

[11] Xiaoqing Guo and Nan Mei. Aloe vera : A review of toxicity and adverse clinical effects. *Journal of Environmental Science and Health, Part C*, 34(2):77–96, 4 2016.

[12] Anket Sharma, Vinod Kumar, Babar Shahzad, Mohsin Tanveer, Gagan Preet Singh Sidhu, Neha Handa, Sukhmeen Kaur Kohli, Poonam Yadav, Aditi Shreeya Bali, Ripu Daman Parihar, Owias Iqbal Dar, Kirpal Singh, Shivam Jasrotia, Palak Bakshi, M. Ramakrishnan, Sandeep Kumar, Renu Bhardwaj, and Ashwani Kumar Thukral. Worldwide pesticide usage and its impacts on ecosystem. *SN Applied Sciences*, 1(11):1446, 11 2019.

[13] Cameron I. Ludemann, Armelle Gruere, Patrick Heffer, and Achim Dobermann. Global data on fertilizer use by crop and by country. *Scientific Data*, 9(1):501, 8 2022.

[14] Bodil Hamborg Jensen, Annette Petersen, Pernille Bjørn Petersen, Tue Christensen, Sisse Fagt, Ellen Trolle, Mette Erecius Poulsen, and Jens Hinge Andersen. Cumulative

dietary risk assessment of pesticides in food for the Danish population for the period 2012–2017. *Food and Chemical Toxicology*, 168:113359, 10 2022.

[15] Changjian Li, Huimin Zhu, Changyan Li, He Qian, Weirong Yao, and Yahui Guo. The present situation of pesticide residues in China and their removal and transformation during food processing. *Food Chemistry*, 354:129552, 8 2021.

[16] Paula Medina-Pastor and Giuseppe Triacchini. The 2018 European Union report on pesticide residues in food. *EFSA Journal*, 18(4):e06057, 4 2020.

[17] Hossein Yousefi and Bahareh Karimi Douna. Risk of Nitrate Residues in Food Products and Drinking Water. *Asian Pacific Journal of Environment and Cancer*, 6(1):69–79, 3 2023.

[18] Ike van der Veen, Anne-Charlotte Hanning, Ann Stare, Pim E.G. Leonards, Jacob de Boer, and Jana M. Weiss. The effect of weathering on per- and polyfluoroalkyl substances (PFASs) from durable water repellent (DWR) clothing. *Chemosphere*, 249:126100, 6 2020.

[19] Jamie J Coleman and Sarah K Pontefract. Adverse drug reactions. *Clinical Medicine*, 16(5):481–485, 10 2016.

[20] Richard K. Harrison. Phase II and phase III failures: 2013–2015. *Nature Reviews Drug Discovery*, 15(12):817–818, 12 2016.

[21] Mohammed Moufid, Carlo Tiebe, Nezha El Bari, Damien Ali Hamada Fakra, Matthias Bartholmai, and Benachir Bouchikhi. Pollution parameters evaluation of wastewater collected at different treatment stages from wastewater treatment plant based on E-nose

and E-tongue systems combined with chemometric techniques. *Chemometrics and Intelligent Laboratory Systems*, 227:104593, 8 2022.

[22] Nikolaos Rousis, Maria Denardou, Nikiforos Alygizakis, Aikaterini Galani, Anna Bletsou, Dimitrios Damalas, Niki Maragou, Kevin Thomas, and Nikolaos Thomaidis. Assessment of Environmental Pollution and Human Exposure to Pesticides by Wastewater Analysis in a Seven-Year Study in Athens, Greece. *Toxics*, 9(10):260, 10 2021.

[23] Elizabeth C. Plunk and Sean M. Richards. Endocrine-Disrupting Air Pollutants and Their Effects on the Hypothalamus-Pituitary-Gonadal Axis. *International Journal of Molecular Sciences*, 21(23):9191, 12 2020.

[24] Agostino Di Ciaula and Piero Portincasa. The Role of Environmental Pollution in Endocrine Diseases. pages 1–31. Springer, Cham, 2019.

[25] Sana Ullah, Shahid Ahmad, Xinle Guo, Saleem Ullah, Sana Ullah, Ghulam Nabi, and Kunyuan Wanghe. A review of the endocrine disrupting effects of micro and nano plastic and their associated chemicals in mammals. *Frontiers in Endocrinology*, 13:1084236, 1 2023.

[26] Massimo Pettoello-Mantovani, Flavia Indrio, Ruggiero Francavilla, and Ida Giardino. The effects of climate change and exposure to endocrine disrupting chemicals on children's health: A challenge for pediatricians. *Global Pediatrics*, 4:100047, 6 2023.

[27] Manoj Kumar, Devojit Kumar Sarma, Swasti Shubham, Manoj Kumawat, Vinod Verma, Anil Prakash, and Rajnarayan Tiwari. Environmental Endocrine-Disrupting Chemical Exposure: Role in Non-Communicable Diseases. *Frontiers in Public Health*, 8:553850, 9 2020.

[28] V.L. Marlatt, S. Bayen, D. Castaneda-Cortès, G. Delbès, P. Grigorova, V.S. Langlois, C.J. Martyniuk, C.D. Metcalfe, L. Parent, A. Rwigemera, P. Thomson, and G. Van Der Kraak. Impacts of endocrine disrupting chemicals on reproduction in wildlife and humans. *Environmental Research*, 208:112584, 5 2022.

[29] Zhanyun Wang, Glen W. Walker, Derek C. G. Muir, and Kakuko Nagatani-Yoshida. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. *Environmental Science & Technology*, 54(5):2575–2584, 3 2020.

[30] Swedish Chemicals Agency. Hazard and risk assessment of chemicals-an introduction GUIDANCE 7. www.kemi.se/en/guidance-on-national-chemicals-control.

[31] Jamie E. Ehrenpreis, Carol DesLauriers, Patrick Lank, P. Keelan Armstrong, and Jerrold B. Leikin. Nutmeg Poisonings: A Retrospective Review of 10 Years Experience from the Illinois Poison Center, 2001–2011. *Journal of Medical Toxicology*, 10(2):148–151, 6 2014.

[32] Nadia Chaouali, Ines Gana, Amira Dorra, Fathia Khelifi, Anouer Nouioui, Wafa Masri, Ines Belwaer, Hayet Ghorbel, and Abderazzek Hedhili. Potential Toxic Levels of Cyanide in Almonds (Prunus amygdalus), Apricot Kernels (Prunus armeniaca), and Almond Syrup. *ISRN Toxicology*, 2013:1–6, 9 2013.

[33] D J Farrell and L Bower. Fatal water intoxication. *Journal of Clinical Pathology*, 56(10):803–a–804, 10 2003.

[34] Gerald T. Ankley, Richard S. Bennett, Russell J. Erickson, Dale J. Hoff, Michael W. Hornung, Rodney D. Johnson, David R. Mount, John W. Nichols, Christine L. Russom, Patricia K. Schmieder, Jose A. Serrano, Joseph E. Tietge, and Daniel L. Villeneuve.

Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. *Environmental Toxicology and Chemistry*, 29(3):730–741, 3 2010.

[35] AOP knowledge base. https://aopkb.oecd.org/.

[36] AOP-Wiki. https://aopwiki.org/.

[37] Terje Svingen, Daniel L Villeneuve, Dries Knapen, Eleftheria Maria Panagiotou, Monica Kam Draskau, Pauliina Damdimopoulou, and Jason M O'Brien. A Pragmatic Approach to Adverse Outcome Pathway Development and Evaluation. *Toxicological Sciences*, 184(2):183–190, 11 2021.

[38] Shihori Tanabe, Jason O'Brien, Knut Erik Tollefsen, Youngjun Kim, Vinita Chauhan, Carole Yauk, Elizabeth Huliganga, Ruthann A. Rudel, Jennifer E. Kay, Jessica S. Helm, Danielle Beaton, Julija Filipovska, Iva Sovadinova, Natalia Garcia-Reyero, Angela Mally, Sarah Søs Poulsen, Nathalie Delrue, Ellen Fritsche, Karsta Luettich, Cinzia La Rocca, Hasmik Yepiskoposyan, Jördis Klose, Pernille Høgh Danielsen, Maranda Esterhuizen, Nicklas Raun Jacobsen, Ulla Vogel, Timothy W. Gant, Ian Choi, and Rex FitzGerald. Reactive Oxygen Species in the Adverse Outcome Pathway Framework: Toward Creation of Harmonized Consensus Key Events. *Frontiers in Toxicology*, 4:887135, 7 2022.

[39] Janani Ravichandran, Bagavathy Shanmugam Karthikeyan, and Areejit Samal. Investigation of a derived adverse outcome pathway (AOP) network for endocrine-mediated perturbations. *Science of The Total Environment*, 826:154112, 6 2022.

[40] Laura Aliisa Saarimäki, Jack Morikka, Alisa Pavel, Seela Korpilähde, Giusy del Giudice, Antonio Federico, Michele Fratello, Angela Serra, and Dario Greco. Toxicogenomics Data for Chemical Safety Assessment and Development of New Approach

Methodologies: An Adverse Outcome Pathway-Based Approach. *Advanced Science*, 10(2):2203984, 1 2023.

[41] Jördis Klose, Lu Li, Melanie Pahl, Farina Bendt, Ulrike Hübenthal, Christian Jüngst, Patrick Petzsch, Astrid Schauss, Karl Köhrer, Ping Chung Leung, Chi Chiu Wang, Katharina Koch, Julia Tigges, Xiaohui Fan, and Ellen Fritsche. Application of the adverse outcome pathway concept for investigating developmental neurotoxicity potential of Chinese herbal medicines by using human neural progenitor cells in vitro. *Cell Biology and Toxicology*, 39(1):319–343, 2 2023.

[42] You Song, Keke Zheng, Dag Anders Brede, Tânia Gomes, Li Xie, Yetneberk Kassaye, Brit Salbu, and Knut Erik Tollefsen. Multiomics Point of Departure (moPOD) Modeling Supports an Adverse Outcome Pathway Network for Ionizing Radiation. *Environmental Science & Technology*, 57(8):3198–3205, 2 2023.

[43] WMS Russell and RL Burch. The principles of humane experimental technique. 1959.

[44] Robert C. Hubrecht and Elizabeth Carter. The 3Rs and Humane Experimental Technique: Implementing Change. *Animals*, 9(10):754, 9 2019.

[45] Aysha Akhtar. The Flaws and Human Harms of Animal Experimentation. *Cambridge Quarterly of Healthcare Ethics*, 24(4):407–419, 10 2015.

[46] R.J. Wall and M. Shani. Are animal models as good as we think? *Theriogenology*, 69(1):2–9, 1 2008.

[47] Harry Olson, Graham Betton, Denise Robinson, Karluss Thomas, Alastair Monro, Gerald Kolaja, Patrick Lilly, James Sanders, Glenn Sipes, William Bracken, Michael Dorato, Koen Van Deun, Peter Smith, Bruce Berger, and Allen Heller. Concordance of

the Toxicity of Pharmaceuticals in Humans and in Animals. *Regulatory Toxicology and Pharmacology*, 32(1):56–67, 8 2000.

[48] Isabella Wy Mak, Nathan Evaniew, and Michelle Ghert. Lost in translation: animal models and clinical trials in cancer treatment. *American journal of translational research*, 6(2):114–8, 2014.

[49] Samuel M. Cohen. The relevance of experimental carcinogenicity studies to human safety. *Current Opinion in Toxicology*, 3:6–11, 4 2017.

[50] Gail A. Van Norman. Limitations of Animal Studies for Predicting Toxicity in Clinical Trials. *JACC: Basic to Translational Science*, 4(7):845–854, 11 2019.

[51] Cathalijn H. C. Leenaars, Carien Kouwenaar, Frans R. Stafleu, André Bleich, Merel Ritskes-Hoitinga, Rob B. M. De Vries, and Franck L. B. Meijboom. Animal to human translation: a systematic scoping review of reported concordance rates. *Journal of Translational Medicine*, 17(1):223, 12 2019.

[52] Lucy Meigs, Lena Smirnova, Costanza Rovida, Marcel Leist, and Thomas Hartung. Animal testing and its alternatives - the most important omics is economics. *ALTEX*, 35(3):275–305, 7 2018.

[53] European Chemicals Agency. New approach methodologies in regulatory science, 4 2016. 10.2823/543644.

[54] Anna J. van der Zalm, João Barroso, Patience Browne, Warren Casey, John Gordon, Tala R. Henry, Nicole C. Kleinstreuer, Anna B. Lowit, Monique Perron, and Amy J. Clippinger. A framework for establishing scientific confidence in new approach methodologies. *Archives of Toxicology*, 96(11):2865–2879, 11 2022.

[55] Andreas O. Stucki, Tara S. Barton-Maclaren, Yadvinder Bhuller, Joseph E. Henriquez, Tala R. Henry, Carole Hirn, Jacqueline Miller-Holt, Edith G. Nagy, Monique M. Perron, Deborah E. Ratzlaff, Todd J. Stedeford, and Amy J. Clippinger. Use of new approach methodologies (NAMs) to meet regulatory requirements for the assessment of industrial chemicals and pesticides for effects on human health. *Frontiers in Toxicology*, 4:964553, 9 2022.

[56] Sebastian Schmeisser, Andrea Miccoli, Martin von Bergen, Elisabet Berggren, Albert Braeuning, Wibke Busch, Christian Desaintes, Anne Gourmelon, Roland Grafström, Joshua Harrill, Thomas Hartung, Matthias Herzler, George E.N. Kass, Nicole Kleinstreuer, Marcel Leist, Mirjam Luijten, Philip Marx-Stoelting, Oliver Poetz, Bennard van Ravenzwaay, Rob Roggeband, Vera Rogiers, Adrian Roth, Pascal Sanders, Russell S. Thomas, Anne Marie Vinggaard, Mathieu Vinken, Bob van de Water, Andreas Luch, and Tewes Tralau. New approach methodologies in human regulatory toxicology – Not if, but how and when! *Environment International*, 178:108082, 8 2023.

[57] mutagenicity | EFSA. https://www.efsa.europa.eu/en/glossary/mutagenicity.

[58] Shahbeg S. Sandhu, Te-Hsiu Ma, Yan Peng, and Xiaodong Zhou. Clastogenicity evaluation of seven chemicals commonly found at hazardous industrial waste sites. *Mutation Research/Genetic Toxicology*, 224(4):437–445, 12 1989.

[59] Hang Yu, Zhihong Chen, Keqi Hu, Zongying Yang, Meiqi Song, Zihuan Li, and Yungang Liu. Potent Clastogenicity of Bisphenol Compounds in Mammalian Cells—Human CYP1A1 Being a Major Activating Enzyme. *Environmental Science & Technology*, 54(23):15267–15276, 12 2020.

[60] Richard J. Albertini and Debra A. Kaden. Mutagenicity monitoring in humans:

Global versus specific origin of mutations. *Mutation Research/Reviews in Mutation Research*, 786:108341, 10 2020.

[61] Joanna Kaplanis, Benjamin Ide, Rashesh Sanghvi, Matthew Neville, Petr Danecek, Tim Coorens, Elena Prigmore, Patrick Short, Giuseppe Gallone, Jeremy McRae, Loukas Moutsianas, Chris Odhams, Jenny Carmichael, Angela Barnicoat, Helen Firth, Patrick O'Brien, Raheleh Rahbari, and Matthew Hurles. Genetic and chemotherapeutic influences on germline hypermutation. *Nature*, 605(7910):503–508, 5 2022.

[62] Andrew Thresher, Robert Foster, David J. Ponting, Susanne A. Stalford, Rachael E. Tennant, and Robert Thomas. Are all nitrosamines concerning? A review of mutagenicity and carcinogenicity data. *Regulatory Toxicology and Pharmacology*, 116:104749, 10 2020.

[63] Takehiko Nohmi and Masahiko Watanabe. Mutagenicity of carcinogenic heterocyclic amines in Salmonella typhimurium YG strains and transgenic rodents including gpt delta. *Genes and Environment*, 43(1):38, 9 2021.

[64] Bennett Van Houten and Neil M. Kad. Single-cell mutagenic responses and cell death revealed in real time. *Proceedings of the National Academy of Sciences*, 115(28):7168–7170, 7 2018.

[65] Aras Türkoğlu, Metin Tosun, and Kamil Haliloğlu. Mutagenic effects of sodium azide on in vitro mutagenesis, polymorphism and genomic instability in wheat (Triticum aestivum L.). *Molecular Biology Reports*, 49(11):10165–10174, 11 2022.

[66] U.S. Department of Health and Human Sservices. Toxicology and Carcinogenesis Studies of Sodium Azide in F344/N Rats. Technical report, 1991.

[67] Bruce N. Ames, William E. Durston, Edith Yamasaki, and Frank D. Lee. Carcinogens are Mutagens: A Simple Test System Combining Liver Homogenates for Activation and Bacteria for Detection. *Proceedings of the National Academy of Sciences*, 70(8):2281–2285, 8 1973.

[68] Bruce N. Ames, Frank D. Lee, and William E. Durston. An Improved Bacterial Test System for the Detection and Classification of Mutagens and Carcinogens. *Proceedings of the National Academy of Sciences*, 70(3):782–786, 3 1973.

[69] Errol Zeiger. The test that changed the world: The Ames test and the regulation of chemicals. *Mutation Research/Genetic Toxicology and Environmental Mutagenesis*, 841:43–48, 5 2019.

[70] B N Ames, J McCann, and E Yamasaki. Methods for detecting carcinogens and mutagens with the Salmonella/mammalian-microsome mutagenicity test. *Mutat. Res.; (Netherlands)*, 31, 1975.

[71] Errol Zeiger. Carcinogenicity of Mutagens: Predictive Capability of the Salmonella Mutagenesis Assay for Rodent Carcinogenicity. *CANCER RESEARCH*, 47:1287–1296, 1987.

[72] Errol Zeiger, Joseph K. Haseman, Michael D. Shelby, Barry H. Margolin, Raymond W. Tennant, and H. E. Holden. Evaluation of four in vitro genetic toxicity tests for predicting rodent carcinogenicity: Confirmation of earlier results with 41 additional chemicals. *Environmental and Molecular Mutagenesis*, 16(S18):1–14, 1 1990.

[73] Romualdo Benigni and Cecilia Bossa. Mechanisms of Chemical Carcinogenicity and Mutagenicity: A Review with Implications for Predictive Toxicology. *Chemical Reviews*, 111(4):2507–2536, 4 2011.

[74] M. Kircher and M. Brendel. DNA alkylation by mustard gas in yeast strains of different repair capacity. *Chemico-Biological Interactions*, 44(1-2):27–39, 4 1983.

[75] Lakshmaiah Sreerama. Alkylating Agents. In *Encyclopedia of Cancer*, pages 132–136. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[76] S. H. D. M. Faria, J. G. Teleschi, L. Teodoro, and M. O. Almeida. Computational investigation of the carmustine (BCNU) alkylation mechanism using the QTAIM, IQA, and NBO models. *Structural Chemistry*, 32(1):79–96, 2 2021.

[77] Lijiao Zhao, Lili Li, Jie Xu, and Rugang Zhong. Comparative investigation of the DNA inter-strand crosslinks induced by ACNU, BCNU, CCNU and FTMS using high-performance liquid chromatography–electrospray ionization tandem mass spectrometry. *International Journal of Mass Spectrometry*, 368:30–36, 7 2014.

[78] E.S. Newlands, M.F.G. Stevens, S.R. Wedge, R.T. Wheelhouse, and C. Brock. Temozolomide: a review of its discovery, chemical properties, pre-clinical development and clinical trials. *Cancer Treatment Reviews*, 23(1):35–61, 1 1997.

[79] Benigno C. Valdez, David Murray, Yago Nieto, Yang Li, Guiyun Wang, Richard E. Champlin, and Borje S. Andersson. Synergistic cytotoxicity of the DNA alkylating agent busulfan, nucleoside analogs and suberoylanilide hydroxamic acid in lymphoma cell lines. *Leukemia & Lymphoma*, 53(5):973–981, 5 2012.

[80] Anastazja Poczta, Aneta Rogalska, and Agnieszka Marczak. Treatment of Multiple Myeloma and the Role of Melphalan in the Era of Modern Therapies—Current Research and Clinical Approaches. *Journal of Clinical Medicine*, 10(9):1841, 4 2021.

[81] Arnab Mukherjee and Wilbee D. Sasikala. Drug–DNA Intercalation. In *Advances in Protein Chemistry and Structural Biology*, volume 92, pages 1–62. Academic Press Inc., 1 2013.

[82] Lynnette R. Ferguson and William A. Denny. Genotoxicity of non-covalent interactions: DNA intercalators. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 623(1-2):14–23, 10 2007.

[83] Kuo-Hsiang Hsu, Bo-Han Su, Yi-Shu Tu, Olivia A. Lin, and Yufeng J. Tseng. Mutagenicity in a Molecule: Identification of Core Structural Features of Mutagenicity Using a Scaffold Analysis. *PLOS ONE*, 11(2):e0148900, 2 2016.

[84] L.S. Lerman. Structural considerations in the interaction of DNA and acridines. *Journal of Molecular Biology*, 3(1):18–IN14, 2 1961.

[85] Gerard Moloney, David Kelly, and P. Mack. Synthesis of Acridine-based DNA Bis-intercalating Agents. *Molecules*, 6(3):230–243, 2 2001.

[86] Hiroyasu Ichikawa, Ronald R. Navarro, Yosuke Iimura, and Kenji Tatsumi. Nature of bioavailability of DNA-intercalated polycyclic aromatic hydrocarbons to Sphingomonas sp. *Chemosphere*, 80(8):866–871, 8 2010.

[87] Stephan Laib, Alexander Krieg, Pascal Häfliger, and Nikos Agorastos. DNA-intercalation on pyrene modified surface coatings. *Chemical Communications*, (44):5566, 11 2005.

[88] Yining Xiong, Junsheng Li, Guoxia Huang, Liujuan Yan, and Ji Ma. Interacting mechanism of benzo(a)pyrene with free DNA in vitro. *International Journal of Biological Macromolecules*, 167:854–861, 1 2021.

[89] Tridib Mahata, Jeet Chakraborty, Ajay Kanungo, Dipendu Patra, Gautam Basu, and Sanjay Dutta. Intercalator-Induced DNA Superstructure Formation: Doxorubicin and a Synthetic Quinoxaline Derivative. *Biochemistry*, 57(38):5557–5563, 9 2018.

[90] Us EPA ORD NCEA Integrated Risk Information System. Toxicological Review of Quinoline. 2001.

[91] Mark J. Hoenerhoff, Molly Boyle, Sheroy Minocherhomji, and Arun R. Pandiri. Carcinogenesis: Mechanisms and Evaluation. In *Haschek and Rousseaux's Handbook of Toxicologic Pathology*, pages 205–254. Elsevier, 1 2022.

[92] Robert Baan, Yann Grosse, Kurt Straif, Béatrice Secretan, Fatiha El Ghissassi, Véronique Bouvard, Lamia Benbrahim-Tallaa, Neela Guha, Crystal Freeman, Laurent Galichet, and Vincent Cogliano. A review of human carcinogens—Part F: Chemical agents and related occupations. *The Lancet Oncology*, 10(12):1143–1144, 12 2009.

[93] Sandeep Tyagi, Saurabh Sharma, Paras Gupta, ArminderSingh Saini, and Chaitnya Kaushal. The peroxisome proliferator-activated receptor: A family of nuclear receptors role in various diseases. *Journal of Advanced Pharmaceutical Technology & Research*, 2(4):236, 2011.

[94] Christian Grommes, Gary E. Landreth, Magdalena Sastre, Martina Beck, Douglas L. Feinstein, Andreas H. Jacobs, Uwe Schlegel, and Michael T. Heneka. Inhibition of in Vivo Glioma Growth and Invasion by Peroxisome Proliferator-Activated Receptor γ Agonist Treatment. *Molecular Pharmacology*, 70(5):1524–1533, 11 2006.

[95] Keiko Fukumoto, Yoshihisa Yano, Nantiga Virgona, Hiromi Hagiwara, Hiromi Sato, Hironobu Senba, Kazuyuki Suzuki, Ryuji Asano, Kazuhiko Yamada, and Tomohiro

Yano. Peroxisome proliferator-activated receptor δ as a molecular target to regulate lung cancer cell growth. *FEBS Letters*, 579(17):3829–3836, 7 2005.

[96] ShouWei Han, Jeffrey D. Ritzenthaler, XiaoJuan Sun, Ying Zheng, and Jesse Roman. Activation of Peroxisome Proliferator–Activated Receptor β/δ Induces Lung Cancer Growth via Peroxisome Proliferator–Activated Receptor Coactivator γ-1α. *American Journal of Respiratory Cell and Molecular Biology*, 40(3):325–331, 3 2009.

[97] Sean R. Pyper, Navin Viswakarma, Songtao Yu, and Janardan K. Reddy. PPARα: Energy Combustion, Hypolipidemia, Inflammation and Cancer. *Nuclear Receptor Signaling*, 8(1):nrs.08002, 1 2010.

[98] D. A. Goodenough and D. L. Paul. Gap Junctions. *Cold Spring Harbor Perspectives in Biology*, 1(1):a002576–a002576, 7 2009.

[99] Wenjing Liu, Yujia Cui, Jieya Wei, Jianxun Sun, Liwei Zheng, and Jing Xie. Gap junction-mediated cell-to-cell communication in oral development and oral diseases: a concise review of research progress. *International Journal of Oral Science*, 12(1):17, 12 2020.

[100] James E Trosko. Cell-cell communication in carcinogenesis. *Frontiers in Bioscience*, 3(4):A275, 2 1998.

[101] Marc Mesnil, Sophie Crespin, José-Luis Avanzo, and Maria-Lucia Zaidan-Dagli. Defective gap junctional intercellular communication in the carcinogenic process. *Biochimica et Biophysica Acta (BBA) - Biomembranes*, 1719(1-2):125–145, 12 2005.

[102] Takanori Harada, Makio Takeda, Sayuri Kojima, and Naruto Tomiyama. Toxicity and

Carcinogenicity of Dichlorodiphenyltrichloroethane (DDT). *Toxicological Research*, 32(1):21–33, 1 2016.

[103] Randall J. Ruch, Ronny Fransson, Sten Flodstrom, Lars Warngard, and James E. Klaunig. Inhibition of hepatocyte gap junctional intercellular communication by endosulfan, chlordane and heptachlor. *Carcinogenesis*, 11(7):1097–1101, 7 1990.

[104] Agency for Toxic Substances and Disease Registry. Toxicological Profile for Chlordane. Technical report, 2018.

[105] Zefferino, Piccoli, Gioia, Capitanio, and Conese. Gap Junction Intercellular Communication in the Carcinogenesis Hallmarks: Is This a Phenomenon or Epiphenomenon? *Cells*, 8(8):896, 8 2019.

[106] S Hiller-Sturmhöfel and A Bartke. The endocrine system: an overview. *Alcohol health and research world*, 22(3):153–64, 1998.

[107] Åke Bergman, Jerrold J. Heindel, Tim Kasten, Karen A. Kidd, Susan Jobling, Maria Neira, R. Thomas Zoeller, Georg Becher, Poul Bjerregaard, Riana Bornman, Ingvar Brandt, Andreas Kortenkamp, Derek Muir, Marie-Noël Brune Drisse, Roseline Ochieng, Niels E. Skakkebaek, Agneta Sundén Byléhn, Taisen Iguchi, Jorma Toppari, and Tracey J. Woodruff. The Impact of Endocrine Disruption: A Consensus Statement on the State of the Science. *Environmental Health Perspectives*, 121(4), 4 2013.

[108] Linda G Kahn, Claire Philippat, Shoji F Nakayama, Rémy Slama, and Leonardo Trasande. Endocrine-disrupting chemicals: implications for human health. *The Lancet Diabetes & Endocrinology*, 8(8):703–718, 8 2020.

[109] Iram Ashaq Kawa, Akbar masood, Qudsia Fatima, Shahnaz Ahmad Mir, Humira Jeelani, Saika Manzoor, and Fouzia Rashid. Endocrine disrupting chemical Bisphenol A and its potential effects on female health. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, 15(3):803–811, 5 2021.

[110] Mao-Hsun Lin, Chun-Ying Lee, Yun-Shiuan Chuang, and Chia-Lung Shih. Exposure to bisphenol A associated with multiple health-related outcomes in humans: An umbrella review of systematic reviews with meta-analyses. *Environmental Research*, 237:116900, 11 2023.

[111] Jéssica C. P. Petrine and Bruno Del Bianco-Borges. The influence of phytoestrogens on different physiological and pathological processes: An overview. *Phytotherapy Research*, 35(1):180–197, 1 2021.

[112] Tyler J. Thompson, Martin A. Briggs, Patrick J. Phillips, Vicki S. Blazer, Kelly L. Smalling, Dana W. Kolpin, and Tyler Wagner. Groundwater discharges as a source of phytoestrogens and other agriculturally derived contaminants to streams. *Science of The Total Environment*, 755:142873, 2 2021.

[113] N. M. Grindler, L. Vanderlinden, R. Karthikraj, K. Kannan, S. Teal, A. J. Polotsky, T. L. Powell, I. V. Yang, and T. Jansson. Exposure to Phthalate, an Endocrine Disrupting Chemical, Alters the First Trimester Placental Methylome and Transcriptome in Women. *Scientific Reports*, 8(1):6086, 4 2018.

[114] Fuchen Xie, Xuedong Chen, Shiqi Weng, Tianxinyu Xia, Xinyi Sun, Tao Luo, and Peng Li. Effects of two environmental endocrine disruptors di-n-butyl phthalate (DBP) and mono-n-butyl phthalate (MBP) on human sperm functions in vitro. *Reproductive Toxicology*, 83:1–7, 1 2019.

[115] Yiyu Qian, Hailing Shao, Xinxin Ying, Wenle Huang, and Ying Hua. The Endocrine Disruption of Prenatal Phthalate Exposure in Mother and Offspring. *Frontiers in Public Health*, 8:548585, 8 2020.

[116] Ayami Matsushima. A Novel Action of Endocrine-Disrupting Chemicals on Wildlife; DDT and Its Derivatives Have Remained in the Environment. *International Journal of Molecular Sciences 2018, Vol. 19, Page 1377*, 19(5):1377, 5 2018.

[117] Allan R Woodward, H Franklin Percival, Michael L Jennings, and Clinton T Moore. Low Clutch Viability of American Alligators on Lake Apopka. *Florida Scientist*, 56(1):52–63, 1993.

[118] L J Guillette, T S Gross, G R Masson, J M Matter, H F Percival, and A R Woodward. Developmental abnormalities of the gonad and abnormal sex hormone concentrations in juvenile alligators from contaminated and control lakes in Florida. *Environmental Health Perspectives*, 102(8):680–688, 8 1994.

[119] Louis J. Guillette Jr., Daniel B. Pickford, D.Andrew Crain, Andrew A. Rooney, and H.Franklin Percival. Reduction in Penis Size and Plasma Testosterone Concentrations in Juvenile Alligators Living in a Contaminated Environment. *General and Comparative Endocrinology*, 101(1):32–42, 1 1996.

[120] Matthew R. Milnes and Louis J. Guillette. Alligator Tales: New Lessons about Environmental Contaminants from a Sentinel Species. *BioScience*, 58(11):1027–1036, 12 2008.

[121] Aleksandra Buha Djordjevic, Evica Antonijevic, Marijana Curcic, Vesna Milovanovic, and Biljana Antonijevic. Endocrine-disrupting mechanisms of polychlorinated biphenyls. *Current Opinion in Toxicology*, 19:42–49, 2 2020.

[122] Shouhua Zhang, Kuai Chen, Weiming Li, Yong Chai, Jian Zhu, Bingfeng Chu, Nuoya Li, Jinlong Yan, Shenglai Zhang, and Yipeng Yang. Varied thyroid disrupting effects of perfluorooctanoic acid (PFOA) and its novel alternatives hexafluoropropylene-oxide-dimer-acid (GenX) and ammonium 4,8-dioxa-3H-perfluorononanoate (ADONA) in vitro. *Environment International*, 156:106745, 11 2021.

[123] Yatao Du, Chaojie Chen, Guangdi Zhou, Zhenzhen Cai, Qiuhong Man, Baolin Liu, and Weiye Charles Wang. Perfluorooctanoic acid disrupts thyroid-specific genes expression and regulation via the TSH-TSHR signaling pathway in thyroid cells. *Environmental Research*, 239:117372, 12 2023.

[124] Adrian Reuben, Holly Tillman, Robert J. Fontana, Timothy Davern, Brendan McGuire, R. Todd Stravitz, Valerie Durkalski, Anne M. Larson, Iris Liou, Oren Fix, Michael Schilsky, Timothy McCashland, J. Eileen Hay, Natalie Murray, Obaid S. Shaikh, Daniel Ganger, Atif Zaman, Steven B. Han, Raymond T. Chung, Alastair Smith, Robert Brown, Jeffrey Crippin, M. Edwyn Harrison, David Koch, Santiago Munoz, K. Rajender Reddy, Lorenzo Rossaro, Raj Satyanarayana, Tarek Hassanein, A. James Hanje, Jody Olson, Ram Subramanian, Constantine Karvellas, Bilal Hameed, Averell H. Sherker, Patricia Robuck, and William M. Lee. Outcomes in Adults With Acute Liver Failure Between 1998 and 2013. *Annals of Internal Medicine*, 164(11):724, 6 2016.

[125] M. C. Donnelly, J. S. Davidson, K. Martin, A. Baird, P. C. Hayes, and K. J. Simpson. Acute liver failure in Scotland: changes in aetiology and outcomes over time (the Scottish Look-Back Study). *Alimentary Pharmacology & Therapeutics*, 45(6):833–843, 3 2017.

[126] Raul J. Andrade, Naga Chalasani, Einar S. Björnsson, Ayako Suzuki, Gerd A. Kullak-

Ublick, Paul B. Watkins, Harshad Devarbhavi, Michael Merz, M. Isabel Lucena, Neil Kaplowitz, and Guruprasad P. Aithal. Drug-induced liver injury. *Nature Reviews Disease Primers*, 5(1):58, 8 2019.

[127] Ki Tae Suk, Dong Joon Kim, Chang Hoon Kim, Seung Ha Park, Jai Hoon Yoon, Yeon Soo Kim, Gwang Ho Baik, Jin Bong Kim, Young Oh Kweon, Byung Ik Kim, Seok Hyun Kim, In Hee Kim, Ju Hyun Kim, Soon Woo Nam, Yong Han Paik, Jeong Ill Suh, Joo Hyun Sohn, Byung Min Ahn, Soon Ho Um, Heon Ju Lee, Mong Cho, Myoung Kuk Jang, Sung Kyu Choi, Seong Gyu Hwang, Ho Taik Sung, Jong Young Choi, and Kwang Hyub Han. A prospective nationwide study of drug-induced liver injury in Korea. *The American journal of gastroenterology*, 107(9):1380–7, 9 2012.

[128] Toby J. Athersuch, Daniel J. Antoine, Alan R. Boobis, Muireann Coen, Ann K. Daly, Lucia Possamai, Jeremy K. Nicholson, and Ian D. Wilson. Paracetamol metabolism, hepatotoxicity, biomarkers and therapeutic interventions: a perspective. *Toxicology Research*, 7(3):347–357, 5 2018.

[129] Anke Kretz-Rommel and Urs A. Boelsterli. Cytotoxic activity of T cells and non-T cells from diclofenac-immunized mice against cultured syngeneic hepatocytes exposed to diclofenac. *Hepatology*, 22(1):213–222, 7 1995.

[130] Ann K. Daly, Guruprasad P. Aithal, Julian B.S. Leathart, Richard A. Swainsbury, Tarana Singh Dang, and Christopher P. Day. Genetic Susceptibility to Diclofenac-Induced Hepatotoxicity: Contribution of UGT2B7, CYP2C8, and ABCC2 Genotypes. *Gastroenterology*, 132(1):272–281, 1 2007.

[131] Kan He, Rasmy E. Talaat, William F. Pool, Michael D. Reily, Jessica E. Reed, Alexander J. Bridges, and Thomas F. Woolf. Metabolic Activation of Troglitazone: Identifica-

tion of a Reactive Metabolite and Mechanisms Involved. *Drug Metabolism and Disposition*, 32(6):639–646, 6 2004.

[132] Martyn T. Smith. Mechanisms of Troglitazone Hepatotoxicity. *Chemical Research in Toxicology*, 16(6):679–687, 6 2003.

[133] Joel Berger and David E. Moller. The Mechanisms of Action of PPARs. *Annual Review of Medicine*, 53(1):409–435, 2 2002.

[134] A.K. Hihi, L. Michalik, and W. Wahli. PPARs: transcriptional effectors of fatty acids and their derivatives. *Cellular and Molecular Life Sciences*, 59(5):790–798, 5 2002.

[135] Christoph Funk, Christiane Ponelle, Gerd Scheuermann, and Michael Pantze. Cholestatic Potential of Troglitazone as a Possible Factor Contributing to Troglitazone-Induced Hepatotoxicity: In Vivo and in Vitro Interaction at the Canalicular Bile Salt Export Pump (Bsep) in the Rat. *Molecular Pharmacology*, 59(3):627–635, 3 2001.

[136] Christoph Funk, Michael Pantze, Linda Jehle, Christiane Ponelle, Gerd Scheuermann, Mirjana Lazendic, and Rodolfo Gasser. Troglitazone-induced intrahepatic cholestasis by an interference with the hepatobiliary export of bile acids in male and female rats. Correlation with the gender difference in troglitazone sulfate formation and the inhibition of the canalicular bile salt export pump (Bsep) by troglitazone and troglitazone sulfate. *Toxicology*, 167(1):83–98, 10 2001.

[137] K Fattinger, C Funk, M Pantze, C Weber, J Reichen, B Stieger, and P J Meier. The endothelin antagonist bosentan inhibits the canalicular bile salt export pump: a potential mechanism for hepatic adverse reactions. *Clinical pharmacology and therapeutics*, 69(4):223–31, 4 2001.

[138] Yuji Mano, Takashi Usui, and Hidetaka Kamimura. Effects of bosentan, an endothelin receptor antagonist, on bile salt export pump and multidrug resistance–associated protein 2. *Biopharmaceutics & Drug Disposition*, 28(1):13–18, 1 2007.

[139] P B Watkins. Drug Safety Sciences and the Bottleneck in Drug Development. *Clinical Pharmacology & Therapeutics*, 89(6):788–790, 6 2011.

[140] Bryan H. Norman. Drug Induced Liver Injury (DILI). Mechanisms and Medicinal Chemistry Avoidance/Mitigation Strategies. *Journal of Medicinal Chemistry*, 63(20):11397–11419, 10 2020.

[141] Timothy E Johnson, Xiaohua Zhang, Kimberly B Bleicher, Gary Dysart, Amy F Loughlin, William H Schaefer, and Diane R Umbenhauer. Statins induce apoptosis in rat and human myotube cultures by inhibiting protein geranylgeranylation but not ubiquinone. *Toxicology and applied pharmacology*, 200(3):237–50, 11 2004.

[142] Frederick Peter Guengerich. Mechanisms of Drug Toxicity and Relevance to Pharmaceutical Development. *Drug Metabolism and Pharmacokinetics*, 26(1):3–14, 1 2011.

[143] Thomas Force and Kyle L. Kolaja. Cardiotoxicity of kinase inhibitors: the prediction and translation of preclinical models to clinical outcomes. *Nature Reviews Drug Discovery*, 10(2):111–126, 2 2011.

[144] Susan Klaeger, Stephanie Heinzlmeir, Mathias Wilhelm, Harald Polzer, Binje Vick, Paul-Albert Koenig, Maria Reinecke, Benjamin Ruprecht, Svenja Petzoldt, Chen Meng, Jana Zecha, Katrin Reiter, Huichao Qiao, Dominic Helm, Heiner Koch, Melanie Schoof, Giulia Canevari, Elena Casale, Stefania Re Depaolini, Annette Feuchtinger, Zhixiang Wu, Tobias Schmidt, Lars Rueckert, Wilhelm Becker, Jan Huenges, Anne-Kathrin Garz, Bjoern-Oliver Gohlke, Daniel Paul Zolg, Gian Kayser, Tonu Vooder,

Robert Preissner, Hannes Hahne, Neeme Tõnisson, Karl Kramer, Katharina Götze, Florian Bassermann, Judith Schlegl, Hans-Christian Ehrlich, Stephan Aiche, Axel Walch, Philipp A. Greif, Sabine Schneider, Eduard Rudolf Felder, Juergen Ruland, Guillaume Médard, Irmela Jeremias, Karsten Spiekermann, and Bernhard Kuster. The target landscape of clinical kinase drugs. *Science*, 358(6367), 12 2017.

[145] Domenic A. Sica. Pharmacokinetics and Pharmacodynamics of Mineralocorticoid Blocking Agents and their Effects on Potassium Homeostasis. *Heart Failure Reviews*, 10(1):23–29, 1 2005.

[146] C. Bonne and J.P. Raynaud. Mode of spironolactone anti-androgenic action: Inhibition of androstanolone binding to rat prostate androgen receptor. *Molecular and Cellular Endocrinology*, 2(1):59–67, 12 1974.

[147] Ronald L. Young, Joseph W. Goldzieher, and Karen Elkind-Hirsch. The endocrine effects of spironolactone used as an antiandrogen. *Fertility and Sterility*, 48(2):223–228, 8 1987.

[148] Deepani Rathnayake and Rodney Sinclair. Innovative Use of Spironolactone as an Antiandrogen in the Treatment of Female Pattern Hair Loss. *Dermatologic Clinics*, 28(3):611–618, 7 2010.

[149] Ryan D. Gabbard, Robert R. Hoopes, and Michael G. Kemp. Spironolactone and XPB: An Old Drug with a New Molecular Target. *Biomolecules*, 10(5):756, 5 2020.

[150] Ann Lin, Christopher J. Giuliano, Ann Palladino, Kristen M. John, Connor Abramowicz, Monet Lou Yuan, Erin L. Sausville, Devon A. Lukow, Luwei Liu, Alexander R. Chait, Zachary C. Galluzzo, Clara Tucker, and Jason M. Sheltzer. Off-target tox-

icity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine*, 11(509):8412, 9 2019.

[151] Fritz Haber. Zur Geschichte des Gaskrieges. *Fünf Vorträge aus den Jahren 1920–1923*, pages 76–92, 1924.

[152] David W. Gaylor. The use of Haber's Law in standard setting and risk assessment. *Toxicology*, 149(1):17–19, 8 2000.

[153] Frederick J Miller, Paul M Schlosser, and Derek B Janszen. Haber's rule: a special case in a family of curves relating concentration and duration of exposure to a fixed level of response for a given endpoint. *Toxicology*, 149(1):21–34, 8 2000.

[154] Arwa B. Raies and Vladimir B. Bajic. In silico toxicology: computational methods for the prediction of chemical toxicity. *WIREs Computational Molecular Science*, 6(2):147–172, 3 2016.

[155] Xiaomei Zhuang and Chuang Lu. PBPK modeling and simulation in drug research and development. *Acta Pharmaceutica Sinica B*, 6(5):430–440, 9 2016.

[156] L Kuepfer, C Niederalt, T Wendl, J-F Schlender, S Willmann, J Lippert, M Block, T Eissing, and D Teutonico. Applied Concepts in PBPK Modeling: How to Build a PBPK/PD Model. *CPT: Pharmacometrics & Systems Pharmacology*, 5(10):516–531, 10 2016.

[157] Sheila Annie Peters. *Physiologically based pharmacokinetic (PBPK) modeling and simulations: principles, methods, and applications in the pharmaceutical industry*. John Wiley & Sons, 2021.

[158] Simulations Plus | Modeling & Simulation Software. https://www.simulations-plus.com/.

[159] Certara Simcyp™ PBPK Simulator | Predicting Drug Performance. https://www.certara.com/software/simcyp-pbpk/.

[160] Jeffrey W Fisher, Jeffery M Gearhart, and Zhoumeng Lin. *Physiologically Based Pharmacokinetic (PBPK) Modeling: Methods and Applications in Toxicology and Risk Assessment*. Academic Press, 2020.

[161] Miyuki Breen, Caroline L Ring, Anna Kreutz, Michael-Rock Goldsmith, and John F Wambaugh. High-throughput PBTK models for in vitro to in vivo extrapolation. *Expert Opinion on Drug Metabolism & Toxicology*, 17(8):903–921, 8 2021.

[162] Emilio Benfenati and Giuseppina Gini. Computational predictive programs (expert systems) in toxicology. *Toxicology*, 119(3):213–225, 5 1997.

[163] Chiara Milan, Onofrio Schifanella, Alessandra Roncaglioni, and Emilio Benfenati. Comparison and Possible Use of In Silico Tools for Carcinogenicity Within REACH Legislation. *Journal of Environmental Science and Health, Part C*, 29(4):300–323, 10 2011.

[164] John Ashby and Raymond W. Tennant. Chemical structure, Salmonella mutagenicity and extent of carcinogenicity as indicators of genotoxic carcinogenesis among 222 chemicals tested in rodents by the U.S. NCI/NTP. *Mutation Research/Genetic Toxicology*, 204(1):17–115, 1 1988.

[165] Romualdo Benigni and Cecilia Bossa. Structure alerts for carcinogenicity, and the

Salmonella assay system: a novel insight through the chemical relational databases technology. *Mutation research*, 659(3):248–61, 9 2008.

[166] Romualdo Benigni, Cecilia Bossa, and Olga Tcheremenskaia. Nongenotoxic Carcinogenicity of Chemicals: Mechanisms of Action and Early Recognition through a New Set of Structural Alerts. *Chemical Reviews*, 113(5):2940–2957, 5 2013.

[167] Toxtree – Toxtree - Toxic Hazard Estimation by decision tree approach. https://toxtree.sourceforge.net/.

[168] Andreas Maunz, Martin Gütlein, Micha Rautenberg, David Vorgrimmler, Denis Gebele, and Christoph Helma. lazar: a modular predictive toxicology framework. *Frontiers in Pharmacology*, 4:40286, 4 2013.

[169] Derek Nexus – Achieving High Accuracy With High Coverage [an Infographic] | Lhasa Limited. https://www.lhasalimited.org/derek-nexus-achieving-high-accuracy-with-high-coverage-an-infographic/.

[170] N. Greene, P. N. Judson, J. J. Langowski, and C. A. Marchant. Knowledge-Based Expert Systems for Toxicity and Metabolism Prediction: DEREK, StAR and METEOR. *SAR and QSAR in Environmental Research*, 10(2-3):299–314, 7 1999.

[171] Alexandr Mikhaylovich Butlerov. Einiges über die chemische Structur der Körper. *Zeitschrift für Chemie*, 4:549–560, 1861.

[172] Frank F. Kluge and David F. Larder. A. M. Butlerov. On the chemical structure of substances. *Journal of Chemical Education*, 48(5):289, 5 1971.

[173] S. Dimitrov and O. Mekenyan. An Introduction to Read-Across for the Prediction of the Effects of Chemicals. In *In Silico Toxicology*, pages 372–384. The Royal Society of Chemistry, 10 2010.

[174] Andrew R. Leach and Valerie J. Gillet. Representation And Manipulation Of 2D Molecular Structures. In *An Introduction To Chemoinformatics*, pages 1–25. Springer Netherlands, Dordrecht, 2007.

[175] T. T. Tanimoto. *An elementary mathematical theory of classification and prediction*. International Business Machines Corporation, New York,, 1958.

[176] Lee R. Dice. Measures of the Amount of Ecologic Association Between Species. *Ecology*, 26(3):297–302, 7 1945.

[177] Amos Tversky. Features of similarity. *Psychological Review*, 84(4):327–352, 7 1977.

[178] Andrew R. Leach and Valerie J. Gillet. Representation And Manipulation Of 3D Molecular Structures. In *An Introduction To Chemoinformatics*, pages 27–52. Springer Netherlands, Dordrecht, 2007.

[179] H. L. Morgan. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation*, 5(2):107–113, 5 1965.

[180] Daylight. https://www.daylight.com/.

[181] David Rogers and Mathew Hahn. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 5 2010.

[182] Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):43, 12 2020.

[183] Shifa Zhong and Xiaohong Guan. Count-Based Morgan Fingerprint: A More Efficient and Interpretable Molecular Representation in Developing Machine Learning-Based Predictive Regression Models for Water Contaminants' Activities and Properties. *Environmental Science & Technology*, 7 2023.

[184] Thanh-Hoang Nguyen-Vo, Quang H. Trinh, Loc Nguyen, Phuong-Uyen Nguyen-Hoang, Thien-Ngan Nguyen, Dung T. Nguyen, Binh P. Nguyen, and Ly Le. iCYP-MFE: Identifying Human Cytochrome P450 Inhibitors Using Multitask Learning and Molecular Fingerprint-Embedded Encoding. *Journal of Chemical Information and Modeling*, 62(21):5059–5068, 11 2022.

[185] Liangxu Xie, Lei Xu, Ren Kong, Shan Chang, and Xiaojun Xu. Improvement of Prediction Performance With Conjoint Molecular Fingerprint in Deep Learning. *Frontiers in Pharmacology*, 11:606668, 12 2020.

[186] Corwin Hansch, Peyton P. Maloney, Toshio Fujita, and Robert M. Muir. Correlation of Biological Activity of Phenoxyacetic Acids with Hammett Substituent Constants and Partition Coefficients. *Nature*, 194(4824):178–180, 4 1962.

[187] Dmitry A. Konovalov, Lyndon E. Llewellyn, Yvan Vander Heyden, and Danny Coomans. Robust Cross-Validation of Linear Regression QSAR Models. *Journal of Chemical Information and Modeling*, 48(10):2081–2094, 10 2008.

[188] Antreas Afantitis, Georgia Melagraki, Haralambos Sarimveis, Panayiotis A. Koutentis, John Markopoulos, and Olga Igglessi-Markopoulou. A novel simple QSAR model

for the prediction of anti-HIV activity using multiple linear regression analysis. *Molecular Diversity*, 10(3):405–414, 8 2006.

[189] David T. Stanton. QSAR and QSPR Model Interpretation Using Partial Least Squares (PLS) Analysis. *Current Computer Aided-Drug Design*, 8(2):107–127, 4 2012.

[190] Hu Mei, Yuan Zhou, Guizhao Liang, and Zhiliang Li. Support vector machine applied in QSAR modelling. *Chinese Science Bulletin*, 50(20):2291–2296, 10 2005.

[191] Pavel G. Polishchuk, Eugene N. Muratov, Anatoly G. Artemenko, Oleg G. Kolumbin, Nail N. Muratov, and Victor E. Kuz'min. Application of random forest approach to QSAR prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*, 49(11):2481–2488, 11 2009.

[192] Subhash Ajmani, Kamalakar Jadhav, and Sudhir A. Kulkarni. Three-Dimensional QSAR Using the k-Nearest Neighbor Method and Its Interpretation. *Journal of Chemical Information and Modeling*, 46(1):24–31, 1 2006.

[193] Mohammad Goodarzi, Bieke Dejaegher, and Yvan Vander Heyden. Feature Selection Methods in QSAR Studies. *Journal of AOAC INTERNATIONAL*, 95(3):636–651, 5 2012.

[194] Martin Eklund, Ulf Norinder, Scott Boyer, and Lars Carlsson. Choosing Feature Selection and Learning Algorithms in QSAR. *Journal of Chemical Information and Modeling*, 54(3):837–843, 3 2014.

[195] Pathan Mohsin Khan and Kunal Roy. Current approaches for choosing feature selection and learning algorithms in quantitative structure–activity relationships (QSAR). *Expert Opinion on Drug Discovery*, 13(12):1075–1089, 12 2018.

[196] Stefano A. Bini. Artificial Intelligence, Machine Learning, Deep Learning, and Cognitive Computing: What Do These Terms Mean and How Will They Impact Health Care? *The Journal of Arthroplasty*, 33(8):2358–2361, 8 2018.

[197] Garrett M. Morris and Marguerita Lim-Wilby. Molecular Docking. In *Methods in Molecular Biology*, volume 443, pages 365–382. Humana Press, 2008.

[198] Alexandre MJJ Bonvin. Flexible protein–protein docking. *Current Opinion in Structural Biology*, 16(2):194–200, 4 2006.

[199] Erney Ramírez-Aportela, José Ramón López-Blanco, and Pablo Chacón. FRODOCK 2.0: fast protein-protein docking server. *Bioinformatics (Oxford, England)*, 32(15):2386–8, 8 2016.

[200] Dima Kozakov, David R Hall, Bing Xia, Kathryn A Porter, Dzmitry Padhorny, Christine Yueh, Dmitri Beglov, and Sandor Vajda. The ClusPro web server for protein–protein docking. *Nature Protocols*, 12(2):255–278, 2 2017.

[201] Yumeng Yan, Huanyu Tao, Jiahua He, and Sheng-You Huang. The HDOCK server for integrated protein–protein docking. *Nature Protocols*, 15(5):1829–1852, 5 2020.

[202] Sharon Sunny and P. B. Jayaraj. Protein–Protein Docking: Past, Present, and Future. *The Protein Journal*, 41(1):1–26, 2 2022.

[203] Sérgio Filipe Sousa, Pedro Alexandrino Fernandes, and Maria João Ramos. Protein–ligand docking: Current status and future challenges. *Proteins: Structure, Function, and Bioinformatics*, 65(1):15–26, 10 2006.

[204] Inbal Halperin, Buyong Ma, Haim Wolfson, and Ruth Nussinov. Principles of docking: An overview of search algorithms and a guide to scoring functions. *Proteins: Structure, Function, and Bioinformatics*, 47(4):409–443, 6 2002.

[205] Garrett M. Morris, Ruth Huey, William Lindstrom, Michel F. Sanner, Richard K. Belew, David S. Goodsell, and Arthur J. Olson. AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry*, 30(16):2785–2791, 12 2009.

[206] Marcel L. Verdonk, Jason C. Cole, Michael J. Hartshorn, Christopher W. Murray, and Richard D. Taylor. Improved protein–ligand docking using GOLD. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623, 9 2003.

[207] Ruben Abagyan, Maxim Totrov, and Dmitry Kuznetsov. ICM—A new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry*, 15(5):488–506, 5 1994.

[208] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 8 2013.

[209] David S. Goodsell and Arthur J. Olson. Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function, and Bioinformatics*, 8(3):195–202, 1 1990.

[210] Xingxing Zhou, Ming Ling, Qingde Lin, Shidi Tang, Jiansheng Wu, and Haifeng Hu. Effectiveness Analysis of Multiple Initial States Simulated Annealing Algorithm, A Case

Study on the Molecular Docking Tool AutoDock Vina. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pages 1–12, 2023.

[211] Jin Li, Ailing Fu, and Le Zhang. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdisciplinary Sciences: Computational Life Sciences*, 11(2):320–328, 6 2019.

[212] T J Ewing, S Makino, A G Skillman, and I D Kuntz. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of computer-aided molecular design*, 15(5):411–28, 5 2001.

[213] C.M. Venkatachalam, X. Jiang, T. Oldfield, and M. Waldman. LigandFit: a novel method for the shape-directed rapid docking of ligands to protein active sites. *Journal of Molecular Graphics and Modelling*, 21(4):289–307, 1 2003.

[214] Hao Fan, Dina Schneidman-Duhovny, John J. Irwin, Guangqiang Dong, Brian K. Shoichet, and Andrej Sali. Statistical Potential for Modeling and Ranking of Protein–Ligand Interactions. *Journal of Chemical Information and Modeling*, 51(12):3078–3092, 12 2011.

[215] Zhong-Ru Xie and Ming-Jing Hwang. An interaction-motif-based scoring function for protein-ligand docking. *BMC Bioinformatics*, 11(1):298, 12 2010.

[216] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 3 2004.

[217] Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1):43, 12 2021.

[218] Bernd Kramer, Matthias Rarey, and Thomas Lengauer. Evaluation of the FLEXX incremental construction algorithm for protein-ligand docking. *Proteins: Structure, Function, and Genetics*, 37(2):228–241, 11 1999.

[219] Maria Kontoyianni, Laura M. McClellan, and Glenn S. Sokol. Evaluation of Docking Performance: Comparative Data on Docking Algorithms. *Journal of Medicinal Chemistry*, 47(3):558–565, 1 2004.

[220] Jason C. Cole, Christopher W. Murray, J. Willem M. Nissink, Richard D. Taylor, and Robin Taylor. Comparing protein–ligand docking programs is difficult. *Proteins: Structure, Function, and Bioinformatics*, 60(3):325–332, 8 2005.

[221] Ruben A. Abagyan and Maxim M. Totrov. Contact area difference (CAD): a robust measure to evaluate accuracy of protein models. *Journal of Molecular Biology*, 268(3):678–685, 5 1997.

[222] Romano T. Kroemer, Anna Vulpetti, Joseph J. McDonald, Douglas C. Rohrer, Jean Yves Trosset, Fabrizio Giordanetto, Simona Cotesta, Colin McMartin, Mats Kihlén, and Pieter F.W. Stouten. Assessment of docking poses: Interactions-based accuracy classification (IBAC) versus crystal structure deviations. *Journal of Chemical Information and Computer Sciences*, 44(3):871–881, 5 2004.

[223] Dina Schneidman-Duhovny, Yuval Inbar, Ruth Nussinov, and Haim J. Wolfson. Geometry-based flexible and symmetric protein docking. *Proteins: Structure, Function, and Bioinformatics*, 60(2):224–231, 8 2005.

[224] Feng Ding, Shuangye Yin, and Nikolay V. Dokholyan. Rapid Flexible Docking Using a Stochastic Rotamer Library of Ligands. *Journal of Chemical Information and Modeling*, 50(9):1623–1632, 9 2010.

[225] Martin Smieško. DOLINA – Docking Based on a Local Induced-Fit Algorithm: Application toward Small-Molecule Binding to Nuclear Receptors. *Journal of Chemical Information and Modeling*, 53(6):1415–1423, 6 2013.

[226] Samuel DeLuca, Karen Khar, and Jens Meiler. Fully Flexible Docking of Medium Sized Ligand Libraries with RosettaLigand. *PLOS ONE*, 10(7):e0132508, 7 2015.

[227] Nabil F. Faruk, Xiangda Peng, Karl F. Freed, Benoît Roux, and Tobin R. Sosnick. Challenges and Advantages of Accounting for Backbone Flexibility in Prediction of Protein-Protein Complexes. *Journal of Chemical Theory and Computation*, 18(3):2016–2032, 3 2022.

[228] P. A. Greenidge, C. Kramer, J.-C. Mozziconacci, and W. Sherman. Improving Docking Results via Reranking of Ensembles of Ligand Poses in Multiple X-ray Protein Conformations with MM-GBSA. *Journal of Chemical Information and Modeling*, 54(10):2697–2717, 10 2014.

[229] Jason B. Cross, David C. Thompson, Brajesh K. Rai, J. Christian Baber, Kristi Yi Fan, Yongbo Hu, and Christine Humblet. Comparison of Several Molecular Docking Programs: Pose Prediction and Virtual Screening Accuracy. *Journal of Chemical Information and Modeling*, 49(6):1455–1474, 6 2009.

[230] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set

of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, 5 2016.

[231] Abdulmujeeb T. Onawole, Temitope U. Kolapo, Kazeem O. Sulaiman, and Rukayat O. Adegoke. Structure based virtual screening of the Ebola virus trimeric glycoprotein using consensus scoring. *Computational Biology and Chemistry*, 72:170–180, 2 2018.

[232] El Hassen Mokrani, Abderrahmane Bensegueni, Ludovic Chaput, Claire Beauvineau, Hanane Djeghim, and Liliane Mouawad. Identification of New Potent Acetylcholinesterase Inhibitors Using Virtual Screening and <i>in vitro</i> Approaches. *Molecular Informatics*, 38(5):1800118, 5 2019.

[233] Giulio Rastelli, Alberto Del Rio, Gianluca Degliesposti, and Miriam Sgobba. Fast and accurate predictions of binding free energies using MM-PBSA and MM-GBSA. *Journal of Computational Chemistry*, 31(4):797–810, 3 2010.

[234] Huiyong Sun, Youyong Li, Mingyun Shen, Sheng Tian, Lei Xu, Peichen Pan, Yan Guan, and Tingjun Hou. Assessing the performance of MM/PBSA and MM/GBSA methods. 5. Improved docking performance using high solute dielectric constant MM/GBSA and MM/PBSA rescoring. *Phys. Chem. Chem. Phys.*, 16(40):22035–22045, 9 2014.

[235] John Mongan, Carlos Simmerling, J. Andrew McCammon, David A. Case, and Alexey Onufriev. Generalized Born Model with a Simple, Robust Molecular Volume Correction. *Journal of Chemical Theory and Computation*, 3(1):156–169, 1 2007.

[236] Alexey V. Onufriev and David A. Case. Generalized Born Implicit Solvent Models for Biomolecules. *Annual Review of Biophysics*, 48(1):275–296, 5 2019.

[237] Lingle Wang, Jennifer Chambers, and Robert Abel. Protein–Ligand Binding Free Energy Calculations with FEP+. In *Methods in Molecular Biology*, volume 2022, pages 201–232. Humana Press Inc., 2019.

[238] Kira A. Armacost, Sereina Riniker, and Zoe Cournia. Exploring Novel Directions in Free Energy Calculations. *Journal of Chemical Information and Modeling*, 60(11):5283–5286, 11 2020.

[239] Joe Z. Wu, Solmaz Azimi, Sheenam Khuttan, Nanjie Deng, and Emilio Gallicchio. Alchemical Transfer Approach to Absolute Binding Free Energy Estimation. *Journal of Chemical Theory and Computation*, 17(6):3309–3319, 6 2021.

[240] Gregory A Ross, Chao Lu, Guido Scarabelli, Steven K. Albanese, Evelyne Houang, Robert Abel, Edward D. Harder, and Lingle Wang. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. 10 2023.

[241] Alexander D. Wade, Agastya P. Bhati, Shunzhou Wan, and Peter V. Coveney. Alchemical Free Energy Estimators and Molecular Dynamics Engines: Accuracy, Precision, and Reproducibility. *Journal of Chemical Theory and Computation*, 18(6):3972–3987, 6 2022.

[242] J. Harry Moore, Christian Margreitter, Jon Paul Janet, Ola Engkvist, Bert L. de Groot, and Vytautas Gapsys. Automated relative binding free energy calculations from SMILES to $\Delta\Delta$G. *Communications Chemistry*, 6(1):82, 4 2023.

[243] Maryam Bagherian, Elyas Sabeti, Kai Wang, Maureen A Sartor, Zaneta Nikolovska-Coleska, and Kayvan Najarian. Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in Bioinformatics*, 22(1):247–269, 1 2021.

[244] A. Suruliandi, T. Idhaya, and S. P. Raja. Drug Target Interaction Prediction Using Machine Learning Techniques – A Review. *International Journal of Interactive Multimedia and Artificial Intelligence*, InPress(InPress):1, 2022.

[245] Yi-Sue Jung, Yoonbee Kim, and Young-Rae Cho. Comparative analysis of network-based approaches and machine learning algorithms for predicting drug-target interactions. *Methods*, 198:19–31, 2 2022.

[246] Heval Atas Guvenilir and Tunca Doğan. How to approach machine learning-based prediction of drug/compound–target interactions. *Journal of Cheminformatics*, 15(1):16, 2 2023.

[247] Hao Ding, Ichigaku Takigawa, Hiroshi Mamitsuka, and Shanfeng Zhu. Similarity-based machine learning methods for predicting drug–target interactions: a brief review. *Briefings in Bioinformatics*, 15(5):734–747, 9 2014.

[248] Yayuan Peng, Jiye Wang, Zengrui Wu, Lulu Zheng, Biting Wang, Guixia Liu, Weihua Li, and Yun Tang. MPSM-DTI: prediction of drug–target interaction <i>via</i> machine learning based on the chemical structure and protein sequence. *Digital Discovery*, 1(2):115–126, 4 2022.

[249] Yanyi Chu, Xiaoqi Shan, Tianhang Chen, Mingming Jiang, Yanjing Wang, Qiankun Wang, Dennis Russell Salahub, Yi Xiong, and Dong-Qing Wei. DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. *Briefings in Bioinformatics*, 22(3):1–15, 5 2021.

[250] M Gribskov, A D McLachlan, and D Eisenberg. Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences*, 84(13):4355–4358, 7 1987.

[251] Kuo-Chen Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 5 2001.

[252] Yanrui Ding, Yujie Cai, Gexin Zhang, and Wenbo Xu. The influence of dipeptide composition on protein thermostability. *FEBS Letters*, 569(1-3):284–288, 7 2004.

[253] Geetha Govindan and Achuthsankar S. Nair. Composition, Transition and Distribution (CTD) &#x2014; A dynamic feature for predictions based on hierarchical structure of cellular sorting. In *2011 Annual IEEE India Conference*, pages 1–6. IEEE, 12 2011.

[254] Zhen Chen, Pei Zhao, Fuyi Li, André Leier, Tatiana T Marquez-Lago, Yanan Wang, Geoffrey I Webb, A Ian Smith, Roger J Daly, Kuo-Chen Chou, and Jiangning Song. iFeature: a Python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics (Oxford, England)*, 34(14):2499–2502, 7 2018.

[255] Vijayakumar Saravanan and Namasivayam Gautham. Harnessing Computational Biology for Exact Linear B-Cell Epitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. *OMICS: A Journal of Integrative Biology*, 19(10):648–658, 10 2015.

[256] Z. R. Li, H. H. Lin, L. Y. Han, L. Jiang, X. Chen, and Y. Z. Chen. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Research*, 34(Web Server):W32–W37, 7 2006.

[257] Shweta Redkar, Sukanta Mondal, Alex Joseph, and K. S. Hareesha. A Machine Learning Approach for Drug-target Interaction Prediction using Wrapper Feature Selection and Class Balancing. *Molecular Informatics*, 39(5):1900062, 5 2020.

[258] Maria Theresa F. Calangian and Vincent Peter C. Magboo. Predicting Drug-Target Interaction (DTI) based on Machine Learning with Lasso Dimensionality Reduction and SMOTE from Protein Sequence and Drug Fingerprint. In *2022 International Conference on Electrical, Computer and Energy Technologies (ICECET)*, pages 1–6. IEEE, 7 2022.

[259] Hengame Abbasi Mesrabadi, Karim Faez, and Jamshid Pirgazi. Drug–target interaction prediction based on protein features, using wrapper feature selection. *Scientific Reports*, 13(1):3594, 3 2023.

[260] Kevin Bleakley, Gérard Biau, and Jean-Philippe Vert. Supervised reconstruction of biological networks with local models. *Bioinformatics*, 23(13):i57–i65, 7 2007.

[261] Kevin Bleakley and Yoshihiro Yamanishi. Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, 25(18):2397–2403, 9 2009.

[262] Masataka Takarabe, Masaaki Kotera, Yosuke Nishimura, Susumu Goto, and Yoshihiro Yamanishi. Drug target prediction using adverse event report systems: a pharmacogenomic approach. *Bioinformatics*, 28(18):i611–i618, 9 2012.

[263] Dzmitry Bahdanau, Jan Chorowski, Dmitriy Serdyuk, Philemon Brakel, and Yoshua Bengio. End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2016-May, pages 4945–4949. IEEE, 3 2016.

[264] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. *Proceedings of Machine Learning Research*, 139:8821–8831, 2 2021.

[265] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2016-December, pages 779–788. IEEE, 6 2016.

[266] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning Hand-Eye Coordination for Robotic Grasping with Deep Learning and Large-Scale Data Collection. *Learning*, 3(9), 3 2016.

[267] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 8 2021.

[268] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 2020-December, 5 2020.

[269] Gemma Conroy. How ChatGPT and other AI tools could disrupt scientific publishing. *Nature*, 622(7982):234–236, 10 2023.

[270] ChatGPT. https://openai.com/chatgpt.

[271] Rikiya Yamashita, Mizuho Nishio, Richard Kinh Gian Do, and Kaori Togashi. Convolutional neural networks: an overview and application in radiology. *Insights into Imaging*, 9(4):611–629, 8 2018.

[272] CC BY 4.0 Deed | Attribution 4.0 International | Creative Commons. https://creativecommons.org/licenses/by/4.0/.

[273] S.M. Hasan Mahmud, Wenyu Chen, Hosney Jahan, Bo Dai, Salah Ud Din, and Anthony Mackitz Dzisoo. DeepACTION: A deep learning-based method for predicting novel drug-target interactions. *Analytical Biochemistry*, 610:113978, 12 2020.

[274] Nelson R. C. Monteiro, Bernardete Ribeiro, and Joel P. Arrais. Drug-Target Interaction Prediction: End-to-End Deep Learning Approach. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 18(6):2364–2374, 11 2021.

[275] Amr H. Mahmoud, Matthew R. Masters, Ying Yang, and Markus A. Lill. Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. *Communications Chemistry*, 3(1):19, 2 2020.

[276] Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 8 2021.

[277] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 9 2016.

[278] Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Joong Choe, Jiyeon Ham, and Woo Youn Kim. Predicting Drug–Target Interaction Using a Novel Graph Neural Network with 3D Structure-Embedded Graph Representation. *Journal of Chemical Information and Modeling*, 59(9):3981–3988, 9 2019.

[279] Davide Castelvecchi. Can we open the black box of AI? : Nature News & Comment. https://www.nature.com/news/can-we-open-the-black-box-of-ai-1.20731.

[280] Seokhyun Moon, Wonho Zhung, Soojung Yang, Jaechang Lim, and Woo Youn Kim. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13(13):3661–3673, 3 2022.

[281] Seokhyun Moon, Sang-Yeon Hwang, Jaechang Lim, and Woo Youn Kim. PIGNet2: A Versatile Deep Learning-based Protein-Ligand Interaction Prediction Model for Binding Affinity Scoring and Virtual Screening. 7 2023.

[282] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *Journal of Medicinal Chemistry*, 65(11):7946–7958, 6 2022.

[283] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 6 2017.

[284] Timo Teräsvirta. Specification, Estimation, and Evaluation of Smooth Transition Autoregressive Models. *Journal of the American Statistical Association*, 89(425):208–218, 3 1994.

[285] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. 6 2022.

[286] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-7:5112–5121, 6 2020.

[287] Kexin Huang, Cao Xiao, Lucas M Glass, and Jimeng Sun. MolTrans: Molecular Interaction Transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 5 2021.

[288] Peiliang Zhang, Ziqi Wei, Chao Che, and Bo Jin. DeepMGT-DTI: Transformer network incorporating multilayer graph information for Drug–Target interaction prediction. *Computers in Biology and Medicine*, 142:105214, 3 2022.

[289] Gan Wang, Xudong Zhang, Zheng Pan, Alfonso Rodríguez Patón, Shuang Wang, Tao Song, and Yuanqiang Gu. Multi-TransDTI: Transformer for Drug–Target Interaction Prediction Based on Simple Universal Dictionaries with Multi-View Strategy. *Biomolecules*, 12(5):644, 4 2022.

[290] Siyuan Liu, Yusong Wang, Yifan Deng, Liang He, Bin Shao, Jian Yin, Nanning Zheng,

Tie-Yan Liu, and Tong Wang. Improved drug–target interaction prediction with intermolecular graph transformer. *Briefings in Bioinformatics*, 23(5):1–10, 9 2022.

[291] Ran Zhang, Zhanjie Wang, Xuezhi Wang, Zhen Meng, and Wenjuan Cui. MHTAN-DTI: Metapath-based hierarchical transformer and attention network for drug–target interaction prediction. *Briefings in Bioinformatics*, 24(2):1–11, 3 2023.

[292] Hongmei Wang, Fang Guo, Mengyan Du, Guishen Wang, and Chen Cao. A novel method for drug-target interaction prediction based on graph transformers model. *BMC Bioinformatics*, 23(1):1–17, 12 2022.

[293] Florinel Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion Models in Vision: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10850–10869, 9 2023.

[294] Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. DiffDock: Diffusion Steps, Twists, and Turns for Molecular Docking. 10 2022.

[295] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A Sutherland, Andrew Hitchcock, C Neil Hunter, Minkyung Baek, Frank DiMaio, and David Baker. Generalized Biomolecular Modeling and Design with RoseTTAFold All-Atom. *bioRxiv*, page 2023.10.09.561603, 10 2023.

[296] Patrick Bryant, Atharva Kelkar, Andrea Guljas, Cecilia Clementi, and Frank Noé. Structure prediction of protein-ligand complexes from sequence information with Umol. *bioRxiv*, page 2023.11.03.565471, 11 2023.

[297] Guoxia Wang, Zhihua Wu, Xiaomin Fang, Yingfei Xiang, Yiqun Liu, Dianhai Yu, and Yanjun Ma. Efficient AlphaFold2 Training using Parallel Evoformer and Branch Parallelism. 10 2022.

[298] Mingyang Hu, Fajie Yuan, Kevin K Yang, Fusong Ju, Jin Su, Hui Wang, Fei Yang Zhejiang Lab, and Qiuyang Ding. Exploring evolution-aware & -free protein language models as protein function predictors. *Advances in Neural Information Processing Systems*, 35:38873–38884, 12 2022.

[299] 10 Nov 1957, Page 65 - The Times at Newspapers.com. https://www.newspapers.com/image/55787725/.

[300] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind Database: Collection of Binding Affinities for Protein–Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 6 2004.

[301] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3):405–412, 2 2015.

[302] Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Stefan Kesselheim, Marie Piraud, Fabian J. Theis, Michael Sattler, and Grzegorz M. Popowicz. MISATO - Machine learning dataset of protein-ligand complexes for structure-based drug discovery. *bioRxiv*, page 2023.05.24.542082, 5 2023.

[303] Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Stefan Kesselheim, Marie Piraud, Fabian J. Theis, Michael Sattler, and Grzegorz M. Popowicz.

MISATO - Machine learning dataset for structure-based drug discovery. *zenodo*, (10.5281/zenodo.7711953), 2023.

[304] public_binding_free_energy_benchmark/fep_benchmark_inputs/structure_inputs at main · schrodinger/public_binding_free_energy_benchmark. https://github.com/schrodinger/public_binding_free_energy_benchmark/tree/main/fep_benchmark_inputs/structure_inputs.

[305] Richard D. Smith, Jordan J. Clark, Aqeel Ahmed, Zachary J. Orban, James B. Dunbar, and Heather A. Carlson. Updates to Binding MOAD (Mother of All Databases): Polypharmacology Tools and Their Utility in Drug Repurposing. *Journal of Molecular Biology*, 431(13):2423–2433, 6 2019.

[306] Swapnil Wagle, Richard D. Smith, Anthony J. Dominic, Debarati DasGupta, Sunil Kumar Tripathi, and Heather A. Carlson. Sunsetting Binding MOAD with its last data update and the addition of 3D-ligand polypharmacology tools. *Scientific Reports*, 13(1):3008, 2 2023.

[307] H. M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000.

[308] Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, Pietro Ciceri, Lisa M Wodicka, Gabriel Pallares, Michael Hocker, Daniel K Treiber, and Patrick P Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29(11):1046–1051, 11 2011.

[309] Jing Tang, Agnieszka Szwajda, Sushil Shakyawar, Tao Xu, Petteri Hintsanen, Krister Wennerberg, and Tero Aittokallio. Making sense of large-scale kinase inhibitor bioactiv-

ity data sets: a comparative and integrative analysis. *Journal of chemical information and modeling*, 54(3):735–43, 3 2014.

[310] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A Shoemaker, Paul A Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 1 2023.

[311] David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, María Gordillo-Marañón, Fiona Hunter, Laura Junco, Grace Mugumbate, Milagros Rodriguez-Lopez, Francis Atkinson, Nicolas Bosc, Chris J Radoux, Aldo Segura-Cabrera, Anne Hersey, and Andrew R Leach. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47(D1):D930–D940, 1 2019.

[312] Michael K. Gilson, Tiqing Liu, Michael Baitaluk, George Nicola, Linda Hwang, and Jenny Chong. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Research*, 44(D1):D1045–D1053, 1 2016.

If the problem has been sloppily programmed, the answer will be just as incorrect. If the programmer made mistakes, the machine will make mistakes.

William D. Mellin

2

Critical Assessment of Dataset Quality for Protein-Ligand Interaction Prediction

As described in the previous chapter, the quality of a dataset is fundamental to training deep learning models. We investigated the effects of different treatments of the PDBbind dataset on the performance of a deep learning model. Additionally, we developed a novel tool for the physics-based identification of protein-ligand interactions and compared it to well-established and commonly used tools. The following article was published as a preprint.[1]

Quality Matters: Deep Learning-Based Analysis of Protein-Ligand Interactions with Focus on Avoiding Bias

Manuel S. Sellner[1,2]
manuel.sellner@unibas.ch

Markus A. Lill[1,2]
markus.lill@unibas.ch

Martin Smieško[1,2*]
martin.smiesko@unibas.ch

[1] Computational Pharmacy, Department of Pharmaceutical Sciences, University of Basel
[2] SIB Swiss Institute of Bioinformatics * Corresponding author

November 16, 2023

2.1 Abstract

The efficient and accurate prediction of protein-ligand binding affinities is an extremely appealing yet still unresolved goal in computational pharmacy. In recent years, many scientists have taken advantage of the remarkable progress of deep learning and applied it to address this issue. Despite all the advances in this field, there is increasing evidence that the typically applied validation of these methods is not suitable for medicinal chemistry applications. This work assesses the importance of dataset quality and proper dataset splitting techniques demonstrated on the example of the PDBbind dataset. We also introduce a new tool for the analysis of protein-ligand complexes, called po-sco. Po-sco allows the extraction of interaction information with much higher detail and comprehensibility than the tools available to date. We trained a transformer-based deep learning model to generate protein-ligand inter-

action fingerprints that can be utilized for downstream predictions, such as binding affinity. When using po-sco, this model generated predictions that were superior to those based on commonly used PLIP and ProLIF tools. We also demonstrate that the quality of the dataset is more important than the number of data points and that suboptimal dataset splitting can lead to a significant overestimation of model performance.

2.2 Introduction

Protein-ligand interactions, which drive molecular recognition processes play a crucial role in many biological processes. The accurate prediction of binding affinities associated with such interactions belongs to the most important tasks in drug design. In recent years, various computational methods have been conceived and developed to predict protein-ligand binding affinities, which can help guide drug discovery efforts and reduce the time and cost associated with experimental screening[2–12].

2.2.1 Prediction of binding affinities

A well-established approach to create protein-ligand binding modes is molecular docking. This method employs algorithms that search for the optimal orientation and conformation of the ligand (referred to as pose) within the active site of the protein and then try to quantify the binding energy based on the interaction between the protein and the ligand[13–17]. While it is relatively simple and fast, it is not very accurate in predicting binding affinities[18–21]. Another method that has gained increasing importance during the last decade is free energy perturbation (FEP). It relies on molecular dynamics simulation to calculate the change in free energy upon binding of the ligand to the protein. This method can deliver predictions of relative and absolute binding affinities of chemical accuracy, but is computationally extremely demanding[7,22–25]. Other computational methods for predicting binding affinities include molecular mechanics Poisson-Boltzmann surface area (MM-PBSA) and molecular mechanics generalized Born surface area (MM-GBSA), which involve the calculation of binding free energies using molecular mechanics force fields and implicit solvation models[26–28]. The development of these computational methods has facilitated the prediction of protein-ligand binding affinities and has the potential to accelerate drug discovery and design efforts. However, most of these methods still suffer from either poor accuracy or high computational cost.

In recent years, deep learning approaches have gained popularity for predicting protein-ligand binding affinities. Deep learning models use artificial neural networks to learn the complex relationships between the molecular features of the protein and ligand and their binding affinity[29]. A commonly used approach is the use of convolutional neural networks (CNNs) to predict binding affinities. These models usually take 1- or 3-dimensional representations of the protein and ligand as input and use convolutional layers to extract local features from the structures. The extracted features are then fed into fully connected layers to predict the binding affinity[30–32]. Although CNNs excel at translational invariance, they are not rotationally invariant, which may cause problems when working with 3-dimensional input data. Another common deep learning approach for the prediction of binding affinities is the use of graph neural networks (GNNs). GNNs operate on molecular graphs, where atoms are considered as nodes and chemical bonds as edges. These models can learn the molecular representation of protein binding sites and ligands from their graph structure. GNNs are commonly used to predict binding affinity by considering binding site residues, ligand atoms, and interactions between them on the molecular graph[33–37]. When implemented correctly, GNNs can have translational, rotational, and reflection equivariance (i.e. E(3) equivariance)[38]. Recently, transformer models[39] have shown great promise in many different fields[40–44]. In protein-ligand interaction prediction, these models take advantage of the power of self- and cross-attention modules to analyze complex interactions between proteins and ligands and predict their binding affinity with seemingly high accuracy[45,46]. A key advantage of transformer models is their ability to model long-range dependencies in sequence-based data and capture complex interactions between amino acids and ligands. This is critical for accurately predicting binding affinities, as these interactions can be highly specific and context-dependent.

So far, deep learning approaches have shown promising results in predicting protein-ligand binding affinity and have the potential to further advance the field of drug discovery and design. However, the development of accurate deep learning models for predicting binding affinity requires large amounts of high-quality data and careful selection of molecular features to incorporate into the models. In this work, we use a transformer-based model that uses self- and cross-attention to learn from residue-ligand interaction fingerprints to predict binding affinities and protein-ligand interaction fingerprints.

2.2.2 Biases in binding affinity prediction

While deep learning approaches have great promise in predicting protein-ligand binding affinities, they are susceptible to various biases that can limit their accuracy and generalizability.

The data used to develop a deep learning model can be a major source of bias. Such bias can have various origins, for example if the training data used to develop the models are not representative of the true distribution of protein-ligand binding affinity data. This can happen due to, e.g. the limited availability of high-quality binding affinity data, the uneven distribution of binding affinity values, or the use of biased selection criteria for the data. To objectively assess the performance of a model, it is absolutely vital to avoid any overlap between training, validation, and test sets. As a result of such bias, trained models may not be able to accurately generalize to new data outside the training set, leading to overfitting. Fan and Shi have demonstrated that the overlap of proteins and ligands in the test and training set greatly influences the apparent performance of the model and leads to a severe overestimation of the capabilities of the model[47]. They also proposed that such a bias has a greater influence on the performance of a model than its architecture.

Another potential bias in deep learning models is feature bias, where the features used to train the models do not capture all the relevant information about protein-ligand interactions. For example, if the features used only capture the geometric properties of the protein and ligand, they may not account for the dynamic changes that occur during the binding process. Feature bias can also occur if the features used contain too much information, causing the model to overfit. Volkov et al. showed that using 3D information of proteins and ligands often introduces bias if molecular structure information is provided directly to the model[48]. For example, if the fully atomistic structure of the ligand and/or binding site, or the protein sequence is provided to a neural network, it often learns to memorize the specific data and does not generalize well. They proposed that instead of directly using structural information, one should rely only on extracted interactions between the protein and the ligand.

To overcome bias, it is important to carefully curate high-quality datasets that accurately represent the range of binding affinity values, meticulously split the data into training, validation, and test sets, and carefully select the input features of the model.

2.2.3 Types of protein-ligand interactions

The analysis of protein-ligand interactions in a structural complex is critical to understanding the binding mechanism and designing new therapeutics. Thus, it seems logical that a detailed analysis of protein-ligand interactions may aid deep learning models to properly learn binding affinities. One way to analyze these interactions is to identify, enumerate, and quantify specific types of molecular interactions recognized in physics and chemistry, such as hydrogen bonds, salt bridges, hydrophobic interactions, pi-pi, sigma-pi, pi-cation interactions, and halogen bonds.

Hydrogen bonds are arguably the most common and important specific molecular interactions in protein-ligand complexes. These interactions occur between a hydrogen atom carrying a partial positive charge (typically due to an electron-withdrawing carrier atom or functional group) and an electronegative atom featuring a lone electron pair, such as oxygen or nitrogen. The strength of a hydrogen bond depends on the distance and angle between the atoms involved, as well as polarization effects. Analyzing hydrogen bonds can not only provide information on the strength of a ligand binding to a protein, but also on the specificity[49].

Charge-assisted interactions like salt bridges are the strongest type of intermolecular interactions. Salt bridges are formed by the electrostatic attraction of oppositely charged atoms or functional groups. This type of interaction can be crucial to stabilizing the protein-ligand complex if the ligand contains charged functional groups, also due to its long range compared to other interactions[50]. Therefore, recognition of salt bridges, but also potentially unsatisfied charged moieties missing counterparts or even repulsive interactions in a protein-ligand complex is imperative for correctly estimating binding affinities.

Although generally weaker than polar interactions if quantified per interacting atom pair, hydrophobic interactions - due to their numerous occurrence - are a key contributor to the strength of a protein ligand complex[49]. These interactions arise from the close proximity of non-polar atoms that shield each other from unfavorable interactions with water (hydrophobic). Ferreira de Freitas and Schapira found that high-efficiency ligands often show numerous hydrophobic interactions when analyzing protein-ligand interactions in the protein data bank (PDB)[51].

Pi-pi interactions are a specific type of dispersion forces and typically occur between large planar functional groups with pronounced electron delocalization like aromatic rings[51,52]. These interactions can be important in stabilizing the protein-ligand complex and can provide additional specificity (besides directional H-bonds and salt bridges) for ligand binding[49]. Since pi-pi interactions have been found to be the third most common protein-ligand interaction in the PDB, identifying different forms of pi-pi interactions (such as face-to-face pi-pi and T-shaped sigma-pi) is important for the analysis of binding complexes[51]. Pi-cation interactions are another type of non-covalent interaction that occurs between a positively charged group, such as a metal cation, protonated amine, or guanidinium group, and an electron-rich π-system. Although pi-cation interactions are not as common as pi-pi stacking, these interactions may be important for stabilizing protein-ligand complexes and can also contribute to ligand binding specificity[51].

Halogen bonds are a type of non-covalent interaction that occurs between a halogen atom beyond fluorine (such as chlorine, bromine, or iodine) and either an electrophile or a nucleophile atom depending on the geometry. Most halogen bonds consist of interactions between a region of the halogen with low electron density (called the sigma hole) and electronegative atoms such as oxygen or nitrogen featuring a free electron pair. However, halogen bonds can also form between a halogen atom's electron-rich belt and electropositive atoms, such as polarized hydrogens. While the former have a linear geometry (i.e. the angle C-X$\cdots$O is close to 180°), the latter halogen bonds prefer a perpendicular setup (i.e. the angle C-X$\cdots$Hpolar is close to 90°)[53,54]. Thus, as with hydrogen bonds, the strength of halogen bonds depends on the distance and angle between the atoms involved.

In general, analysis of specific molecular interactions in protein-ligand complexes can provide valuable insight into the binding mechanism and can inform the design of novel therapeutics. By understanding the specific interactions involved in protein-ligand binding, researchers can design more effective drugs with higher binding affinity and better selectivity. The detailed analysis of molecular interactions can also help computational scientists develop tools and models for various predictions in the field of drug design or toxicology.

There are a few tools that allow the generation of protein-ligand interaction fingerprints. The two most widely used tools are PLIP[55] and ProLIF[56]. PLIP allows the detection of hydrophobic interactions, hydrogen bonds, aromatic stacking, pi-cation interactions, salt bridges, water-bridged hydrogen bonds, halogen bonds, and metal interactions. The newer ProLIF tool allows for detecting hydrophobic interactions, hydrogen bonds (distinguishing between accepting and donating groups), pi-pi stacking (edge-to-face and face-to-face), pi-cation and cation-pi interactions, salt bridges (with differentiation between cationic and anionic groups in the ligand), donating and accepting halogen bonds, as well as donating and accepting metal interactions. Although these tools are sufficient for the analysis of the most common interactions, they lack more sophisticated functionalities: in particular, the differentiation between backbone and side-chain hydrogen bonds, linear (to nucleophiles) and perpendicular (to electrophiles) halogen bonds, the detection of polarized hydrogen bonds, and repulsive interactions. To address these points, we developed a custom tool for protein-ligand interaction analysis, which is the most sophisticated tool to our knowledge so far. We call our tool po-sco (as an abbreviation of pose-scorer).

2.2.4 The PDBbind dataset

At present, several datasets are commonly used to train artificial intelligence models to predict protein-ligand binding affinities. These datasets vary in size, diversity, and quality, and each has its own advantages and disadvantages.

One of the most widely used datasets is the PDBbind database, which harbors experimentally determined binding affinities for a diverse set of protein-ligand complexes[57]. The PDBbind database has been used to train and evaluate a wide range of machine learning and deep learning models, and is frequently used as a benchmark to assess the performance of new models[3,58–61]. One advantage of the PDBbind database is its size and diversity, which allows for the development of models that can potentially generalize well to new protein-ligand complexes. Another strong feature of the PDBbind dataset is that it links the three-dimensional protein-ligand structural information (including their PDB ID) with the corresponding binding affinity data. This allows the use of three-dimensional features, such as molecular interactions, in prediction models.

The major disadvantage of the PDBbind database is that it contains a relatively small number of high-quality complexes. This may be limiting for models that need a high level of detail. Also, it is not trivial to achieve a good dataset split limiting the bias as much as possible. The PDBbind dataset is divided into a general, refined, and core set. The general set represents the largest part and contains complexes of lower quality. The refined set is a curated set that contains only complexes that meet certain quality criteria. These criteria include quality checks of the 3D structures (e.g., a resolution < 2.5 Å, no covalently bound ligands, and no steric clashes) and the binding data (e.g., K_d and K_i instead of IC50, exact values instead of ranges, and affinity data only within a pharmaceutically relevant range). The core set is the smallest of the three and contains only high-quality complexes of high diversity. Some researchers use the general, refined, and core set as training, validation, and test set, respectively[3]. Others randomly split one of the sets to construct the datasets used for training[60]. Since there are many overlaps of proteins (i.e. Uniprot IDs) and ligands (with the same 2D structure) among the different sets of PDBbind, both approaches probably lead to the introduction of a large bias, as discussed in Section 2.2.2. Thus, a great deal of attention must be paid to using the PDBbind dataset for developing tools for protein-ligand binding affinity prediction.

The PDBbind dataset contains a multitude of protein-ligand complexes. Each complex is provided in the form of a protein structure file, a file containing only the binding pocket of the protein, and the ligand in two different formats. While the ligand file contains all explicit hydrogen atoms, the protein structure contains only polar hydrogen atoms, and water molecules contain no hydrogen atoms. This may be problematic, e.g., because the orientation of co-crystallized water molecules is unknown, although water molecules are considered imperative for protein-ligand binding[32,62–64]. In addition, the ligand structures in the PDBbind dataset are provided in neutral state. Thus, the generation of the correct protonation states of the ligand (and the protein) at the desired pH must be performed by the user.

2.2.5 Our contribution

In this work, we developed an in silico tool called po-sco to comprehensively analyze native protein-ligand interactions as they appear in crystals and create residue-based interaction fingerprints. These fingerprints are designed to be constructed in a way that excludes any explicit information about the ligand structure or binding site geometry in order to avoid any structural bias. An attention-based deep neural network is then employed to learn the corresponding experimentally determined protein-ligand binding affinities. Furthermore, the model converts residue-based interaction fingerprints into a single protein-ligand interaction fingerprint that can be used for downstream tasks.

As we use the PDBbind dataset to train and validate our model, we also show how different splitting procedures affect the apparent model performance, while already having reduced structural bias by focusing on molecular interactions rather than fully atomistic structures.

With this work, we hope to shed light into the murky waters of deep learning-based protein-ligand binding affinity prediction and highlight the importance of high-quality datasets and features.

2.3 Results and Discussion

2.3.1 Po-sco interaction analysis

Po-sco provides a great level of detail when it comes to the analysis of protein-ligand complexes. It is capable of detecting the large majority of currently recognized types of intermolecular interactions found in protein-ligand complexes. Additionally, its detection algorithms implement a high degree of physics and medicinal chemistry rules to provide the user with very fine-grained analyses. While traditional protein-ligand interaction analysis tools are limited to binary fingerprints, po-sco extends binary fingerprints with continuous values, allowing even more information to be stored in a single fingerprint. The combination of all these features leads to 28 binary and 6 continuous features extracted by po-sco (see Table 2.1 for a full list of included features).

To demonstrate the binary interactions identified by po-sco, we analyzed the crystal structure of human tryoptophan hydroxylase type 1 co-crystallized with the ligand LP-533401 (PDB ID 3HF8). An overview of the identified interactions can be found in Figure 2.1. For better visibility, we split the interactions in polar (Figure 2.1A), hydrophobic (Figure 2.1B), and exotic (Figure 2.1D) interactions. Furthermore, we show the identified unsaturated polar functional groups in Figure 2.1C. The figure shows that po-sco was able to reliably identify the most important interactions in high detail. We believe that not only the detailed analysis of the existing interactions, but also the information of missing interactions as shown in Figure 2.1C are important for assessing the quality of a protein-ligand complex.

Table 2.1: Protein-ligand interactions calculated by PLIP, ProLIF, and po-sco. Where not otherwise stated, the interactions are listed from the perspective of the protein residue. Except for the interaction types under "Po-sco continuous", the presence or absence of an interaction is indicated by a binary value. For "Po-sco continuous", the values are continuous, ranging from 0 to 1.

PLIP	ProLIF	Po-sco binary	Po-sco continuous
Hydrophobic	Hydrophobic	Hydrophobic	Normalized lipophilic energy
Hydrogen bond	Hydrogen bond acceptor	H-bond donor in side chain	Normalized H-bonding energy
	Hydrogen bond donor	H-bond donor in backbone	Fraction of unsaturated H-bond donors
		H-bond acceptor in backbone	Fraction of unsaturated H-bond acceptors
		H-bond acceptor in side chain	
		Unsaturated H-bond donor	
		Unsaturated H-bond acceptor	
		Polarized H-bond	
Salt bridge	Cationic salt bridge	Cationic salt bridge	
	Anionic salt bridge	Anionic salt bridge	
		Unsaturated cation	
		Unsaturated anion	
Water bridge		Residue is water	
Pi-pi stacking	Pi-pi interaction T-shaped	Pi-pi interaction T-shaped	
	Pi-pi interaction face-to-face	Pi-pi interaction face-to-face	
		Sigma-pi interaction	
		Pi-pi interaction w/o rings	
Pi-cation interaction	Pi-cation interaction	Pi-cation perpendicular	
	Cation-pi interaction	Cation-pi perpendicular	
		Pi-cation parallel	
		Cation-pi parallel	
Halogen bond	Halogen bond	Halogen bond to sigma hole	
		Halogen bond to sigma hole from aromatic ring	
		Halogen bond to electron belt	
Metal interaction	Metal interaction	Metal interaction[a]	
		Residue is metal	
	VdW contact		
		Residue is co-factor	Fraction of polar atoms contributing to intermolecular interactions
		Repulsive interaction	Fraction of polar atoms contributing to intramolecular interactions

[a] Residue is a metal and is interacting with the ligand

2.3.2 Attention-based prediction model

The first goal of this work was to find out if more detailed interaction analysis benefits the prediction of binding affinities. As shown in Table 2.1, the level of detail provided by PLIP,

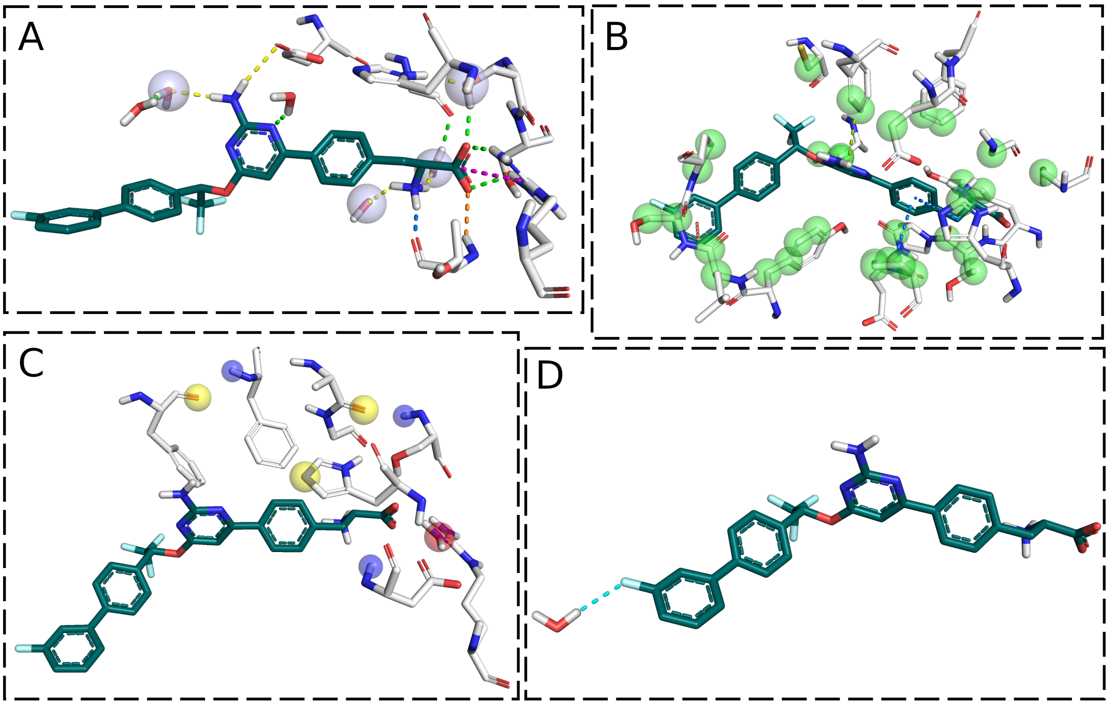


Figure 2.1: Visualization of the binary interactions identified by po-sco in PDB ID 3HF8. A) Polar interactions where yellow, green, blue, and orange dashed lines represent interactions with H-bond acceptors in side chains, H-bond donors in side chains, H-bond acceptors in the backbone, and H-bond donors in the backbone, respectively. Salt bridges are shown as dashed magenta lines, and polarized H-bond acceptors are marked with a semi-transparent sphere. B) Hydrophobic interactions where blue, green, and orange dashed lines represent face-to-face pi-pi, T-shaped pi-pi, and pi-pi interactions to smaller elements, respectively. Green spheres mark residues with hydrophobic contacts with the ligand. C) Unsaturated polar functional groups. The yellow, blue, and red spheres represent unsaturated uncharged acceptors, unsaturated uncharged donors, and unsaturated cationic groups, respectively. D) A single exotic group (halogen bond to electron-rich belt) is shown.

ProLIF, and po-sco differs significantly. Therefore, we trained multiple versions of the same model with interaction fingerprint data extracted from the three analytic tools. For po-sco, we trained two models, one with explicit information on the ligand happiness and one without it. In order to compare the model performance, we calculated the Pearson correlation coefficient (PCC), the mean unsigned error (MUE), and the root mean squared error (RMSE) between predicted and true, experimentally determined affinities. It is important to note here that the goal was not to get a perfect prediction of binding affinities, but to compare how different model input data affect the quality (predictivity) of results while using the same model

Table 2.2: Comparison of predictions with different interaction analysis tools. The performance metrics used to compare PLIP, ProLIF, and po-sco are the PCC, the MUE, and the RMSE. The shown metrics apply to the external test set only. The best values are highlighted in bold font. All models were trained on a diverse split of the refined PDBbind set (see Section 2.4.1).

Tool	PCC	MUE	RMSE
PLIP	0.37	2.20	2.71
ProLIF	0.53	2.03	2.48
Po-sco w/o L	0.63	1.92	2.36
Po-sco	**0.66**	**1.84**	**2.34**

architecture.

Table 2.2 serves as the performance comparison of the different models. All three performance metrics are worst for PLIP and best for po-sco (with ligand information). They also clearly correlate with the level of detail provided by the different tools. ProLIF distinguishes between hydrogen bond donors and acceptors as well as between anionic and cationic salt bridges. These details are missing in PLIP. Although ProLIF lacks detection of water bridges, it performed better than PLIP. This may be due to the lower occurrence of these kinds of interaction and, therefore, to their lower relative importance for the prediction of binding affinities. Hydrogen bonds and salt bridges, on the other hand, are among the most common protein-ligand interactions and are therefore considered highly important. Thus, it seems reasonable that a higher level of detail in these common interaction types leads to better results.

Both po-sco models with and without ligand information outperform the PLIP and ProLIF models, while the model with ligand information performed the best. This improved performance is likely due to the very high level of detail provided by po-sco. Not only does po-sco provide fine details about individual interaction types (e.g. hydrogen bond donor/acceptor in backbone/side chain), but it also extends binary fingerprints by continuous values, allowing the storage of a higher amount of information. The results in Table 2.2 show that this attention to detail is beneficial for the prediction of binding affinities using a deep learning model.

Adding information on the ligand “happiness”, which describes the level of saturation of various kinds of interaction partners by protein counterparts, seems to also benefit the model. However, it must be noted that in order to implement the ligand information, the prediction model used a slightly adapted architecture (see Section 2.4.3).

2.3.3 Impact of dataset split

The second goal of this work was to investigate the influence of dataset splitting, specifically the PDBbind dataset, on model performance. Here, we prepared three different splittings, two based on the complete PDBbind dataset and one based on the refined set. See Section 2.4.1 for details on the construction of the data set.

In a PDBbind splitting that is often seen in the literature, the general set is used for training, the refined set for validation, and the core set for testing. We analyzed the occurrence of proteins (i.e. Uniprot IDs) across the different sets in this split to identify potential overlaps. There were 2388, 1494, and 64 different proteins in the general, refined, and core set, respectively. Of the 1494 different proteins in the refined set, 74% were also part of the general set, and in the core set it was even 97%. Thus, even though there were different PDB entries in the training, validation, and test set in this split, there was a substantial overlap of proteins. This means that, while the exact conformation of the binding site residues likely differs in the different PDB entries of the same protein, the general similarity between different conformations of the same protein binding site introduces a large bias. This may allow a deep learning model to memorize the binding sites without learning much from the analyzed interactions and still perform well on similar binding sites.

To avoid such bias, we created a custom split that randomly assigns complexes to the train-

Table 2.3: Comparison of the po-sco-based prediction model using different splittings of the PDBbind dataset. The performance is shown only for the external test set. The best performances are highlighted in bold.

Model	Dataset	Split	PCC	MUE	RMSE
Model 1	Complete	Classical	**0.66**	**1.78**	**2.23**
Model 2	Complete	Diverse	0.48	1.92	2.42
Model 3	Refined	Diverse	**0.66**	1.84	2.34

ing, validation, and test set, while avoiding spreading complexes containing the same protein across different sets.

We trained three versions of the same model. Model 1 was trained on the complete PDBbind dataset with the "classical" splitting, i.e. the general set for training, the refined set for validation, and the core set for testing. Model 2 was also trained on the complete PDBbind dataset, but with our custom split that enforces diverse proteins between sets. Model 3 was trained on the refined set only with the same diversity-optimized splitting as in model 2. Due to the best performance of the po-sco based prediction model compared to the ones based on PLIP and ProLIF, we focus here on models trained with po-sco inputs only. Data for the same analyses performed with PLIP and ProLIF-based models can be found in the Supporting Information in Tables A2.1 and A2.2, respectively.

Table 2.3 shows the performance of the model trained on the different datasets. It can be seen that, with respect to the prediction error, model 1 performs better on the test set than the other two models. However, this seemingly high performance is misleading, since 97% of the proteins in the test set were already seen during training. After removing overlapping proteins between training, validation, and test set, the performance drops significantly, indicating that the model has much more difficulty actually learning from the data (model 2).

Finally, when using only the refined set with a diverse split (model 3), the performance is higher compared to training on the complete dataset with the same splitting. Although the training set of model 2 was more than twice the size of that of model 3, the additional data points were not beneficial in model training. This indicates that for such tasks, the higher quality of the structure data in the refined set is clearly beneficial for meaningful (generally applicable and transferable) learning of the models.

2.3.4 Real-world example

We decided to test the trained models on a real-world example. For this, we chose the human 5HT receptor 1B (Uniprot ID P28222) because it is a pharmaceutically relevant target and is not part of the PDBbind dataset.

We docked more than 700 compounds with known binding affinities to the 5HT receptor using smina and Glide[13,16]. The generated poses were re-scored with the three trained po-sco models and with gnina[6]. Again, we calculated the PCC, MUE, and RMSE of the different methods. The exact details can be found in Section 2.4.4.

Table 2.4 shows the results of this study. No tested method yielded satisfactory results, as the correlations between predicted and true affinities were quite weak. Judging by the MUE and RMSE, gnina does the best job in predicting binding affinities. Based on the performance reported in Table 2.3, it could be expected that model 1 performs the best out of all po-sco models. However, it in fact performed much worse than model 3 which was only

Table 2.4: Comparison of different methods for the prediction of binding affinities towards the human 5HT 1B receptor (Uniprot ID P28222). The po-sco models 1, 2, and 3 were trained on the complete PDBbind dataset with a classical split, a diverse split, and on the refined subset with a diverse split, respectively.

Method	PCC	MUE	RMSE
Smina	0.39	1.38	1.72
Glide	0.47	1.58	1.94
Gnina	0.41	1.25	1.50
Po-sco model 1	0.39	1.97	2.36
Po-sco model 2	0.37	1.37	1.64
Po-sco model 3	0.40	1.58	1.95

trained on the refined subset of the PDBbind dataset. We think that model 1 mainly learned to memorize the structures in the training set. Due to the high overlap between the training, validation, and test set, the reported performance is accordingly high. When presented with a new protein structure, the model has difficulty in recognizing it and is unable to make an accurate prediction. Model 3, in contrast, was trained on data of superior quality with no overlap between the training, validation, and test sets, thus requiring it to gain more knowledge from the data itself. This clarifies why this model is more effective than model 1, even though the results in Table 2.3 may suggest otherwise.

Interestingly, model 2 which was trained on the complete PDBbind dataset with a split that removed all protein overlaps between training, validation, and test set, performed the best of the three po-sco models in this study. At first glance, this is surprising since the performance on the test set (Table 2.3) was much worse. One explanation could be that there are differences in the training or test set between models 2 and 3, leading to results that cannot be compared. However, 85% of the complexes that are in the training set of model 3 are also in the training set of model 2. For the test set, it is more than 75%. Also, as the protein at hand is not part of the PDBbind dataset at all, these effects will likely not have a big impact. Another possible explanation is that the performance on the test set as reported in Table 2.3 was a fluke and the true performance is better. This can be ruled out because the model performs equally as bad (or even worse) on the validation set (PCC 0.51, MUE 2.04, RMSE 2.47). A third explanation could be that the high performance of model 2 in this example was a "lucky shot" and is not representative for the general performance of the model which is better represented by the results in Table 2.3. Further investigation showed that while the human 5HT 1B is not part of the PDBbind dataset, the general set contains the turkey 5HT 1B receptor. In fact, this protein was part of the training set of models 1 and 2. Visual inspection showed that the binding sites of the turkey and human 5HT 1B receptors are not the same, although they are very similar. Thus, model 2 appeared to have an easier job predicting these compounds than model 3 because it has been trained on a very similar structure, which explains the higher performance. The fact that model 1 was also trained with this structure but still performed the worst of the three models further highlights that it probably did not have to learn actually meaningful information to still achieve a seemingly high performance due to the overlap between the data sets.

Another point that needs to be kept in mind is that in this experiment, the model was applied to poses from molecular docking whereas it was trained on crystal structures. This could also have an influence on the outcome.

We strongly advocate that researchers be very mindful about the data set used to train a model and how it was handled. This enables the early detection of reported performances that have been exaggerated and prevents the scientific literature from being inundated with models that look good on paper, but fail in real-world scenarios. It is essential to critically evaluate the reported performance of a model.

2.4 Methods

2.4.1 Dataset construction

As described in Section 2.2.4, the protein and ligand structures provided in the PDBbind dataset need to be prepared before using them for structure-based modeling approaches. We used the protein preparation wizard in Schrödinger's Maestro to prepare all structures[65]. The preparation workflow consisted of adding all explicit hydrogen atoms, assigning bond orders, generating protonation states at the pH present at crystallization (for the protonation of the ligand, Epik was used[66–68]), generating water orientations, and running a restrained minimization with an RMSD cutoff of 0.3 Å.

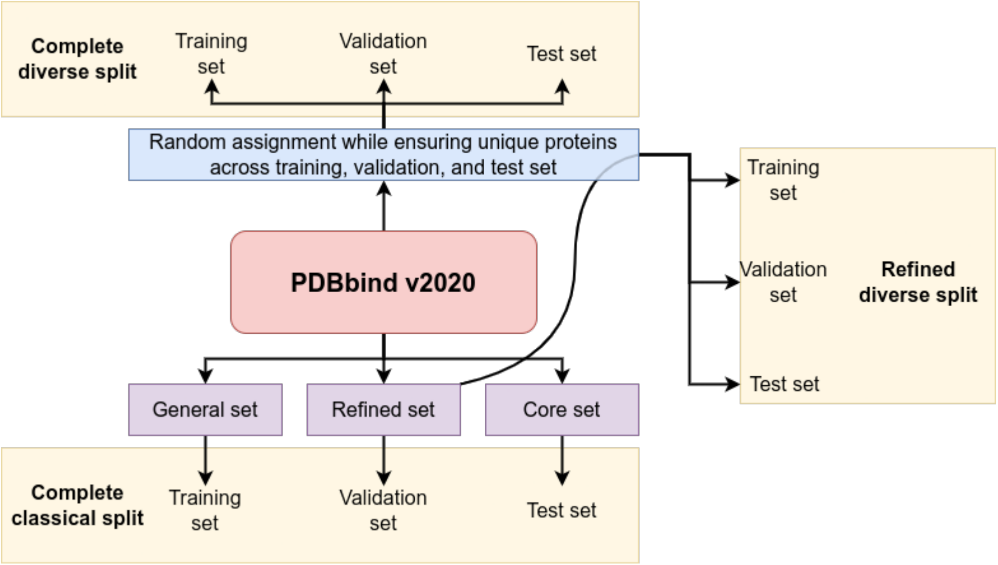


Figure 2.2: The splitting scenarios applied to the PDBbind dataset. The complete dataset was split in two ways, one in which the general, refined, and core set are used for training, validation, and testing, respectively (complete classical split), and one in which the complexes were split randomly while ensuring that there are no overlaps of proteins between the training, validation, and test set (complete diverse splitting). The same diverse splitting was applied to the refined set only (refined diverse splitting).

In this work, we used several different splits of the PDBbind dataset version 2020 (Figure 2.2). The complete dataset (consisting of the general, refined, and core set) contained 19,443 protein-ligand complexes. For the general set, there were affinities of Ki, Kd, and IC50 types. Since Ki/Kd and IC50 cannot be directly compared, we removed samples with affinities of type IC50. This left us with 6558 complexes in the general set, 5041 in the refined set, and 256 in the test set after removing all duplicate complexes.

The complete dataset was split in two ways: 1) The classical split, where the general set was used for training, the refined set for validation, and the core set for testing; 2) the diverse split, where the complexes were randomly assigned to either the training, validation, or test set while ensuring that there are no overlaps regarding the protein (i.e. Uniprot ID) between the datasets. This means that, while there were several complexes containing the same protein in one dataset, the proteins were unique between different datasets. The training set contained around 80% of the complexes, while the validation and test set contained around 10% of the complexes each. For the refined set (containing complexes of higher quality compared to the general set), the same technique for generating a diverse split was used.

The model architecture used here, combined with the provided input, does not allow conclusions to be drawn about the structure of the ligand. This was specifically chosen to avoid the introduction of bias and thus prevent the model from simply memorizing structures of ligands.

2.4.2 Po-sco interaction analysis

The po-sco interaction analyzer tool contains algorithms to detect the most significant intermolecular protein-ligand interactions currently recognized by the modeling community. The list of all supported interaction types can be found in Table 2.1.

Detection algorithms have been implemented in such a way that they reliably identify specified interactions within reasonable structural deviations (see below). While we trained our models on crystallographic data, where realistic interaction patterns (e.g. interatomic distances, angles, ring and π-system planes) are naturally expected, we aim at using po-sco for scoring of docked poses, where moderate deviations from ideal parameters can be frequently observed, even with “good” poses that we typically refer to as those within 2.0 Å from the native pose. Therefore, e.g. for distance thresholds, a general factor allowing for a 20% deviation from the optimum is applied. In terms of linearity, a deviation of up to 45 degrees is accepted.

Compared to similar tools, in many cases, our routines are parameterized with more stringent criteria. For example, for detecting pi-pi face-to-face stacking and T-shape interactions between rings we use different cut-off values for centroid distances, which helps minimizing false positive as well as false negative cases. We implemented our expertise from modeling and docking small molecules to a multitude of different targets covering membrane bound and soluble proteases, metalloenzymes, cytochromes, GPCRs, nuclear receptors, ion channels, etc. During the development we relied heavily on benchmarking with real-world systems. The H-bonding parameters, both in terms of geometry and energetics, are based on the Yeti force field by Vedani et al.[69]. The same applies for ligand-metal interactions[70].

The algorithm for the atom-type-dependent quantification of hydrophobic interaction was parameterized based on the relative solvation free energies of matched molecular pairs of unsubstituted benzene, toluene, fluro-, chloro-, bromo-, iodo- benzene and thiophenol in the organic solvent hexadecane (cf. Figure S3). The underlying data were extracted from the Minnesota Solvation database[71]. Toluene was selected as the reference substance, providing a weighting coefficient of 1.0 for the sp3-hybridized carbon atom. Equation 2.1 shows the calculation of the hydrophobic interaction energy between atoms a_i and a_j where w_i and w_j are the weights according to Table A2.3 for the respective atoms, k is an additional factor according to Li et al.[72], $r_{vdw}(\cdot)$ is the Van der Waals radius of a given atom, n is a slope parameter, and d_{ij} is the distance between atoms i and j. The slope parameter n was set to -6.0 and following the article by Li et al., we defined $k = -0.3$.

$$E(\text{hydrophobic}) = \frac{k \cdot w_i \cdot w_j}{1 + \exp\left(n \cdot \left(d_{ij} - 1 - r_{vdw}(a_i) - r_{vdw}(a_j)\right)\right)} \tag{2.1}$$

Ring stacking interaction is detected up to the ring centroid distance of 5.0 Å. Depending on the ring character (aromatic or aliphatic), a detailed analysis of ring interaction patterns is performed. In case of aromatic rings, the pair-wise distances are calculated between all atoms in the first ring and all atoms in the other ring counting the number of hydrogen and heavy atoms at or below the sum of the Van der Waals radii. Based on the prevailing interactions (hydrogen-to-heavy or heavy-to-heavy), the type of interaction is determined (T-shaped or parallel). In our experience, this procedure works more reliably than calculating the angle of ring planes.

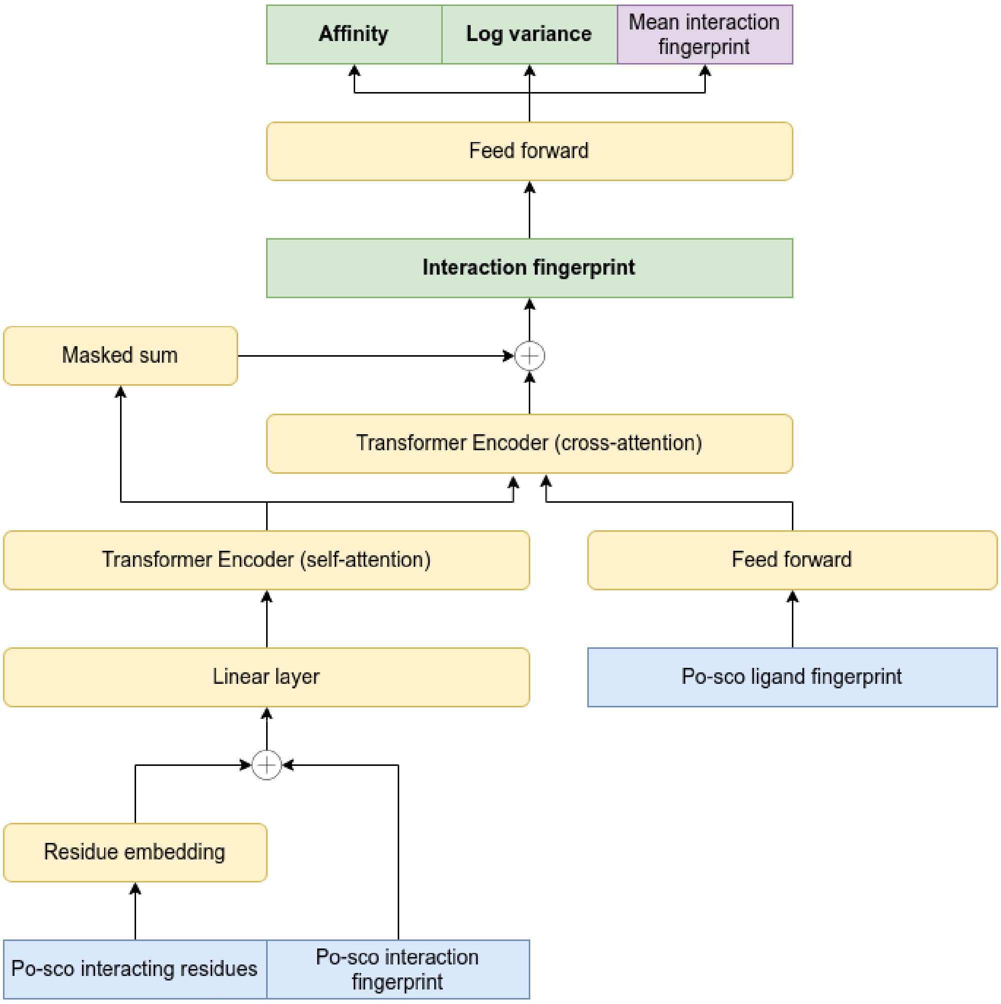
Detection of complex interaction patterns, e.g. polarized H atoms interacting with an aromatic ring, is based on a thorough analysis of the local environment of interacting atoms, i.e. for switching the relevant fingerprint bit on, such a polar hydrogen atom must be within acceptable interaction distance (below the sum of VdW radii) with at least three heavy atoms forming the target pi system of the ring, and simultaneously the hydrogen exit vector must be co-linear (with the ring normal with a maximum allowed deviation, here 30 degrees).

Similarly, pi-pi stacking of smaller, non-ring systems is detected for at least three consecutive sp2 hybridized atoms (e.g. peptide bond π-system). After confirming the validity of geometric criteria, interactions are classified according to the chemical character of atoms in higher order interacting functional groups (neutral, positively or negatively charged π-system and all combinations thereof).

Halogen bonding is also detected at interatomic distances below or at the sum of Van der Waals radii of the analyzed atom pair. Depending on the angle with respect to the atom, which carries the halogen atom, interactions are classified to the sigma hole (angle carrier atom - halogen - heteroatom is greater than 135 degrees) or the belt (angle carrier atom - halogen - hydrogen in the range of 45 to 135 degrees). Halogen sigma hole - aromatic ring interactions are detected if at least two ring atoms engage with the halogen atom and the ring normal is within 45 degrees of the halogen sigma hole bond vector.

Some geometric parameters (e.g. H-bond donor or acceptor saturation, and happiness of hydrophobic atoms) can be correctly evaluated only by considering their buriedness, i.e. accessibility for interaction with the bulk water. Buriedness is calculated for each ligand and protein atom. First, a grid of points (1.0 Å spacing) is generated around the ligand extending 12 Å ligand extremes in each Cartesian axis direction (-x, +x, -y, +y, -z/+z). Next, all grid points within the distance below 3.0 Å from either protein or ligand atoms are deleted. Finally, the remaining points are analyzed for their vicinity to ligand and protein. Atoms appearing within the threshold value of 6.0 Å from any grid point are marked as not buried.

2.4.3 Attention-based prediction model



Affinity
Log variance
Mean interaction fingerprint
Feed forward
Interaction fingerprint
Masked sum
Transformer Encoder (cross-attention)
Transformer Encoder (self-attention)
Feed forward
Linear layer
Po-sco ligand fingerprint
Residue embedding
Po-sco interacting residues
Po-sco interaction fingerprint

Figure 2.3: The architecture of the prediction model. Blue: Output obtained from the po-sco interaction analyzer. Yellow: Neural network layers. Green: Usable model outputs. Purple: Only used for training to increase the conservation of interaction information.

The po-sco prediction model uses three outputs from the po-sco interaction analyzer as input. The first output is the type of interacting residues, the second output is the interaction fingerprint between protein residues (including waters and metals) and the ligand, and the third output is the ligand “happiness” fingerprint describing the level of saturation of interaction partners in the ligand molecule by the protein counterparts. Here, we describe the architecture of the prediction model based on po-sco (Figure 2.3).

The type of interacting residues is first embedded in an 8-dimensional space and concatenated with the corresponding interaction fingerprint. This combined input is then passed through a linear layer to embed the data in a 64-dimensional space. The embedded data are then used to compute self-attention with a transformer encoder layer.

The ligand fingerprint is also embedded into a 64-dimensional space using a fully connected layer. The embedded ligand is then used to calculate cross-attention with the encoded residue information using a transformer encoder layer.

To conserve information on the interactions in the complex, we added the encoded residue information to the output from protein-ligand cross-attention. Since the output from the cross-attention layer is a vector, whereas the encoded residue information is a matrix, we applied a masked sum on the encoded residue information over the first dimension of the matrix. The mask is used to mask out padding elements in samples of unequal length. The output of this sum operation can be used as a non-binary single vector interaction fingerprint describing the protein-ligand complex. This vector could also be used for down-stream predictions other than the binding affinity.

The generated interaction fingerprint is then passed through a feed forward network to predict the affinity, the log variance of the affinity, and the mean po-sco interaction fingerprint. The predicted affinity and log variance are used to train the model using maximum likelihood estimation. In addition to the maximum likelihood estimation, the mean po-sco interaction fingerprint can be used to train the reconstruction of the interactions, which is intended to increase the amount of information (about the interactions in the complex) stored in the predicted interaction fingerprint. This reconstruction is trained using a mean squared error loss. The complete loss function is therefore a combination of a negative log likelihood loss and a mean squared error loss:

$$L(\hat{y}, v, \hat{p}, y, p) = NLL(\hat{y}, \sqrt{e^{v}}, y) + MSE(\hat{p}, p) \tag{2.2}$$

Where $\hat{y}$ is the predicted affinity, y is the true affinity, v is the predicted log variance of the affinity, $\hat{p}$ is the predicted mean po-sco interaction fingerprint, and p is the true mean po-sco interaction fingerprint. The negative log likelihood loss is defined as:

$$NLL(\hat{y}, \sigma, y) = -\sum_{i=1}^{N} \log\left(P(\hat{y}_i; y_i, \sigma)\right) \tag{2.3}$$

where $P(\hat{y}; y, \sigma)$ is the probability density function, i.e. the probability of finding $\hat{y}$ under a normal distribution given by the mean y and standard deviation σ, defined as:

$$P(\hat{y}; y, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y - \hat{y})^2}{2\sigma^2}\right) \tag{2.4}$$

The mean squared error loss is defined as:

$$MSE(\hat{p}, p) = \frac{1}{N}\sum_{i=1}^{N}(p_i - \hat{p}_i)^2 \tag{2.5}$$

The model was trained in PyTorch using an Adam optimizer[73]. The model was trained with 2 transformer encoder layers for the self- and cross-attention each, an embedding dimension of 64, 2 heads per attention layer, a batch size of 64, a learning rate of $1e^{-4}$, and a dropout rate of 0.15. All models were trained for 1000 epochs, and the epoch with the lowest validation loss was used to assess the model's performance (on the test set). All models were trained with the same hyperparameters where applicable.

Ablation study

In order to test the contribution of the ligand fingerprint calculated by posco, we performed an ablation study in which the ligand fingerprint was not used in model training. In this model, the predicted interaction fingerprint was calculated directly as the masked sum of the encoded residue information. Thus, there was no cross-attention with the ligand. The rest of the model remained the same.

Model based on PLIP

Since PLIP is commonly used to calculate protein-ligand interactions, we trained a model using the information obtained from PLIP. In our tests, we used PLIP version 2.2.2. The interactions calculated using PLIP can be found in Table 2.1.

The architecture of the model used with PLIP followed that of the one used in the ablation study, because no ligand fingerprint was used.

Model based on ProLIF

We trained an additional model using ProLIF version 1.1.0, one of the other tools commonly used for the analysis of protein-ligand interactions. The interaction types used from ProLIF can be found in Table 2.1. Like the PLIP-based model, the architecture of this model followed that in the ablation study without ligand information.

2.4.4 Real-world example

We downloaded compounds with known activity toward the human 5HT 1B receptor (Uniprot ID P28222) from PubChem[74]. The obtained substances were further filtered to exclude all samples that do not have an activity type of Ki or Kd, that do not have an exact affinity value, or that do not have unambiguous stereochemistry. This left us with a total of 726 compounds. Before docking, all compounds were prepared with LigPrep to generate protonation states at pH 7.4[75].

The compounds were docked with smina against the PDB IDs 4IAQ, 5V54, and 6G79 and with Glide against the PDB IDs 5V54, 6G79, and 7C61. For each compound, up to 9 poses were generated. For Glide, the Glide SP protocol was used, and the binding site was defined based on the location of the co-crystallized ligand. For smina, the binding site was defined based on the co-crystallized ligands with a default buffer of 4 Å on all six sides and the compounds were docked with an exhaustiveness of 16 and a seed of 42.

All protein structures were prepared by adding explicit hydrogen atoms, assigning bond orders, converting selenomethionines to methionines, filling in missing side chains and loops, generating protonation states at pH 7.4, optimizing the water network, and performing restrained minimization with an RMSD cutoff at 0.3 Å. All preparation was carried out in Schrödinger's Maestro using the protein preparation wizard[65].

The generated docking poses were re-scored with gnina and three trained po-sco models. Gnina was used with the `-score_only` flag, 2 CNN rotations, a seed of 42, and an exhaustiveness of 16. The best predicted score for all poses of a compound was used to calculate the performance statistics.

2.5 Conclusion

Recently, large language models trained on large amounts of data have gained increasing attention[40,76,77]. In this work, we show that the quality of the data used to train a model is more important than the quantity when it comes to the prediction of binding affinities based on molecular interactions. We demonstrate how overlaps between the training, validation, and test set, specifically in the PDBbind dataset, can lead to severe overestimation of model performance. We also suspect that many recently published models for binding affinity prediction may suffer from such a bias. In this regard, we urge researchers to critically assess the data used to train deep learning models and to avoid bias introduced by poor dataset splits. We also suggest working with smaller datasets of very high quality rather than large datasets of poor quality.

Furthermore, we introduced po-sco, a sophisticated tool for the analysis of protein-ligand complexes that provides a great level of detail. We showed how the high content of information extracted by po-sco benefits deep learning models. A transformer-based model that utilized analyses from po-sco was more successful in predicting binding affinities than models that were trained with data from PLIP and ProLIF. We propose that a very detailed and physics-based analysis of protein-ligand interactions allows deep learning models to better learn from structural data and improves downstream predictions.

Finally, we promote the use of structure-agnostic models, i.e. models that do not explicitly know the structure of the ligand or binding site. In this way, the risk of creating biased models can be reduced to a minimum.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of two RTX A5000 GPUs used for this research.

2.6 Appendix

Table A2.1: Comparison of the PLIP-based prediction model using different splittings of the PDBbind dataset. The performance is shown only for the external test set. The best performances are highlighted in bold.

Model	Dataset	Split	PCC	MUE	RMSE
Model 1	Complete	Classical	**0.48**	**2.05**	**2.56**
Model 2	Complete	Diverse	0.18	2.14	2.68
Model 3	Refined	Diverse	0.37	2.20	2.71

Table A2.2: Comparison of the ProLIF-based prediction model using different splittings of the PDBbind dataset. The performance is shown only for the external test set. The best performances are highlighted in bold.

Model	Dataset	Split	PCC	MUE	RMSE
Model 1	Complete	Classical	**0.57**	**1.91**	**2.41**
Model 2	Complete	Diverse	-0.03	2.16	2.71
Model 3	Refined	Diverse	0.53	2.03	2.48

Table A2.3: Weight factors for the calculation of hydrophobic contributions. The weight w of a solute x is calculated as $w = \frac{\Delta G_s(x)}{\Delta G_s(\text{Me-benzene})}$ where ΔG_s is the free energy of solvation of solute x in the organic solvent hexadecane and Me-benzene is toluene.

Solute	$\Delta G_s(hexadecane)$	weight
Benzene	-3.80	0.837
Me-benzene	-4.54	1.000
F-benzene	-4.03	0.888
Cl-benzene	-4.99	1.099
Br-benzene	-5.51	1.214
I-benzene	-6.25	1.377
SH-benzene	-5.61	1.236

References

[1] Manuel S. Sellner, Markus A. Lill, and Martin Smieško. Quality matters: Deep learning-based analysis of protein-ligand interactions with focus on avoiding bias. *bioRxiv*, page 2023.11.13.566916, 11 2023.

[2] Zhi Jin, Tingfang Wu, Taoning Chen, Deng Pan, Xuejiao Wang, Jingxin Xie, Lijun Quan, and Qiang Lyu. CAPLA: improved prediction of protein–ligand binding affinity by a deep learning approach based on a cross-attention mechanism. *Bioinformatics*, 39(2), 2 2023.

[3] Derek Jones, Hyojin Kim, Xiaohua Zhang, Adam Zemla, Garrett Stevenson, William D. Bennett, Daniel Kirshner, Sergio Wong, Felice Lightstone, and Jonathan E. Allen. Improved Protein-ligand Binding Affinity Prediction with Structure-Based Deep Fusion Inference. *Journal of Chemical Information and Modeling*, 61(4):1583–1592, 5 2021.

[4] Yuliang Gu, Xiangzhou Zhang, Anqi Xu, Weiqi Chen, Kang Liu, Lijuan Wu, Shenglong Mo, Yong Hu, Mei Liu, and Qichao Luo. Protein–ligand binding affinity prediction with edge awareness and supervised attention. *iScience*, 26(1):105892, 1 2023.

[5] Huiwen Wang, Haoquan Liu, Shangbo Ning, Chengwei Zeng, and Yunjie Zhao. DLSSAffinity: protein–ligand binding affinity prediction via a deep learning model. *Physical Chemistry Chemical Physics*, 24(17):10124–10133, 5 2022.

[6] Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1):1–20, 12 2021.

[7] Hiraku Oshima, Suyong Re, and Yuji Sugita. Prediction of protein-ligand binding pose and affinity using the gREST+FEP method. *Journal of Chemical Information and Modeling*, 60(11):5382–5394, 11 2020.

[8] Himanshu Goel, Anthony Hazel, Vincent D. Ustach, Sunhwan Jo, Wenbo Yu, and Alexander D. MacKerell. Rapid and accurate estimation of protein–ligand relative binding affinities using site-identification by ligand competitive saturation. *Chemical Science*, 12(25):8844–8858, 7 2021.

[9] Gregory A. Ross, Chao Lu, Guido Scarabelli, Steven K. Albanese, Evelyne Houang, Robert Abel, Edward D. Harder, and Lingle Wang. The maximal and current accuracy of rigorous protein-ligand binding free energy calculations. *Communications Chemistry*, 6(1):222, 10 2023.

[10] Vivek Govind Kumar, Adithya Polasa, Shilpi Agrawal, Thallapuranam Krishnaswamy Suresh Kumar, and Mahmoud Moradi. Binding affinity estimation from restrained umbrella sampling simulations. *Nature Computational Science*, 3(1):59–70, 12 2022.

[11] Y. Khalak, G. Tresadern, M. Aldeghi, H. M. Baumann, D. L. Mobley, B. L. de Groot, and V. Gapsys. Alchemical absolute protein–ligand binding free energies for drug design. *Chemical Science*, 12(41):13958–13971, 10 2021.

[12] Son Tung Ngo, Khanh B. Vu, Le Minh Bui, and Van V. Vu. Effective Estimation of Ligand-Binding Affinity Using Biased Sampling Method. *ACS Omega*, 4(2):3887–3893, 2 2019.

[13] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K.

Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 3 2004.

[14] Oleg Trott and Arthur J. Olson. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry*, 31(2):455–461, 1 2010.

[15] Martin Smieško. DOLINA - Docking based on a local induced-fit algorithm: Application toward small-molecule binding to nuclear receptors. *Journal of Chemical Information and Modeling*, 53(6):1415–1423, 6 2013.

[16] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 8 2013.

[17] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, 5 2016.

[18] Tatu Pantsar and Antti Poso. Binding Affinity via Docking: Fact and Fiction. *Molecules*, 23(8):1899, 7 2018.

[19] Tiejun Cheng, Xun Li, Yan Li, Zhihai Liu, and Renxiao Wang. Comparative Assessment of Scoring Functions on a Diverse Test Set. *Journal of Chemical Information and Modeling*, 49(4):1079–1093, 4 2009.

[20] Yan Li, Li Han, Zhihai Liu, and Renxiao Wang. Comparative Assessment of Scoring Functions on an Updated Benchmark: 2. Evaluation Methods and General Results. *Journal of Chemical Information and Modeling*, 54(6):1717–1736, 6 2014.

[21] Nataraj S. Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical Reviews*, 9(2):91–102, 4 2017.

[22] Robert Abel, Lingle Wang, Edward D. Harder, B. J. Berne, and Richard A. Friesner. Advancing Drug Discovery through Enhanced Free Energy Calculations. *Accounts of Chemical Research*, 50(7):1625–1632, 7 2017.

[23] Dibyendu Mondal, Jacob Florian, and Arieh Warshel. Exploring the Effectiveness of Binding Free Energy Calculations. *The Journal of Physical Chemistry B*, 123(42):8910–8915, 10 2019.

[24] Joe Z. Wu, Solmaz Azimi, Sheenam Khuttan, Nanjie Deng, and Emilio Gallicchio. Alchemical Transfer Approach to Absolute Binding Free Energy Estimation. *Journal of Chemical Theory and Computation*, 17(6):3309–3319, 6 2021.

[25] Solmaz Azimi, Sheenam Khuttan, Joe Z. Wu, Rajat K. Pal, and Emilio Gallicchio. Relative Binding Free Energy Calculations for Ligands with Diverse Scaffolds with the Alchemical Transfer Method. *Journal of Chemical Information and Modeling*, 62(2):309–323, 1 2022.

[26] Ercheng Wang, Huiyong Sun, Junmei Wang, Zhe Wang, Hui Liu, John Z.H. Zhang, and Tingjun Hou. End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design. *Chemical Reviews*, 119(16):9478–9508, 8 2019.

[27] Susu Zhong, Kaifang Huang, Song Luo, Shuheng Dong, and Lili Duan. Improving the performance of the MM/PBSA and MM/GBSA methods in recognizing the native structure of the Bcl-2 family using the interaction entropy method. *Physical Chemistry Chemical Physics*, 22(7):4240–4251, 2 2020.

[28] Edward King, Erick Aitchison, Han Li, and Ray Luo. Recent Developments in Free Energy Calculations for Drug Discovery. *Frontiers in Molecular Biosciences*, 8:775, 8 2021.

[29] Lingling Zhao, Yan Zhu, Junjie Wang, Naifeng Wen, Chunyu Wang, and Liang Cheng. A brief review of protein–ligand interaction prediction. *Computational and Structural Biotechnology Journal*, 20:2831–2838, 1 2022.

[30] Asad Ahmed, Bhavika Mam, and Ramanathan Sowdhamini. DEELIG: A Deep Learning Approach to Predict Protein-Ligand Binding Affinity. *Bioinformatics and Biology Insights*, 15:117793222110303, 1 2021.

[31] Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. ML-DTI: Mutual Learning Mechanism for Interpretable Drug–Target Interaction Prediction. *The Journal of Physical Chemistry Letters*, 12(17):4247–4261, 5 2021.

[32] Amr H. Mahmoud, Matthew R. Masters, Ying Yang, and Markus A. Lill. Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. *Communications Chemistry*, 3(1):19, 2 2020.

[33] Ziduo Yang, Weihe Zhong, Lu Zhao, and Calvin Yu-Chian Chen. MGraphDTA: deep multiscale graph neural network for explainable drug–target binding affinity prediction. *Chemical Science*, 13(3):816–833, 1 2022.

[34] Shrimon Mukherjee, Madhusudan Ghosh, and Partha Basuchowdhuri. DeepGLSTM: Deep Graph Convolutional Network and LSTM based approach for predicting drug-target binding affinity. In *Proceedings of the 2022 SIAM International Conference on Data Mining (SDM)*, pages 729–737. Society for Industrial and Applied Mathematics, Philadelphia, PA, 1 2022.

[35] Shuke Zhang, Yanzhao Jin, Tianmeng Liu, Qi Wang, Zhaohui Zhang, Shuliang Zhao, and Bo Shan. SS-GNN: A Simple-Structured Graph Neural Network for Affinity Prediction. *arXiv*, 5 2022.

[36] Yuan Jin, Jiarui Lu, Runhan Shi, and Yang Yang. EmbedDTI: Enhancing the Molecular Representations via Sequence Embedding and Graph Convolutional Network for the Prediction of Drug-Target Interaction. *Biomolecules*, 11(12):1783, 11 2021.

[37] Carter Knutson, Mridula Bontha, Jenna A. Bilbrey, and Neeraj Kumar. Decoding the protein–ligand interactions using parallel graph neural networks. *Scientific Reports*, 12(1):7624, 5 2022.

[38] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, 5 2022.

[39] Ashish Vaswani, Google Brain, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. *Advances in Neural Information Processing Systems*, 30, 2017.

[40] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell,

Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 2020-Decem, 5 2020.

[41] Qiming Zhang, Yufei Xu, Jing Zhang, and Dacheng Tao. ViTAEv2: Vision Transformer Advanced by Exploring Inductive Bias for Image Recognition and Beyond. *International Journal of Computer Vision*, 131(5):1141–1162, 5 2023.

[42] Fahad Shamshad, Salman Khan, Syed Waqas Zamir, Muhammad Haris Khan, Munawar Hayat, Fahad Shahbaz Khan, and Huazhu Fu. Transformers in medical imaging: A survey. *Medical Image Analysis*, page 102802, 4 2023.

[43] Abel Chandra, Laura Tünnermann, Tommy Löfstedt, and Regina Gratz. Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12, 1 2023.

[44] Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *Journal of the American Chemical Society*, 145(5):2958–2967, 2 2023.

[45] Nelson R.C. Monteiro, José L. Oliveira, and Joel P. Arrais. DTITR: End-to-end drug–target binding affinity prediction with transformers. *Computers in Biology and Medicine*, 147:105772, 8 2022.

[46] Ingoo Lee and Hojung Nam. Sequence-based prediction of protein binding regions and drug–target interactions. *Journal of Cheminformatics*, 14(1):5, 12 2022.

[47] Frankie J. Fan and Yun Shi. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Bioorganic & Medicinal Chemistry*, 72:117003, 10 2022.

[48] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *Journal of Medicinal Chemistry*, 65(11):7946–7958, 6 2022.

[49] Caterina Bissantz, Bernd Kuhn, and Martin Stahl. A Medicinal Chemist's Guide to Molecular Interactions. *Journal of Medicinal Chemistry*, 53(14):5061–5084, 7 2010.

[50] Danislav S. Spassov, Mariyana Atanasova, and Irini Doytchinova. A role of salt bridges in mediating drug potency: A lesson from the N-myristoyltransferase inhibitors. *Frontiers in Molecular Biosciences*, 9:1405, 1 2023.

[51] Renato Ferreira de Freitas and Matthieu Schapira. A systematic analysis of atomic protein–ligand interactions in the PDB. *MedChemComm*, 8(10):1970–1981, 10 2017.

[52] Emilio M. Pérez and Nazario Martín. π–π interactions in carbon nanostructures. *Chemical Society Reviews*, 44(18):6425–6433, 9 2015.

[53] Gabriella Cavallo, Pierangelo Metrangolo, Roberto Milani, Tullio Pilati, Arri Priimagi, Giuseppe Resnati, and Giancarlo Terraneo. The Halogen Bond. *Chemical Reviews*, 116(4):2478–2601, 2 2016.

[54] Paulo J. Costa. The halogen bond: Nature and applications. *Physical Sciences Reviews*, 2(11), 11 2017.

[55] Sebastian Salentin, Sven Schreiber, V. Joachim Haupt, Melissa F. Adasme, and Michael Schroeder. PLIP: fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, 43(W1):W443–W447, 7 2015.

[56] Cédric Bouysset and Sébastien Fiorucci. ProLIF: a library to encode molecular interactions as fingerprints. *Journal of Cheminformatics*, 13(1):72, 12 2021.

[57] Renxiao Wang, Xueliang Fang, Yipin Lu, and Shaomeng Wang. The PDBbind Database: Collection of Binding Affinities for Protein-Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980, 6 2004.

[58] Maha A. Thafar, Mona Alshahrani, Somayah Albaradei, Takashi Gojobori, Magbubah Essack, and Xin Gao. Affinity2Vec: drug-target binding affinity prediction through representation learning, graph mining, and machine learning. *Scientific Reports*, 12(1):4751, 3 2022.

[59] Sangmin Seo, Jonghwan Choi, Sanghyun Park, and Jaegyoon Ahn. Binding affinity prediction for protein–ligand complex using deep attention mechanism based on intermolecular interactions. *BMC Bioinformatics*, 22(1):542, 11 2021.

[60] Surendra Kumar and Mi-hyun Kim. SMPLIP-Score: predicting ligand binding affinity from simple and interpretable on-the-fly interaction fingerprint pattern descriptors. *Journal of Cheminformatics*, 13(1):28, 3 2021.

[61] Debby D. Wang and Moon-Tong Chan. Protein-ligand binding affinity prediction based on profiles of intermolecular contacts. *Computational and Structural Biotechnology Journal*, 20:1088–1096, 1 2022.

[62] John E. Ladbury. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology*, 3(12):973–980, 12 1996.

[63] Manuela Maurer and Chris Oostenbrink. Water in protein hydration and ligand recognition. *Journal of Molecular Recognition*, 32(12):e2810, 12 2019.

[64] Joel Wahl and Martin Smieško. Thermodynamic Insight into the Effects of Water Displacement and Rearrangement upon Ligand Modifications using Molecular Dynamics Simulations. *ChemMedChem*, 13(13):1325–1335, 7 2018.

[65] LLC Schrödinger. Schrödinger Release 2021-1: Maestro, 2021.

[66] LLC Schrödinger. Schrödinger Release 2021-1: Epik, 2021.

[67] Jeremy R. Greenwood, David Calkins, Arron P. Sullivan, and John C. Shelley. Towards the comprehensive, rapid, and accurate prediction of the favorable tautomeric states of drug-like molecules in aqueous solution. *Journal of Computer-Aided Molecular Design*, 24(6-7):591–604, 6 2010.

[68] John C. Shelley, Anuradha Cholleti, Leah L. Frye, Jeremy R. Greenwood, Mathew R. Timlin, and Makoto Uchimaya. Epik: a software program for pK a prediction and protonation state generation for drug-like molecules. *Journal of Computer-Aided Molecular Design*, 21(12):681–691, 12 2007.

[69] Angelo Vedani. YETI: An interactive molecular mechanics program for small-molecule protein complexes. *Journal of Computational Chemistry*, 9(3):269–280, 4 1988.

[70] Angelo Vedani and David W. Huhta. A new force field for modeling metalloproteins. *Journal of the American Chemical Society*, 112(12):4759–4767, 6 1990.

[71] Aleksandr V. Marenich, Casey P. Kelly, Jason D. Thompson, Gregory D. Hawkins, Candee C. Chambers, David J. Giesen, Paul Winget, Christopher J. Cramer, and Donald G. Truhlar. Minnesota Solvation Database (MNSOL) version 2012, 2020. https://doi.org/10.13020/3eks-j059.

[72] Jianing Li, Robert Abel, Kai Zhu, Yixiang Cao, Suwen Zhao, and Richard A. Friesner. The VSGB 2.0 model: A next generation energy model for high resolution protein structure modeling. *Proteins: Structure, Function, and Bioinformatics*, 79(10):2794–2812, 10 2011.

[73] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury Google, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf Xamla, Edward Yang, Zach Devito, Martin Raison Nabla, Alykhan Tejani, Sasank Chilamkurthy, Qure Ai, Benoit Steiner, Lu Fang Facebook, Junjie Bai Facebook, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. 2019.

[74] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 1 2023.

[75] LLC Schrödinger. Schrödinger Release 2021-1: Ligprep, 2021.

[76] OpenAI. GPT-4 Technical Report. 3 2023.

[77] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. 2 2023.

The most dangerous phrase in the English language is 'we have always done it this way.'

Grace Hopper

3

Automated Prediction of Off-Target Binding of Small Molecules

The safety of chemicals, in a pharmaceutical or environmental setting, is of paramount importance. Since adverse effects in humans often begin with a small molecule binding to a protein structure (an off-target), we developed an automated platform for the prediction of these molecular initiating events.

PanScreen, our prediction platform, is a modular online platform that is freely available to the public. It was specifically designed to make the implementation of new off-targets and new methods as easy as possible. The PanScreen online platform essentially consists of three parts, the front-end, the back-end, and a bridging database.

The front-end was designed to be simple and user-friendly. It is built using of a combination of the Python-based Django web framework and an nginx webserver. For easy deployment, both parts are contained in a Docker container. This allows maximum flexibility with minimum effort.

The back-end is the core of PanScreen. We developed it as a portable Python package that can be installed with a single command. This allows for easy distribution and management. This package can be used not only to predict binding to implemented off-targets, but also to automatically train new models when providing the necessary data. To implement a new off-target, it is therefore only necessary to prepare an ensemble of crystal structures and compile a set of ligands with known binding affinities to the desired off-target. The docking, hyperparameter search, and training and integration of the models are then performed automatically by the package. To allow for easy installation on different machines, we deploy the back-end in a Docker container, together with a job listener connecting to the bridging database.

The bridging database receives jobs from the front-end. These jobs can then be accessed by the back-end for processing. Once a job is completed, its results are being fed back to the database, which serves them to the front-end again.

The following article describes the detailed methods and performance of PanScreen and was published as a preprint.[1] We hope that our tools can complement existing methods for chemical safety assessment and, ultimately, help reduce the use of animals in toxicology.

PanScreen: A Comprehensive Approach to Off-Target Liability Assessment

Manuel S. Sellner[1,2]
manuel.sellner@unibas.ch

Markus A. Lill[1,2]
markus.lill@unibas.ch

Martin Smieško[1,2*]
martin.smiesko@unibas.ch

[1] Computational Pharmacy, Department of Pharmaceutical Sciences, University of Basel
[2] SIB Swiss Institute of Bioinformatics
[*] Corresponding author

November 17, 2023

3.1 Abstract

Drug development projects are getting increasingly more expensive while their success rate is stagnating. Safety issues attributed to off-target binding represent a major reason for the failure of new drugs. Besides desired on-target binding, small molecules may interact with off-targets, triggering adverse effects. Therefore, the development of novel methods for early recognition of such issues that are resource-efficient and cost-effective becomes vital. Here, we introduce PanScreen, an online platform for the automated assessment of off-target liabilities. PanScreen combines structure-based modeling techniques with state-of-the-art deep learning methods to not only predict accurate binding affinities but also give insight into potential modes of action. We show that the predictions are approaching experimental accuracy found in public datasets and that the same technology can also be used for other research areas, such as drug repurposing. Such fast and inexpensive methods allow researchers to test

not only drug candidates, but all small molecules that might come into contact with a human organism for potential safety concerns very early in the development process. PanScreen is publicly available at www.panscreen.ch.

3.2 Introduction

Chemicals are omnipresent in the environment due to the use of drugs, pesticides, fertilizers, combustion engines, waste water, and industrial by-products. Humans are in constant contact with their environment, leading to unevitable exposure to a wide variety of chemicals. The most common sources of human exposure to chemicals include food, air, personal care and pharmaceutical products, clothing, and household products[2]. Environmental exposure to chemicals has long been known to have adverse effects on humans, such as various cancers, infertility, other reproductive disorders, respiratory diseases, and allergic reactions[3–5]. Pharmaceutical products are specifically made to be ingested, injected, inhaled, or topically applied by humans. Thus, it is especially important to ensure their safety by recognizing and avoiding toxic effects.

Most small molecule drugs are designed to interact with one or more proteins in the human body by modulating their physiological behavior[6–8]. Sometimes, modulating the target protein inevitably leads to adverse effects e.g. by the disruption of essential cellular pathways. This effect is known as on-target toxicity[9,10]. Often, however, drugs not only interact with their intended target protein, but also with other so-called off-targets, leading to possible side effects[8]. Such toxicities are estimated to account for up to a third of the attrition of drugs[11,12]. In some exceptional cases, off-target binding can even be beneficial[13].

Investigation and identification of potential off-target toxicities is therefore highly important. This is true not only for the pharmaceutical industry but also for environmental chemicals that may end up in the human body. Experimental off-target profiling is usually expensive, slow, and labour- and resource-intensive[14–18]. On the other hand, computational methods are cheap and fast. The immense increase of available computing power during the past decade combined with the continuous improvement of computational methods has enabled in-silico tools to become a viable alternative to experimental testing. It is therefore not surprising that several tools aiming to predict off-target toxicities have been developed in recent years[19–24].

In-silico methods are used to predict not only off-target toxicities but also various endpoints. Some tools predict assay outcomes such as mutagenicity or skin sensitization[25–28], others predict clinical outcome[29,30]. In off-target liability prediction, the underlying mechanism is usually based on undesired interactions between a small molecule and an off-target. This falls within the scope of drug-target interaction prediction[31–34]. In a toxicology setting, drug–off-target interactions usually represent molecular initiating events in an adverse outcome pathway[35]. However, drug-target interaction prediction is not restricted to toxicology and can also be used in drug development.

3.2.1 Ligand-based methods

One of the principles frequently applied in drug discovery, as well as predictive toxicology, is that chemically similar molecules exhibit similar properties. Similarity can thereby be defined in various ways, such as 2D similarity or 3D shape overlap[36,37].

The advantage of ligand-based methods is that they only need a seed (or template) ligand structure as input. This allows them to be used in most drug development projects with at least one known initial hit. Additionally, ligand-based methods such as similarity searches are usually computationally inexpensive, leading to fast results. For these reasons, it is no surprise that ligand-based methods are routinely used in off-target prediction and drug development in general[38–41].

Although these methods work well in many cases, they have some inherent disadvantages. Two-dimensional approaches may be limited to a particular molecular scaffold, which can lead to a similarity search missing out on hits with a different structure. Furthermore, two molecules can be highly similar in 2D structure and/or 3D shape but still exhibit completely different activities to a given target. This phenomenon is known as an activity cliff[42,43]. In such cases, more detailed analyses are necessary to accurately assess the potency of a molecule.

3.2.2 Structure-based methods

In contrast to ligand-based methods, structure-based methods require the 3D structure or the primary sequence of the protein to be known. Structural data, especially well-resolved experimentally determined complexes, allow the protein-ligand complementarity to be decoded in very fine detail, enabling the methods to overcome the drawbacks of ligand-based methods[44,45]. However, because of the generally higher computational cost of structure-based methods, they tend to be much slower than ligand-based methods.

One of the most commonly used methods in structure-based drug development is molecular docking, in which a small molecule is placed into the binding site of a protein while optimizing the molecular interactions between the two entities[46–49]. In docking, proteins are often treated as rigid bodies, while the ligands are allowed to be flexible. This has the disadvantage that induced fit effects cannot be captured and the result of the docking depends on the input conformation of the protein. While it is possible to allow the side chains (or even the backbone) of the binding site residues of the protein to be flexible, this introduces many more degrees of freedom, leading to an explosion of possible combinations and therefore computational cost.

One way of tackling this problem is the use of ensemble docking[50–52]. In ensemble docking, a ligand is docked to an ensemble of protein structures, usually coming either from molecular dynamics simulations or experimentally determined (X-ray, cryo-EM, or NMR) structures. Usually, the ensemble is selected to represent different conformational states of the protein binding site. This approach implicitly accounts for the protein flexibility while minimizing computational cost.

Our group has previously developed VirtualToxLab, a platform accessible via a simple Java application for the automated assessment of the toxic potential of small molecules[23,24]. Relying on the concepts of structure-based modeling, it features a portfolio of 16 well-known and comprehensively prepared off-targets, an induced-fit-enabled docking program, and optimized scoring functions for each target. VirtualToxLab has been extensively used by academia, regulatory agencies, and industry partners for its predictions, especially for CYP450 enzymes and nuclear receptors.

3.2.3 Machine learning-based methods

In recent years, machine learning has emerged as a major challenger to classical ligand- and structure-based methods[53–55]. Many machine learning models such as random forest, support vector machine, or naive bayes have been developed to predict drug-target interactions[56–58]. With increasing computational power, deep learning models have become more popular. With the right architecture, deep learning models have the ability to outperform classical machine learning models[59]. Thus, many deep learning-based models have recently emerged that aim to predict drug-target interactions[60–64]. A popular method to improve the performance of deep learning models is ensemble deep learning, in which an ensemble of models is trained with the goal of improving the generalizability of the combined ensemble[65,66].

Although deep learning models have great potential to substantially improve the predictive power of in silico tools, they are not trivial to train. The construction of the data set and the processing of input features are imperative for a robust and unbiased model. The incorrect handling of data sets and the use of too large molecular input feature vectors contaminated with irrelevant information have been shown to lead to an overestimation of model performance[67–69]. Therefore, it is essential to thoroughly evaluate all components of the model training process to create a reliable and accurate model.

3.2.4 Our contribution

In this work, we introduce PanScreen, an online platform for the prediction of off-target liability that is publicly available. Similar to its predecessor, VirtualToxLab, PanScreen features a portfolio of off-targets. All implemented off-targets are thoroughly prepared and validated. PanScreen applies an ensemble docking approach using multiple docking programs and processes the output with deep learning models. This not only allows accurate predictions of off-target interactions, but also provides structural insight into the potential mechanism of action. The platform presents a user-friendly web interface that allows easy access to researchers with various degrees of experience with in silico structure-based modeling. It is available free of charge for academic and non-commercial use.

Although the number of implemented off-targets is still limited, highly standardized processes of preparing protein structures and training deep learning models greatly facilitate the addition of new off-targets. We anticipate a rapid growth of the off-target portfolio in the very near future.

3.3 Results and Discussion

PanScreen is available as an online service at www.panscreen.ch. The web application was developed using the Django web framework and is served using an nginx webserver[70]. At the

Table 3.1: (Off-)targets currently implemented in PanScreen.

Uniprot ID	Name	Family
O60674	Tyrosine-protein kinase JAK2	Kinase
P03372	Estrogen receptor alpha	Nuclear receptor
P04150	Glucocorticoid receptor	Nuclear receptor
P07550	Beta-2 adrenergic receptor	GPCR
P10275	Androgen receptor	Nuclear receptor
P14416	Dopamine receptor D2	GPCR
P23458	Tyrosine-protein kinase JAK1	Kinase
P25103	Substance-P receptor	GPCR
P28222	5HT receptor 1B	GPCR
P37231	PPARγ	Nuclear receptor
P49286	Melatonin receptor 1B	GPCR
Q08499	Phosphodiesterase 4D	Hydrolase
Q92731	Estrogen receptor beta	Nuclear receptor
Q9Y233	Phosphodiesterase 10A	Hydrolase

time of publication of this article, the platform contained 14 implemented off-targets (see Table 3.1 for a complete overview). Each implemented off-target consists of an ensemble of thoroughly curated protein structures (see Section 3.4.1 for more information).

Users can upload query molecules in various, commonly used data formats (e.g., SDF, MOL2, or SMILES) and select the off-targets against which the submitted molecules should be screened. The uploaded molecule is converted to canonical SMILES using openbabel version 3.0.0[71]. The canonical SMILES format allows for the unique encoding of molecular structures while conserving stereochemistry and protonation states. The canonical SMILES are then stored in a PostgreSQL database that connects the front end and the back end. If the same molecule has already been processed before, the results are fetched from the database and provided to the user without need of re-running the simulations. In case the submitted molecule has not been processed before, the back-end reads the canonical SMILES from the database and starts processing it.

First, the canonical SMILES is used to generate a 3D conformation of the molecule using Schrödinger's LigPrep[72]. It is thereby up to the user whether the protonation states present in the input molecule should be preserved or whether protonation states at physi-

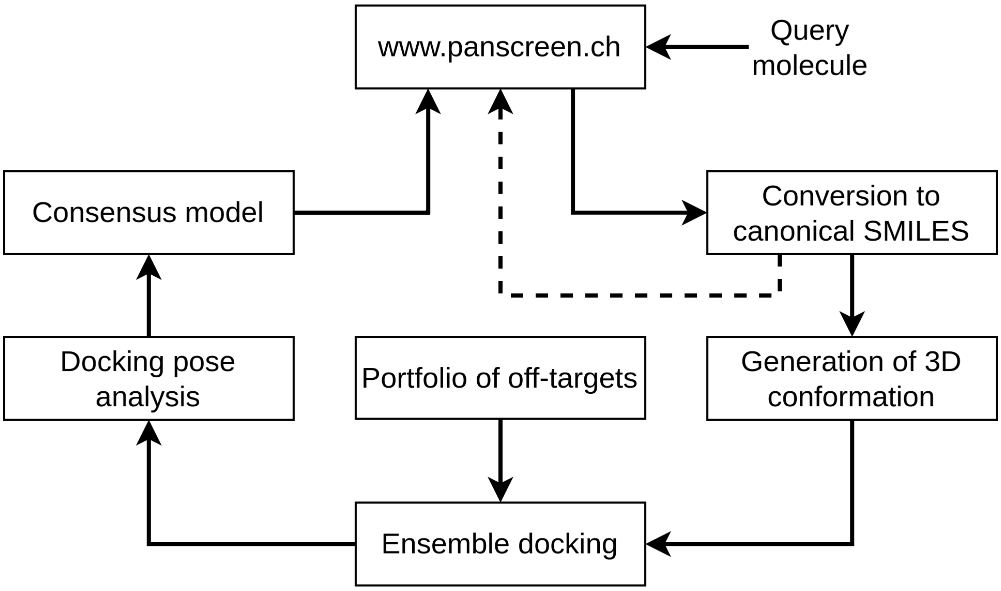


Figure 3.1: Flow of data in PanScreen. Once a compound is uploaded to PanScreen, it is converted to canonical SMILES and stored in a database. If the compound was processed before, its results are returned back to the user (dashed line). If the compound has not been processed before, it will be completely processed by the back end before the results are fed into the database and presented to the user.

ological pH should be automatically generated. The prepared molecule is then docked to the protein ensembles of the desired off-targets. Currently, PanScreen implements 3 different docking programs to generate poses and 2 programs to re-score and analyze the generated poses. Detailed information on this process can be found in Section 3.4.2. The information from molecular docking and the analysis of the generated poses are then fed into a consensus model (described in Section 3.4.3). For each implemented off-target, we developed a specialized consensus model. This was done to increase the performance of individual off-targets while avoiding the bias that could be introduced by providing the model with structural information of the protein[68]. Once all calculations are completed, the results are stored in the PostgreSQL database. Since the front-end is also connected to this database, the user will have immediate access to the results of the computations. An overview of the data flow in PanScreen is shown in Figure 3.1. The protein-ligand complexes generated by PanScreen can be viewed online and downloaded.

3.3.1 Performance analysis

The validation performance of all implemented models can be found in Table 3.2 (the same analysis for smina, Glide, LeDock, and gnina can be found in the Supporting Information in Tables A3.2, A3.3, A3.4, and A3.5, respectively). The Pearson correlation coefficient (PCC) was above 0.70 for all models (except PPARγ) with an average of 0.79. With exception of the melatonin receptor 1B and the estrogen receptor beta, all mean unsigned errors (MUE) were below 1.0 kcal/mol. The average MUE and root mean squared error (RMSE) were 0.83 kcal/mol and 1.12 kcal/mol, respectively. The area under the receiver operating characteristics curve (AUROC), calculated at an active/inactive threshold of 1.0 μM, was above 0.80 for all implemented off-targets with a mean of 0.89. This indicates very good performance for all models regardless of their protein family.

Table 3.2: Validation metrics of the implemented models. Shown are the Pearson correlation coefficient (PCC; higher is better), the mean unsigned error (MUE; lower is better), the root mean squared error (RMSE; lower is better), and the area under the receiver operating characteristics curve (AUROC; higher is better).

Protein name	PCC	MUE [kcal/mol]	RMSE [kcal/mol]	AUROC
Tyrosine-protein kinase JAK2	0.81	0.68	1.10	0.94
Estrogen receptor alpha	0.84	0.94	1.18	0.89
Glucocorticoid receptor	0.79	0.70	0.95	0.89
Beta-2 adrenergic receptor	0.79	0.92	1.17	0.91
Androgen receptor	0.81	0.79	1.03	0.88
Dopamine receptor D2	0.75	0.66	0.87	0.88
Tyrosine-protein kinase JAK1	0.81	0.55	0.85	0.94
Substance-P receptor	0.80	0.77	1.02	0.91
5HT receptor 1B	0.77	0.83	1.06	0.85
PPARγ	0.68	0.89	1.33	0.84
Melatonin receptor 1B	0.72	1.06	1.35	0.84
Phosphodiesterase 4D	0.80	0.95	1.36	0.84
Estrogen receptor beta	0.75	1.00	1.29	0.84
Phosphodiesterase 10A	0.80	0.88	1.18	0.93
Mean	0.79 ± 0.05	0.83 ± 0.15	1.12 ± 0.17	0.88 ± 0.04

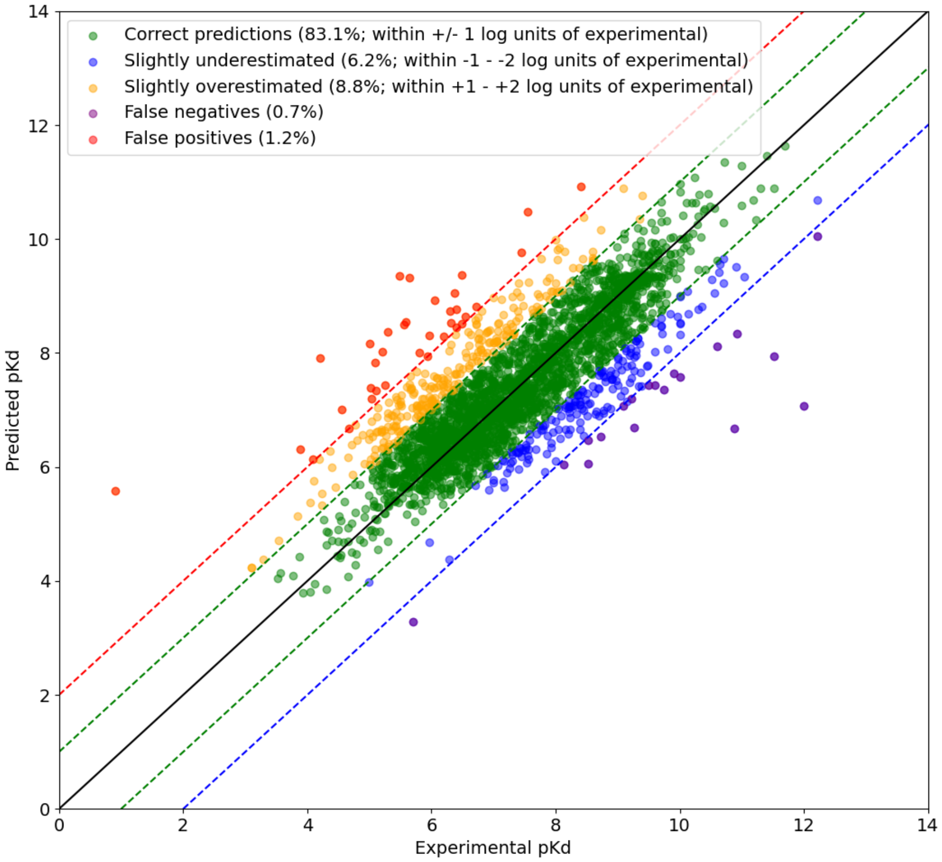


Figure 3.2: Correlation between predicted and experimental pKd values. The solid black line represents a perfect correlation. The green dashed lines represent a +/- 1 log unit deviation from the experimentally determined pKd values, the red and blue dashed lines represent a 2 log unit deviation from the experimental pKd values.

Plotting the predicted binding affinities against the experimentally determined affinities over all implemented off-targets (a total of more than 2800 predictions) revealed a PCC of 0.83 (see Figure 3.2). Moreover, more than 83% of the predictions were within 1 log unit of the true affinity. 15% of the predictions were between 1 and 2 log units from the experimentally determined binding affinity and only around 2% had a deviation of more than 2 log units.

To investigate the meaning of these numbers, we analyzed the experimental accuracy in the datasets used to train the models. For this, we investigated all the data points with measured

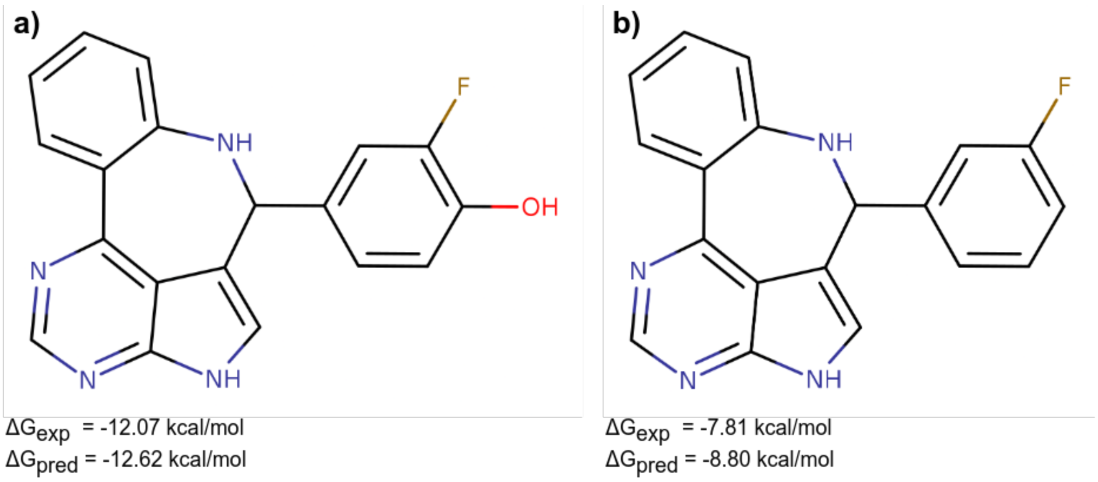

a)
F
NH
OH
N
N
NH
ΔG$_{exp}$ = -12.07 kcal/mol
ΔG$_{pred}$ = -12.62 kcal/mol
b)
F
NH
N
N
NH
ΔG$_{exp}$ = -7.81 kcal/mol
ΔG$_{pred}$ = -8.80 kcal/mol

Figure 3.3: Example for activity cliff. a) Potent inhibitor of the Janus kinase 2, b) Weak inhibitor of the Janus kinase 2.

K_i and K_d values for every off-target currently implemented in PanScreen. We filtered the data for compounds that have been tested at least twice (around 2600 individual compounds) and calculated the maximum spread between the individual measurements. We found that 2083 compounds (80%) had a spread of less than 1 log unit, 375 (14%) were spread between 1 and 2 log units and 135 (5%) had a spread of more than 2 log units. These findings align very well with the accuracy of our predictions. In fact, when considering only compounds that have been measured at least 4 times (a total of 305 different compounds), we found a median and mean spread of 1.2 and 2.7 log units, respectively. This shows that our predictions reach the accuracy found in publicly available experimental datasets.

Regarding the previously discussed shortcomings of the ligand-based methods, we performed an in-depth investigation of how the structure-based approach implemented in PanScreen copes with matched molecular pairs (MMPs)[73–76]. MMPs are pairs of highly similar molecules that differ in only a few atoms. In some cases, MMPs have very different activities (binding affinities) despite their high degree of structural similarity. One such example is shown in Figure 3.3 where the molecule in subfigure a) is a very potent inhibitor of the Janus kinase 2 whereas the molecule in subfigure b) inhibits the Janus kinase 2 only very weakly. This is a prime example of an activity cliff caused by the removal of the hydroxyl group in Figure 3.3 b). Importantly, these compounds were not used in the training set of the respective model. For this example, there is an experimentally determined $\Delta\Delta G$ of 4.26 kcal/mol. The predicted binding free energies of Panscreen were within 1.0 kcal/mol of the experimentally determined values for both molecules and the predicted $\Delta\Delta G$ was 3.82 kcal/mol.

Figure 3.4 depicts a comprehensive analysis of all MMPs found in the validation sets of the models for all implemented off-targets. Here, we defined MMPs as molecules with a Tanimoto similarity of at least 0.7. A confusion matrix containing the results can be found in Table A3.1. Of the 3466 identified MMPs that were not part of the training sets, 2926 (84.4%) had a predicted $\Delta\Delta G$ within ± 1.0 kcal of the experimental $\Delta\Delta G$. Only very few MMPs (38; 1.1%) were overestimated by the models. However, a total of 502 MMPs (14.5%) were underestimated, whereof 110 (3.2%) had a predicted $\Delta\Delta G$ that was more than 2 kcal/mol lower than the experimentally determined one.

In these cases, our models were not able to correctly predict the activity cliffs. Our analyses showed that in most of these MMPs, the docking programs were unable to correctly account for the key structural difference and thus distinguish between the two molecules. The docking scores for these MMPs usually had a $\Delta\Delta G$ of less than 1 kcal/mol and our models were not able to correct the predictions. Thus, these shortcomings are mainly due to limitations of the implemented docking programs.

3.3.2 Screening performance

In order to further evaluate the quality of the predictions of our models, we screened the compounds contained in the Drugbank (after excluding ions and fragments) against the es-

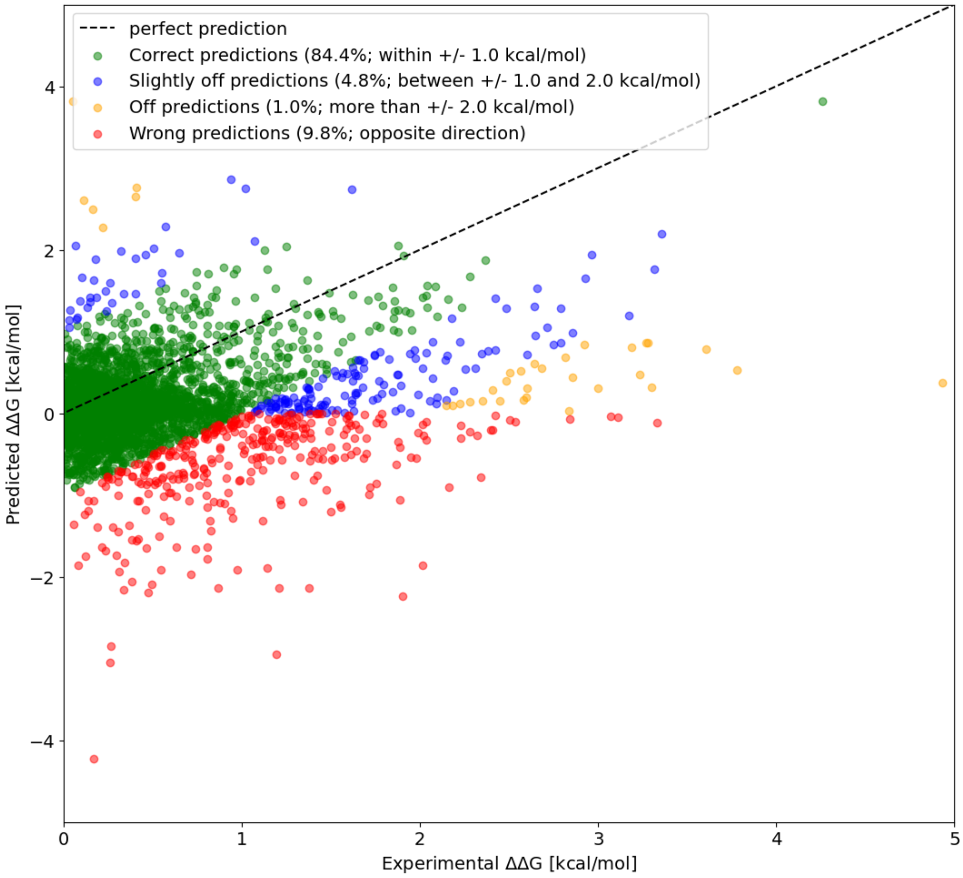


Figure 3.4: Analysis of matched molecular pairs. The black dashed line represents a perfect correlation between predicted and experimental $\Delta\Delta$G values.

trogen receptor alpha using our workflow[77]. We paid extra attention to only include compounds that were neither in the training nor in the validation set of the model. We filtered out all hits with an applicability score below 0.2 or with raised warning flags according to Section 3.4.3. The top 10 hits are shown in Table 3.3. The results showed that 5 of the top 10 hits have literature-confirmed activity on the estrogen receptor alpha. These compounds include steroids as well as non-steroidal structures. For the remaining 5 compounds of the top 10 hits, no reference for the activity at the estrogen receptor alpha could be found in the literature. We therefore analyzed the binding modes of these compounds and found that most of them could form reasonable interactions with the receptor, but most importantly had a good complementarity with the binding cavity (see Figures A3.1-A3.5 for the binding modes). This complementarity has been found to be an important factor in determining the quality of a binding pose[78]. Thus, there is a good chance that these compounds indeed bind to the estrogen receptor alpha, but experimental testing would be necessary to confirm these hits.

One of the compounds (DB16139) is under investigation for the treatment of schizophrenia by acting as a dopamine D1 agonist[79]. It has been shown that estrogen receptor modulators such as raloxifene also have some activity on dopamine receptors[80,81]. This indicates that there may be the possibility that compounds designed to bind to the dopamine receptor also interact with the estrogen receptor.

DB15449 (citarinostat) is a histone deacetylase inhibitor. It consists of a triphenylamine analog, a linker, and a zinc binding group (Figure A3.6). The triphenylamine analog has a shape similar to that of cyclofenil analogues and therefore may also possess the ability to bind to the estrogen receptor alpha. If this was the case, citarinostat could act as a histone deacetylase inhibitor and estrogen receptor modulator hybrid[82–84].

Table 3.3: Top 10 hits found by screening the Drugbank compounds against the estrogen receptor alpha. Only compounds that were neither in the training nor in the validation set used to train the model are shown.

Compound name	ΔG_{pred} [kcal/mol]	Confirmed ERα activity
DB06249	-13.45	Yes
DB08309	-12.75	n.a.
DB16139	-12.52	n.a.
DB01524	-12.05	Yes
DB02187	-12.00	Yes
DB00345	-11.97	n.a.
DB13866	-11.96	Yes
DB03882	-11.96	Yes
DB13591	-11.74	n.a.
DB15449	-11.74	n.a.

In total, the top 10 predicted hits were satisfactory with 5 confirmed estrogen receptor alpha modulators, two interesting compounds that could be compelling for further investigation, and several compounds with legitimate binding modes. This shows that the tested model has a very good enrichment of the top N hits with confirmed or plausible molecules. Therefore, we believe that PanScreen could be advantageously applied for use cases other than off-target assessment, e.g. drug repurposing.

3.4 Methods

3.4.1 Selection and preparation of protein structures

All protein structures implemented in PanScreen have been experimentally determined and computationally prepared. We identified off-targets based on their Uniprot ID and used the associated crystal structures listed on Uniprot as the starting position[85]. The obtained crystal structures were then manually assessed using their entry in the PDB[86]. Only structures with co-crystallized ligands were considered while excluding fragments. We checked for mutations in the vicinity of the binding site and visually inspected the electron densities of the binding site residues and the co-crystallized ligands. All crystal structures with non-covalently bound co-crystallized ligands, an acceptable electron density at the binding site, and no mutations in the binding site were selected as potential ensemble candidates.

The goal of the ensemble selection was to minimize the size of the ensemble while maximizing the diversity of the contained structures. This was achieved by aligning all binding sites using the "align_binding_sites" routine that comes with Schrodinger Maestro version 2021-2 and selecting up to 4 structures with the highest binding site RMSD to each other[87]. The selected structures were then thoroughly prepared.

For the preparation of the protein structures, we used Schrodinger Maestro version 2021-

2[87]. We regenerated the crystal mates to ensure that there were no crystallization artifacts introduced by neighboring proteins in the crystal structure. In case there were binding site-remote mutations detected in the protein that could not affect the ligand binding mode, they were reverted to wild-type. We removed all crystallization adjuvants, but kept all physiological co-factors within a 12 Å radius around the ligand. The Protein Preparation Wizard within Maestro was used to assign bond orders, add explicit hydrogens, create zero-bond orders to metals, create disulfide bonds, convert selenomethionines to methionines, fill in missing side chains and loops, and generate protonation states at pH 7.4 ± 0.1[88]. We then optimized the H-bond network at physiological pH and ran a minimization restrained to 0.3 Å. Finally, the structures were visually checked for any problems and fixed where necessary. A special focus was placed on the protonation states of aspartic acids, glutamic acids, and histidines, as well as flips of histidines, asparagines and glutamines.

It is well known that water can significantly influence the strength of a ligand binding to a protein[89–92]. Therefore, we modeled the binding site of the ensemble candidates in several different solvation states, depending on the availability of co-crystallized waters. When no co-crystallized water molecules were resolved, no solvation states were modeled. To select the final ensemble, we cross-docked all co-crsytallized ligands for a protein to the ensemble candidates in different solvation states. We calculated the lowest RMSD for each ligand-structure pair and selected the ensemble with the lowest average RMSD over all cross-docked ligands. It was therefore possible to get ensembles with more than one solvation state of a crystal structure, but we made sure that there were always at least 2 different crystal structures used in an ensemble. Figure 3.5 shows an overview of the complete ensemble generation process.

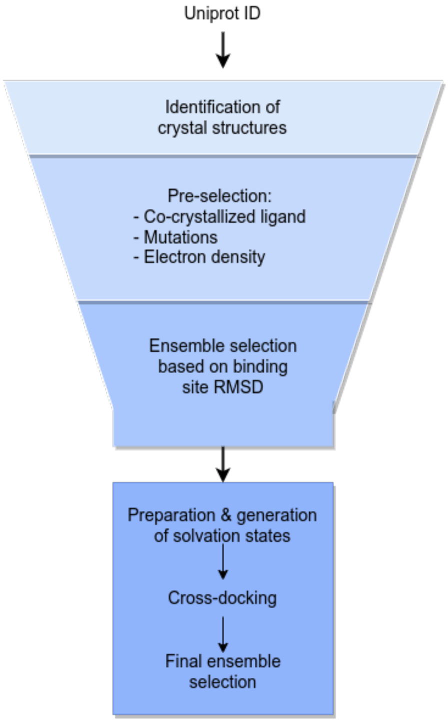

Uniprot ID
Identification of crystal structures
Pre-selection:
- Co-crystallized ligand
- Mutations
- Electron density
Ensemble selection based on binding site RMSD
Preparation & generation of solvation states
Cross-docking
Final ensemble selection

Figure 3.5: Process for the generation of ensembles. Crystal structures are selected based on a Uniprot ID and pre-filtered based on the electron density, co-crystallized ligands, and mutations. An initial crystal structure ensemble is generated by maximizing the binding site RMSD. The selected crystal structures are prepared in Schrodinger's Maestro and different solvation states are modeled. Cross-docking to the solvation states is used to get the final ensemble.

3.4.2 Ensemble docking

PanScreen currently uses smina, Glide, and LeDock to generate docking poses and calculate accompanying scores[47,48,93]. For smina, the co-crystallized ligand was used to identify the binding site and the default buffer of 4 Å was added. We chose to generate up to 9 docking poses with an exhaustiveness of 16. For glide, we used single-precision docking to generate up to 10 docking poses. With LeDock, we generated up to 20 docking poses with a box constructed with a buffer of 6 Å around the co-crystallized ligand. Each ligand was docked to each structure in the ensemble using all 3 docking programs.

After generating protein-ligand complexes with the programs mentioned above, we used gnina to rescore all poses[94]. Gnina was run with the default model, the score_only flag, with

2 CNN rotations, and an exhaustiveness of 16. Additionally, we used a model trained to predict binding affinities and generate protein-ligand interaction fingerprints based on po-sco. The training of this model followed the original publication[69]. This model was used to analyze all protein-ligand complexes generated by smina, Glide, and LeDock.

3.4.3 Consenus prediction

The calculated docking scores as well as the interaction fingerprints from the po-sco model were used to compute the final consensus prediction. This was done by training an individual consensus model for each implemented off-target. The docking scores of smina, Glide, LeDock, and gnina, as well as the affinity predicted by the po-sco model, were first converted to kcal/mol where necessary. We made sure that all affinities were less than or equal to zero by capping positive scores. In addition to the docking scores, we also calculated the standard deviation of the calculated docking scores over all generated poses for each program to estimate the uncertainty of the docking programs. The po-sco model also predicts an uncertainty estimation which was used for the same purpose. The docking scores and uncertainties were then passed through a radial basis function (RBF) expansion $r(x)$ as shown in Equation 3.1, where x is the binding affinity predicted by a docking program or the po-sco model, the binning threshold set c is defined as $x_{min} = c_1 < c_2 < \cdots < c_m = x_{max}$ with $x_{min} = -15$ and $x_{max} = 0$, and m is the number of bins. This number is subject to hyperparameter optimization and varies between models.

$$r(x) = \left(e^{-\frac{(x-c_1)^2}{\sigma^2}}, e^{-\frac{(x-c_2)^2}{\sigma^2}}, \ldots, e^{-\frac{(x-c_m)^2}{\sigma^2}} \right) \tag{3.1}$$

The definition of σ follows Equation 3.2 where s is also subject to hyperparameter optimization.

$$\sigma := s\sqrt{|c_1 - c_2|} > 0. \tag{3.2}$$

This means that we used in total 10 RBF representations as inputs which were all concatenated: 5 docking scores and 5 corresponding uncertainties. The interaction fingerprints for the best complexes from smina, glide, and LeDock were then concatenated with the expanded docking scores and uncertainties. Since it is not easily possible to objectively determine the "best" complex, we chose the one that had the best affinity predicted by the po-sco model. An overview of the input processing can be found in Figure 3.6A.

The consensus model itself is a simple feed-forward neural network. A visual representation of its architecture can be found in Figure 3.6B. The processed inputs were passed through N feed-forward blocks. One block consisted of a linear layer, a leaky ReLU activation function, layer normalization, and a dropout node. The number of blocks (N) is subject to hyperparameter optimization and was in the range of 1 to 3. After the N feed-forward blocks, a single linear layer predicted the binding affinity and the log of the variance. The width of the linear layers was determined by a hyperparameter optimization for each target individually. During training, the predicted affinity and log variance were used to train the model using maximum likelihood estimation (minimization of the negative log likelihood loss). This is defined in Equation 3.3 where θ represents the model parameters, n is the number of samples in a batch, y_i is the true label for sample i, x_i is the input of sample i, and $p(\cdot)$ is the probability density function that gives the conditional probability of y_i given x_i and θ.

$$\mathcal{L}(\theta) = -\sum_{i=1}^{n} \log p(y_i|x_i;\theta) \tag{3.3}$$

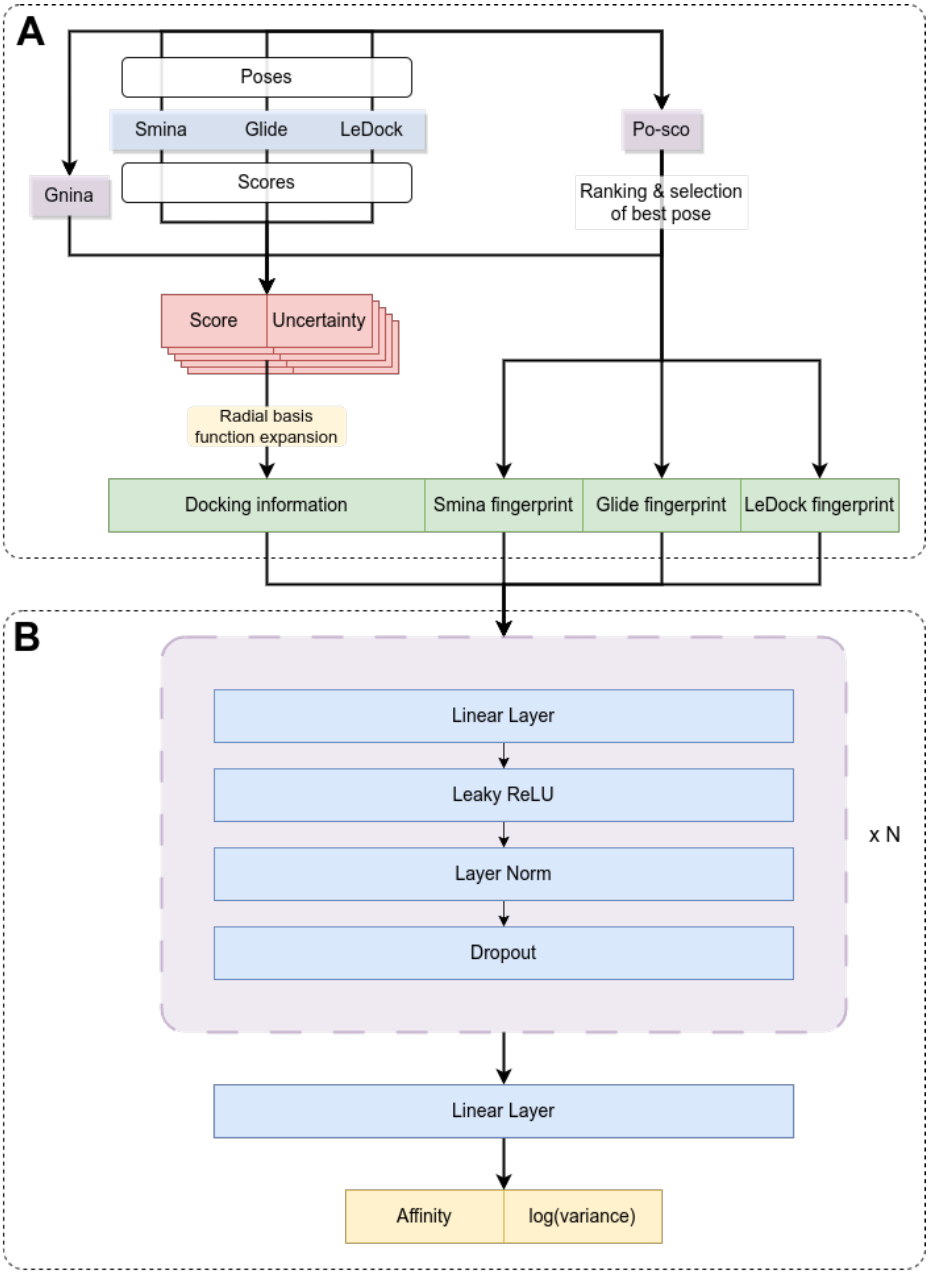


Figure 3.6: The full consensus model architecture. **A)** Processing of the inputs for the consensus model. Smina, Glide, and LeDock (blue) are used to generate docking poses and calculate docking scores and uncertainties (red). Gnina and the po-sco model (purple) are used to re-score the generated complexes, and their scores are combined with the ones of the docking programs (red). The scores and uncertainties are passed through a radial basis function expansion (yellow) to obtain the final docking information. The po-sco model is also used to generate interaction fingerprints for the best complexes generated by smina, Glide, and LeDock. The processed scores and the fingerprints are concatenated to form the final input of the model (green). **B)** Architecture of the consensus model itself. The processed input is passed N times through a linear layer followed by a leaky ReLU activation function, layer normalization, and a dropout node (purple). For the last layer, no activation function, layer normalization, or dropout is applied. The model predicts the affinity as well as the log variance.

The neural network $\mathrm{NN}(\cdot)$ predicts a distribution as $\mathrm{NN}(x_i) = (\hat{y}_i, \log \sigma_i^2)$ where $\hat{y}$ is the predicted mean and $\log \sigma^2$ is the predicted log variance. The conditional probability $p(y_i|x_i;\theta)$ is then calculated as defined in Equation 3.4.

$$p(y_i|x_i;\theta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-\hat{y})^2}{2\sigma^2}\right) \tag{3.4}$$

To increase the robustness of the model, we trained a total of 100 models with different seeds for the weight initialization for each off-target. The predictions of individual models were aggregated into a final prediction using a weighted mean based on compound similarities. This step was only done once all 100 models were trained. For each compound in the training set, we calculated the optimal weights for the 100 models using a multi-linear regression. For each unseen compound (from the test set or during inference), we calculated the Tanimoto similarities to all compounds in the training set based on Morgan fingperprints with a radius of 2 and size of 1024 bits. All training compounds with a similarity of > 0.75 to the unseen compound were selected as reference compounds. In case there were less than 5 training compounds with a similarity > 0.75 to the unseen compounds, the 5 training compounds with the highest similarity were chosen. The reference compounds and their similarities were then used to calculate a weighted average of the model weights according to Equation 3.5 where w_j is a vector containing the model weights for the unseen compound j, w_i are the optimal model weights for reference compound i, N is the number of reference compounds, and s_{ij} is the similarity between reference compound i and unseen compound j.

$$w_j = \frac{\sum_{i=1}^{N} w_i s_{ij}}{\sum_{i=1}^{N} s_{ij}} \tag{3.5}$$

The data available for training and validation of the consensus models was usually limited, and the molecules did not cover the entire chemical space. Thus, estimating the applicability domain of the models is an essential part of the prediction. We did this by developing an applicability score. To calculate the applicability score, we first mapped all molecules used to train and validate a consensus model into 4-dimensional space using their smina, glide, LeDock, and gnina scores. We then constructed a convex hull around all data points in 4D space. When evaluating a query molecule, we mapped its docking scores into the same 4D space and checked whether it was within the hull. If it was, the applicability score was set to 1.0. If the query molecule fell outside the convex hull, we applied an exponential decay to the shortest distance between the molecule and the surface of the hull. The applicability score is therefore in the interval $[0, 1]$ where higher scores indicate a better overlap with the applicability domain of the model.

$$a = \exp\left(-||q - S_H(X)||\right) \tag{3.6}$$

The calculation of the applicability score a is shown in Equation 3.6 where q is a query molecule, X represents all molecules used to train and validate the model, $S_H(\cdot)$ is the surface of the convex hull around the set X, and $|| \cdot ||$ is the Euclidean norm.

In addition to the applicability score, we also introduced warning flags for our predictions. In total, there are 3 flags that could be raised: i) The compound is not binding (sum of smina, glide, and LeDock scores is > -8), ii) The compound could not be docked by all implemented docking programs, iii) The compound has unfavorable docking scores (sum of smina, glide, and LeDock scores is > -14).

3.4.4 Dataset generation

To train a model for a specific (off-)target, a dataset containing compounds with experimentally determined binding affinities are needed. For reproducibility, we developed a standardized routine to obtain and process these data. In the first step, the Uniprot ID of the target of interest is used to search for tested compounds on PubChem[95]. The obtained list is then filtered to only include data points with an activity type of either K_i or K_d and with an absolute affinity value (no values with indication "less than" or "greater than"). For compounds with multiple measurements, we first calculated the mean affinity and excluded all data points with measured affinities that deviate more than $\pm$ 30% from the mean affinity. Finally, we downloaded the 3D SDF files from PubChem for all remaining compounds. In some cases, there was no 3D structure available. This was mostly the case for very large and flexible compounds or for compounds with ambiguous stereochemistry. These compounds were excluded from the final dataset.

The final dataset was then sorted by decreasing affinity and every 5th element was added to the validation set while the remaining elements were used for the training set. This approach was chosen to ensure a similar distribution of affinities between the training and validation set. Since the consensus model that was used to make the final predictions is agnostic of ligand structures, we did not pay attention to any structural similarities between the training and validation set. To deal with imbalances of high- and low-affinity compounds, we clustered all training compounds into 3 clusters with affinity thresholds of < 100 nM, < 1 μM, and > 1 μM. We then used a weighted random sampler to ensure the same numbers of high-, medium-, and low-affinity compounds per mini-batch.

3.5 Conclusion

With the rise of increasingly accurate computational methods, in silico prediction of off-target interactions has become a viable tool to complement classical in vitro testing. In light of the FDA Modernization Act 2.0, we believe that it is the right time to further promote in silico methods due to their advantages in resource efficiency and cost effectiveness[96].

In this article, we present PanScreen, an online platform for the automated testing of off-target liabilities. At the time of writing, PanScreen features 14 (off-)targets of various protein families. Using a combination of structure-based modeling and artificial intelligence, all backed by profound knowledge in structural biology and medicinal chemistry, PanScreen is able to accurately predict binding affinities for diverse molecules. In addition to the predicted binding affinities, PanScreen offers possible binding modes as an explanation for the predictions. We also showed that PanScreen has the potential to detect activity cliffs between highly similar molecules. Due to the underlying technology, which is independent of a specific use case, our platform can be used not only for toxicology studies, but also for drug repurposing, selectivity assessment, and a wide range of other applications in the pharmaceutical and biomedical fields. To our knowledge, PanScreen is the first online platform that combines structure-based methods with deep learning to assess off-target interactions in a portfolio of highly curated proteins.

By providing PanScreen as a publicly available online platform, we hope to enable scientists of various backgrounds to use in silico off-target analysis with minimal effort and integrate the results in their own research.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of two RTX A5000 GPUs used for this research.

3.6 Appendix

Table A3.1: Confusion matrix for the MMP analysis.

	$\Delta\Delta G_{pred} < 1.0$	$\Delta\Delta G_{pred} < 2.0$	$\Delta\Delta G_{pred} > 2.0$	**total**
$\Delta\Delta G_{exp} < 1.0$	2899	85	10	2994
$\Delta\Delta G_{exp} < 2.0$	329	43	5	377
$\Delta\Delta G_{exp} > 2.0$	71	22	2	95
total	3299	150	17	

Table A3.2: Performance metrics for smina. Shown are the Pearson correlation coefficient (PCC; higher is better), the mean unsigned error (MUE; lower is better), the root mean squared error (RMSE; lower is better), and the area under the receiver operating characteristics (AUROC; higher is better). Note that smina sometimes produced positive scores. Hence, the MUE and RMSE can get very high for some targets.

Protein name	PCC	MUE [kcal/mol]	RMSE [kcal/mol]	AUROC
Tyrosine-protein kinase JAK2	0.43	1.89	2.15	0.72
Estrogen receptor alpha	0.46	1.59	2.37	0.75
Glucocorticoid receptor	0.16	1.57	1.93	0.64
Beta-2 adrenergic receptor	0.25	1.69	2.05	0.60
Androgen receptor	0.17	3.87	6.01	0.66
Dopamine receptor D2	0.11	1.49	1.82	0.58
Tyrosine-protein kinase JAK1	0.46	2.12	2.31	0.75
Substance-P receptor	0.08	1.59	1.93	0.58
5HT receptor 1B	0.43	1.34	1.68	0.70
PPARγ	0.35	1.57	2.09	0.63
Melatonin receptor 1B	0.37	2.18	2.63	0.68
Phosphodiesterase 4D	0.59	1.47	1.96	0.74
Estrogen receptor beta	0.25	11.12	14.70	0.69
Phosphodiesterase 10A	0.22	2.14	2.61	0.65
Mean	0.31 ± 0.15	2.48 ± 2.59	3.09 ± 2.70	0.67 ± 0.06

Table A3.3: Performance metrics for Glide. Shown are the Pearson correlation coefficient (PCC; higher is better), the mean unsigned error (MUE; lower is better), the root mean squared error (RMSE; lower is better), and the area under the receiver operating characteristics (AUROC; higher is better). Note that Glide failed to dock certain compounds. In these cases, the docking score was set to 0.

Protein name	PCC	MUE [kcal/mol]	RMSE [kcal/mol]	AUROC
Tyrosine-protein kinase JAK2	0.06	3.32	3.73	0.53
Estrogen receptor alpha	0.56	1.85	2.54	0.73
Glucocorticoid receptor	0.14	1.87	2.45	0.57
Beta-2 adrenergic receptor	0.37	1.57	1.97	0.65
Androgen receptor	0.27	3.68	5.18	0.69
Dopamine receptor D2	0.09	1.60	2.01	0.60
Tyrosine-protein kinase JAK1	-0.03	3.72	4.16	0.47
Substance-P receptor	0.19	2.10	2.63	0.67
5HT receptor 1B	0.51	1.59	1.93	0.71
PPARγ	0.13	2.89	3.64	0.58
Melatonin receptor 1B	0.34	2.25	2.86	0.69
Phosphodiesterase 4D	0.02	2.31	2.60	0.52
Estrogen receptor beta	0.31	7.31	8.26	0.69
Phosphodiesterase 10A	0.29	2.55	3.02	0.77
Mean	0.23 ± 0.18	2.76 ± 1.51	3.36 ± 1.69	0.63 ± 0.09

Table A3.4: Performance metrics for LeDock. Shown are the Pearson correlation coefficient (PCC; higher is better), the mean unsigned error (MUE; lower is better), the root mean squared error (RMSE; lower is better), and the area under the receiver operating characteristics (AUROC; higher is better).

Protein name	PCC	MUE [kcal/mol]	RMSE [kcal/mol]	AUROC
Tyrosine-protein kinase JAK2	0.49	2.87	3.10	0.79
Estrogen receptor alpha	0.53	2.54	2.96	0.74
Glucocorticoid receptor	0.31	3.24	3.56	0.69
Beta-2 adrenergic receptor	0.29	2.01	2.46	0.59
Androgen receptor	0.41	3.48	3.84	0.71
Dopamine receptor D2	0.20	2.25	2.61	0.63
Tyrosine-protein kinase JAK1	0.62	2.74	2.89	0.81
Substance-P receptor	0.01	3.41	3.88	0.63
5HT receptor 1B	0.47	2.66	3.03	0.70
PPARγ	0.40	1.65	2.04	0.69
Melatonin receptor 1B	0.17	4.76	5.11	0.57
Phosphodiesterase 4D	0.40	2.34	2.82	0.70
Estrogen receptor beta	0.20	5.36	5.77	0.62
Phosphodiesterase 10A	0.59	4.20	4.49	0.82
Mean	0.36 ± 0.18	3.11 ± 1.06	3.47 ± 1.06	0.69 ± 0.08

Table A3.5: Performance metrics for gnina. Shown are the Pearson correlation coefficient (PCC; higher is better), the mean unsigned error (MUE; lower is better), the root mean squared error (RMSE; lower is better), and the area under the receiver operating characteristics (AUROC; higher is better).

Protein name	PCC	MUE [kcal/mol]	RMSE [kcal/mol]	AUROC
Tyrosine-protein kinase JAK2	0.65	1.53	1.77	0.87
Estrogen receptor alpha	0.49	2.07	2.60	0.71
Glucocorticoid receptor	0.33	1.23	1.48	0.70
Beta-2 adrenergic receptor	0.32	1.54	1.86	0.63
Androgen receptor	0.49	1.57	1.92	0.57
Dopamine receptor D2	0.23	1.13	1.39	0.61
Tyrosine-protein kinase JAK1	0.39	1.55	1.72	0.73
Substance-P receptor	0.08	1.54	1.86	0.67
5HT receptor 1B	0.44	1.25	1.48	0.71
PPARγ	0.40	1.65	2.04	0.69
Melatonin receptor 1B	0.22	2.32	2.78	0.62
Phosphodiesterase 4D	0.50	2.09	2.52	0.68
Estrogen receptor beta	0.06	2.08	2.47	0.50
Phosphodiesterase 10A	0.53	1.56	1.94	0.75
Mean	0.37 ± 0.17	1.65 ± 0.36	1.99 ± 0.44	0.67 ± 0.09

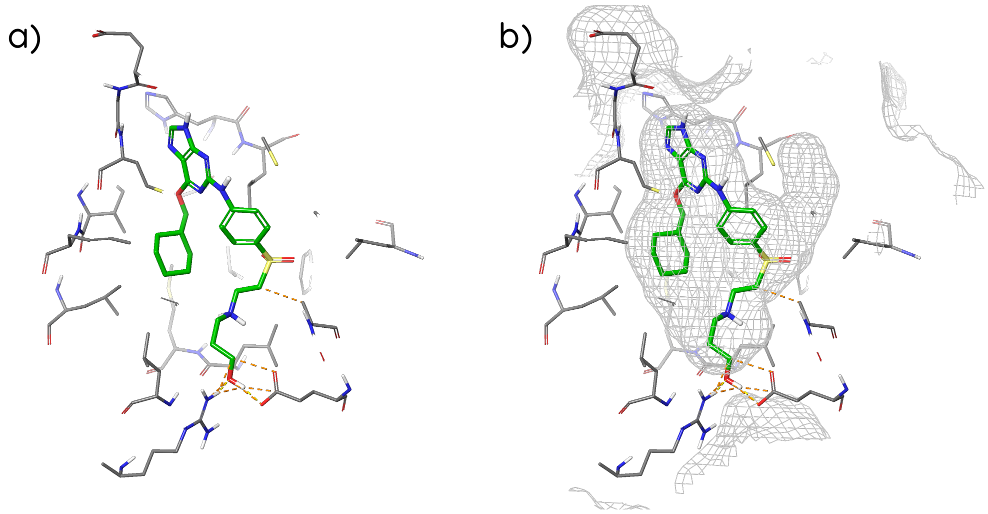


Figure A3.1: Binding mode of DB08309 at the estrogen receptor alpha generated by Glide. a) without and b) with the binding site surface displayed as a mesh.

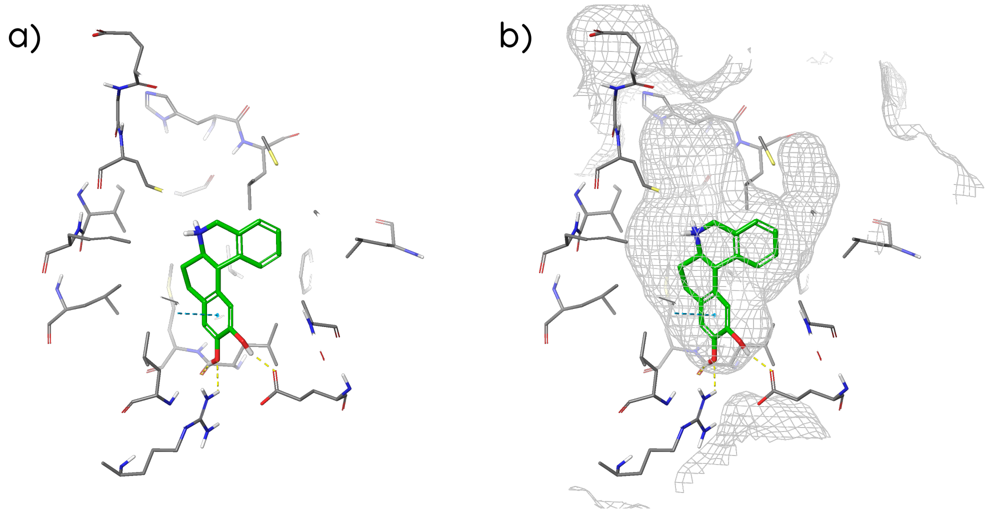


Figure A3.2: Binding mode of DB16139 at the estrogen receptor alpha generated by Glide. a) without and b) with the binding site surface displayed as a mesh.

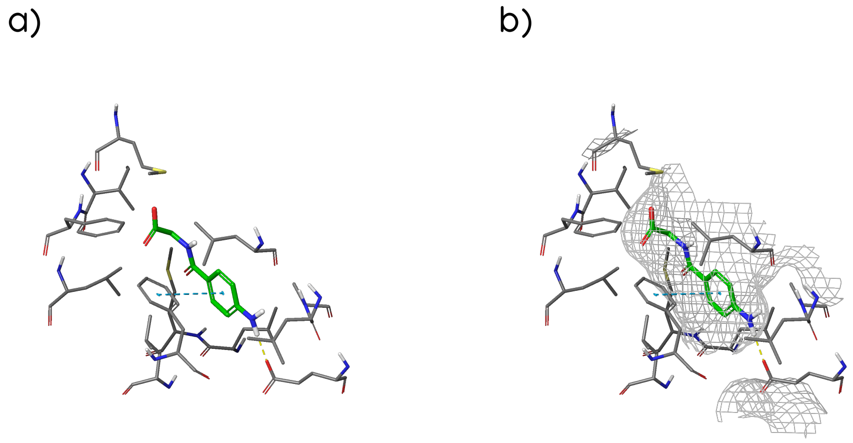


Figure A3.3: Binding mode of DB00345 at the estrogen receptor alpha generated by Glide. a) without and b) with the binding site surface displayed as a mesh.

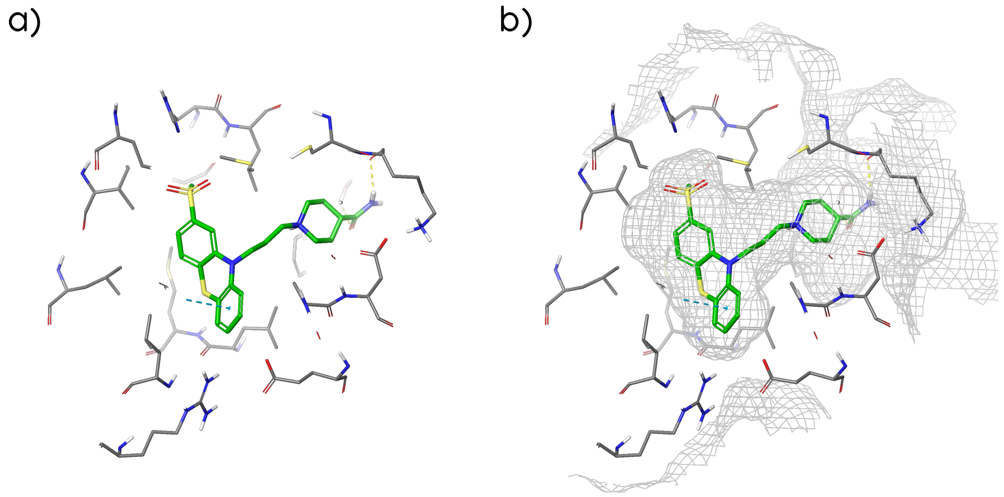


Figure A3.4: Binding mode of DB13591 at the estrogen receptor alpha generated by Glide. a) without and b) with the binding site surface displayed as a mesh.

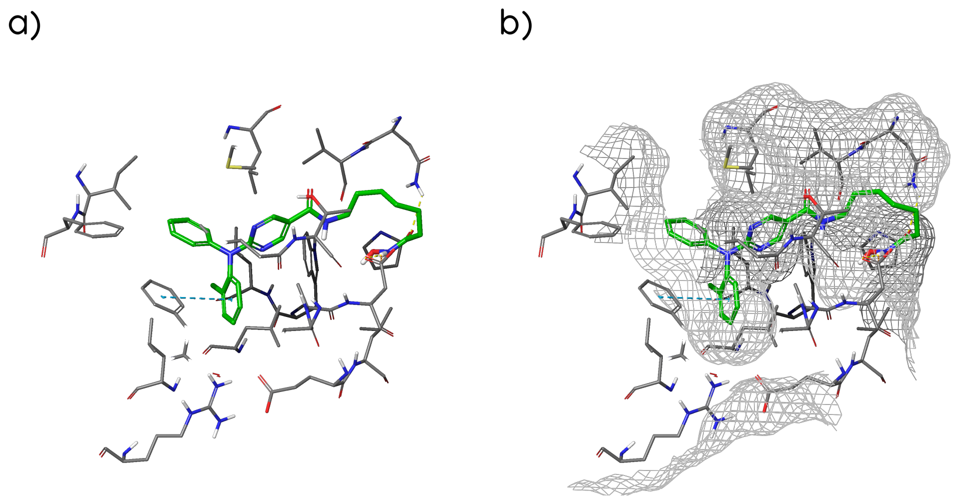


Figure A3.5: Binding mode of DB15449 at the estrogen receptor alpha generated by Glide. a) without and b) with the binding site surface displayed as a mesh.

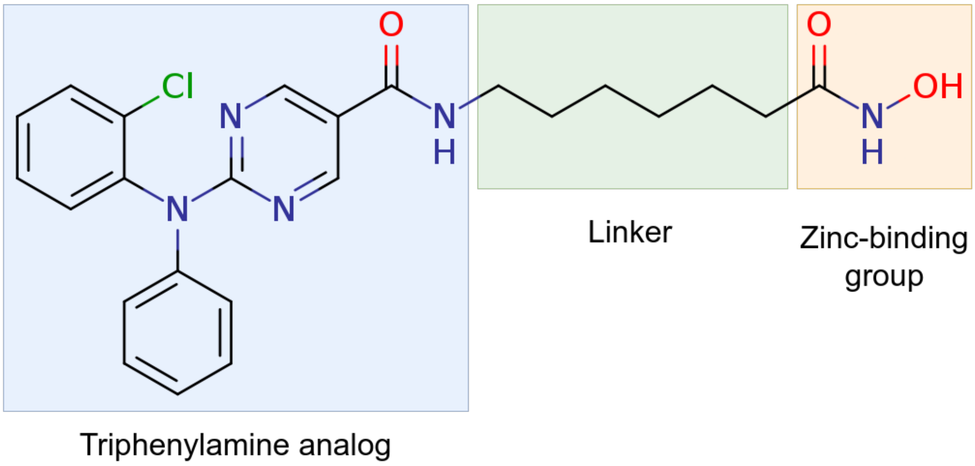


Figure A3.6: 2D structure of citarinostat.

References

[1] Manuel S. Sellner, Markus A. Lill, and Martin Smiesko. Panscreen: A comprehensive approach to off-target liability assessment. *bioRxiv*, page 2023.11.16.567496, 11 2023.

[2] Irfan A. Rather, Wee Yin Koh, Woon K. Paek, and Jeongheui Lim. The Sources of Chemical Contaminants in Food and Their Health Implications. *Frontiers in Pharmacology*, 8(NOV):308465, 11 2017.

[3] Stavros Sifakis, Vasilis P. Androutsopoulos, Aristeidis M. Tsatsakis, and Demetrios A. Spandidos. Human exposure to endocrine disrupting chemicals: effects on the male and female reproductive systems. *Environmental Toxicology and Pharmacology*, 51:56–70, 4 2017.

[4] Giuseppe Genchi, Alessia Carocci, Graziantonio Lauria, Maria Stefania Sinicropi, and Alessia Catalano. Nickel: Human Health and Environmental Toxicology. *International Journal of Environmental Research and Public Health*, 17(3):679, 1 2020.

[5] Muhammad Shahid, Muhammad Nadeem, and Hafiz Faiq Bakhat. Environmental toxicology and associated human health risks. *Environmental Science and Pollution Research*, 27(32):39671–39675, 11 2020.

[6] Hartmut Beck, Michael Härter, Bastian Haß, Carsten Schmeck, and Lars Baerfacker. Small molecules and their impact in drug discovery: A perspective on the occasion of the 125th anniversary of the Bayer Chemical Research Laboratory. *Drug Discovery Today*, 27(6):1560–1574, 6 2022.

[7] Qingxin Li and Congbao Kang. Mechanisms of Action for Small Molecules Revealed

by Structural Biology in Drug Discovery. *International Journal of Molecular Sciences*, 21(15):5262, 7 2020.

[8] Daniel G Rudmann. On-target and Off-target-based Toxicologic Effects. *Toxicologic Pathology*, 41(2):310–314, 2 2013.

[9] Christina Buchanan, Kate Lee, and Peter Shepherd. For Better or Worse: The Potential for Dose Limiting the On-Target Toxicity of PI 3-Kinase Inhibitors. *Biomolecules*, 9(9):402, 8 2019.

[10] József Tímár and Andrea Uhlyarik. On-Target Side Effects of Targeted Therapeutics of Cancer. *Pathology and Oncology Research*, 28:1610694, 9 2022.

[11] F. Peter Guengerich. Mechanisms of Drug Toxicity and Relevance to Pharmaceutical Development. *Drug Metabolism and Pharmacokinetics*, 26(1):3–14, 1 2011.

[12] Geoffrey Kabue Kiriiri, Peter Mbugua Njogu, and Alex Njoroge Mwangi. Exploring different approaches to improve the success of drug discovery and development projects: a review. *Future Journal of Pharmaceutical Sciences*, 6(1):27, 12 2020.

[13] Ann Lin, Christopher J. Giuliano, Ann Palladino, Kristen M. John, Connor Abramowicz, Monet Lou Yuan, Erin L. Sausville, Devon A. Lukow, Luwei Liu, Alexander R. Chait, Zachary C. Galluzzo, Clara Tucker, and Jason M. Sheltzer. Off-target toxicity is a common mechanism of action of cancer drugs undergoing clinical trials. *Science Translational Medicine*, 11(509):8412, 9 2019.

[14] Oliver Schoor, Jens Fritsche, Sarah Kutscher, Andrea Mahr, Lea Stevermann, Annika Sonntag, Franziska Hoffgaard, Dominik Vahrenhorst, Julia Leibold, Valentina

Goldfinger, Leonie Alten, Sebastian Bunk, Dominik Maurer, Steffen Walter, Hans-Georg Rammensee, Harpreet Singh-Jasuja, and Toni Weinschenk. Abstract 2291: On- and off target toxicity profiling for adoptive cell therapy by mass spectrometry-based immunopeptidome analysis of primary human normal tissues. *Cancer Research*, 76(14_Supplement):2291–2291, 7 2016.

[15] Daniel S. Bejan and Michael S. Cohen. Methods for profiling the target and off-target landscape of PARP inhibitors. *Current Research in Chemical Biology*, 2:100027, 1 2022.

[16] André Mateus, Nils Kurzawa, Jessica Perrin, Giovanna Bergamini, and Mikhail M. Savitski. Drug Target Identification in Tissues by Thermal Proteome Profiling. *Annual Review of Pharmacology and Toxicology*, 62(1):465–482, 1 2022.

[17] Richard C. Kevin, Elizabeth A. Cairns, Rochelle Boyd, Jonathon C. Arnold, Michael T. Bowen, Iain S. McGregor, and Samuel D. Banister. Off-target pharmacological profiling of synthetic cannabinoid receptor agonists including AMB-FUBINACA, CUMYL-PINACA, PB-22, and XLR-11. *Frontiers in Psychiatry*, 13:1048836, 12 2022.

[18] Yili Xu, Ona Barauskas, Cynthia Kim, Darius Babusis, Eisuke Murakami, Dmytro Kornyeyev, Gary Lee, George Stepan, Michel Perron, Roy Bannister, Brian E. Schultz, Roman Sakowicz, Danielle Porter, Tomas Cihlar, and Joy Y. Feng. Off-Target In Vitro Profiling Demonstrates that Remdesivir Is a Highly Selective Antiviral Agent. *Antimicrobial Agents and Chemotherapy*, 65(2), 1 2021.

[19] Meenakshi Mishra, Hongliang Fei, and Jun Huan. Computational prediction of

toxicity. In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 686–691. IEEE, 12 2010.

[20] Rodolpho C. Braga, Vinicius M. Alves, Meryck F. B. Silva, Eugene Muratov, Denis Fourches, Luciano M. Lião, Alexander Tropsha, and Carolina H. Andrade. Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Molecular Informatics*, 34(10):698–701, 10 2015.

[21] Andreas Mayr, Günter Klambauer, Thomas Unterthiner, and Sepp Hochreiter. DeepTox: Toxicity Prediction using Deep Learning. *Frontiers in Environmental Science*, 3(FEB):167215, 2 2016.

[22] Francis E. Agamah, Gaston K. Mazandu, Radia Hassan, Christian D. Bope, Nicholas E. Thomford, Anita Ghansah, and Emile R. Chimusa. Computational/in silico methods in drug target and lead prediction. *Briefings in Bioinformatics*, 21(5):1663–1675, 9 2020.

[23] Angelo Vedani, Max Dobler, and Martin Smieško. VirtualToxLab — A platform for estimating the toxic potential of drugs, chemicals and natural products. *Toxicology and Applied Pharmacology*, 261(2):142–153, 6 2012.

[24] Angelo Vedani, Max Dobler, Zhenquan Hu, and Martin Smieško. OpenVirtualToxLab—A platform for generating and exchanging in silico toxicity data. *Toxicology Letters*, 232(2):519–532, 1 2015.

[25] Masamitsu Honma. An assessment of mutagenicity of chemical substances by (quantitative) structure–activity relationship. *Genes and Environment*, 42(1):23, 12 2020.

[26] Toshio Kasamatsu, Airi Kitazawa, Sumie Tajima, Masahiro Kaneko, Kei-ichi Sugiyama, Masami Yamada, Manabu Yasui, Kenichi Masumura, Katsuyoshi Horibata, and Masamitsu Honma. Development of a new quantitative structure–activity relationship model for predicting Ames mutagenicity of food flavor chemicals using StarDrop™ auto-Modeller™. *Genes and Environment*, 43(1):16, 12 2021.

[27] Anke Wilm, Marina Garcia de Lomana, Conrad Stork, Neann Mathai, Steffen Hirte, Ulf Norinder, Jochen Kühnl, and Johannes Kirchmair. Predicting the Skin Sensitization Potential of Small Molecules with Machine Learning Models Trained on Biologically Meaningful Descriptors. *Pharmaceuticals*, 14(8):790, 8 2021.

[28] Emily Golden, Daniel C. Ukaegbu, Peter Ranslow, Robert H. Brown, Thomas Hartung, and Alexandra Maertens. The Good, The Bad, and The Perplexing: Structural Alerts and Read-Across for Predicting Skin Sensitization Using Human Data. *Chemical Research in Toxicology*, 36(5):734–746, 5 2023.

[29] Hongyi Zhou, Hongnan Cao, Lilya Matyunina, Madelyn Shelby, Lauren Cassels, John F. McDonald, and Jeffrey Skolnick. MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action. *Molecular Pharmaceutics*, 17(5):1558–1574, 5 2020.

[30] Bara A Badwan, Gerry Liaropoulos, Efthymios Kyrodimos, Dimitrios Skaltsas, Aristotelis Tsirigos, and Vassilis G Gorgoulis. Machine learning approaches to predict drug efficacy and toxicity in oncology. *Cell Reports Methods*, 3(2):100413, 2 2023.

[31] Shanmugam Anusuya, Manish Kesherwani, K. Vishnu Priya, Antonydhason Vimala, Gnanendra Shanmugam, Devadasan Velmurugan, and M. Michael Gromiha. Drug-

Target Interactions: Prediction Methods and Applications. *Current Protein & Peptide Science*, 19(6):537–561, 4 2018.

[32] Zhan-Heng Chen, Zhu-Hong You, Zhen-Hao Guo, Hai-Cheng Yi, Gong-Xu Luo, and Yan-Bin Wang. Prediction of Drug–Target Interactions From Multi-Molecular Network Based on Deep Walk Embedding Model. *Frontiers in Bioengineering and Biotechnology*, 8:528253, 6 2020.

[33] Maha A. Thafar, Rawan S. Olayan, Somayah Albaradei, Vladimir B. Bajic, Takashi Gojobori, Magbubah Essack, and Xin Gao. DTi2Vec: Drug–target interaction prediction using network embedding and ensemble learning. *Journal of Cheminformatics*, 13(1):71, 12 2021.

[34] Qing Ye, Chang-Yu Hsieh, Ziyi Yang, Yu Kang, Jiming Chen, Dongsheng Cao, Shibo He, and Tingjun Hou. A unified drug–target interaction prediction framework based on knowledge graph and recommendation system. *Nature Communications*, 12(1):6775, 11 2021.

[35] Timothy E. H. Allen, Jonathan M. Goodman, Steve Gutsell, and Paul J. Russell. Defining Molecular Initiating Events in the Adverse Outcome Pathway Framework for Risk Assessment. *Chemical Research in Toxicology*, 27(12):2100–2112, 12 2014.

[36] Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, 11(2):137–148, 2 2016.

[37] Ashutosh Kumar and Kam Y. J. Zhang. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Frontiers in Chemistry*, 6(JUL):383861, 7 2018.

[38] Su-Qing Yang, Qing Ye, Jun-Jie Ding, Yin, Ai-Ping Lu, Xiang Chen, Ting-Jun Hou, and Dong-Sheng Cao. Current advances in ligand-based target prediction. *WIREs Computational Molecular Science*, 11(3):e1504, 5 2021.

[39] Tao Huang, Hong Mi, Cheng-yuan Lin, Ling Zhao, Linda L. D. Zhong, Feng-bin Liu, Ge Zhang, Ai-ping Lu, and Zhao-xiang Bian. MOST: most-similar ligand based approach to target prediction. *BMC Bioinformatics*, 18(1):165, 12 2017.

[40] Faraz Shaikh, Hio Kuan Tai, Nirali Desai, and Shirley W. I. Siu. LigTMap: ligand and structure-based target identification and activity prediction for small molecular compounds. *Journal of Cheminformatics*, 13(1):44, 12 2021.

[41] Yanqing Yang, Zhengdan Zhu, Xiaoyu Wang, Xinben Zhang, Kaijie Mu, Yulong Shi, Cheng Peng, Zhijian Xu, and Weiliang Zhu. Ligand-based approach for predicting drug targets and for virtual screening against COVID-19. *Briefings in Bioinformatics*, 22(2):1053–1064, 3 2021.

[42] Dagmar Stumpfe, Huabin Hu, and Jürgen Bajorath. Evolving Concept of Activity Cliffs. *ACS Omega*, 4(11):14360–14368, 9 2019.

[43] Markus Dablander, Thierry Hanser, Renaud Lambiotte, and Garrett M. Morris. Exploring QSAR models for activity-cliff prediction. *Journal of Cheminformatics*, 15(1):47, 4 2023.

[44] Amy C. Anderson. The Process of Structure-Based Drug Design. *Chemistry & Biology*, 10(9):787–797, 9 2003.

[45] Maria Batool, Bilal Ahmad, and Sangdun Choi. A Structure-Based Drug Discovery Paradigm. *International Journal of Molecular Sciences*, 20(11):2783, 6 2019.

[46] Jerome Eberhardt, Diogo Santos-Martins, Andreas F. Tillack, and Stefano Forli. AutoDock Vina 1.2.0: New Docking Methods, Expanded Force Field, and Python Bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898, 8 2021.

[47] David Ryan Koes, Matthew P. Baumgartner, and Carlos J. Camacho. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *Journal of Chemical Information and Modeling*, 53(8):1893–1904, 8 2013.

[48] Richard A. Friesner, Jay L. Banks, Robert B. Murphy, Thomas A. Halgren, Jasna J. Klicic, Daniel T. Mainz, Matthew P. Repasky, Eric H. Knoll, Mee Shelley, Jason K. Perry, David E. Shaw, Perry Francis, and Peter S. Shenkin. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749, 3 2004.

[49] Martin Smieško. DOLINA – Docking Based on a Local Induced-Fit Algorithm: Application toward Small-Molecule Binding to Nuclear Receptors. *Journal of Chemical Information and Modeling*, 53(6):1415–1423, 6 2013.

[50] Mengang Xu and Markus A. Lill. Induced fit docking, and the use of QM/MM methods in docking. *Drug Discovery Today: Technologies*, 10(3):e411–e418, 9 2013.

[51] Wilfredo Evangelista Falcon, Sally R. Ellingson, Jeremy C. Smith, and Jerome Baudry. Ensemble Docking in Drug Discovery: How Many Protein Configurations from Molecular Dynamics Simulations are Needed To Reproduce Known Ligand Binding? *The Journal of Physical Chemistry B*, 123(25):5189–5195, 6 2019.

[52] Sara Mohammadi, Zahra Narimani, Mitra Ashouri, Rohoullah Firouzi, and Mohammad Hossein Karimi-Jafari. Ensemble learning from ensemble docking: revisiting the optimum ensemble size problem. *Scientific Reports*, 12(1):410, 1 2022.

[53] John B. O. Mitchell. Machine learning methods in chemoinformatics. *WIREs Computational Molecular Science*, 4(5):468–481, 9 2014.

[54] Antonio Lavecchia. Machine-learning approaches in drug discovery: methods and applications. *Drug Discovery Today*, 20(3):318–331, 3 2015.

[55] Hongming Chen, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6):1241–1250, 6 2018.

[56] Han Shi, Simin Liu, Junqi Chen, Xuan Li, Qin Ma, and Bin Yu. Predicting drug-target interactions using Lasso with random forest based on evolutionary information and chemical structure. *Genomics*, 111(6):1839–1852, 12 2019.

[57] Yayuan Peng, Jiye Wang, Zengrui Wu, Lulu Zheng, Biting Wang, Guixia Liu, Weihua Li, and Yun Tang. MPSM-DTI: prediction of drug–target interaction via machine learning based on the chemical structure and protein sequence. *Digital Discovery*, 1(2):115–126, 4 2022.

[58] Sangjin Ahn, Si Eun Lee, and Mi-hyun Kim. Random-forest model for drug–target interaction prediction via Kullback–Leibler divergence. *Journal of Cheminformatics*, 14(1):67, 10 2022.

[59] Alexandru Korotcov, Valery Tkachenko, Daniel P. Russo, and Sean Ekins. Comparison of Deep Learning With Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Molecular Pharmaceutics*, 14(12):4462–4475, 12 2017.

[60] Yan-Bin Wang, Zhu-Hong You, Shan Yang, Hai-Cheng Yi, Zhan-Heng Chen, and Kai Zheng. A deep learning-based method for drug-target interaction prediction based on long short-term memory neural network. *BMC Medical Informatics and Decision Making*, 20(S2):49, 3 2020.

[61] Ming Wen, Zhimin Zhang, Shaoyu Niu, Haozhi Sha, Ruihan Yang, Yonghuan Yun, and Hongmei Lu. Deep-Learning-Based Drug–Target Interaction Prediction. *Journal of Proteome Research*, 16(4):1401–1409, 4 2017.

[62] Guannan Liu, Manali Singha, Limeng Pu, Prasanga Neupane, Joseph Feinstein, Hsiao-Chun Wu, J. Ramanujam, and Michal Brylinski. GraphDTI: A robust deep learning predictor of drug-target interactions from multiple heterogeneous data. *Journal of Cheminformatics*, 13(1):58, 8 2021.

[63] Jackson G de Souza, Marcelo A. C. Fernandes, and Raquel de Melo Barbosa. A Novel Deep Neural Network Technique for Drug–Target Interaction. *Pharmaceutics*, 14(3):625, 3 2022.

[64] S.M. Hasan Mahmud, Wenyu Chen, Hosney Jahan, Bo Dai, Salah Ud Din, and Anthony Mackitz Dzisoo. DeepACTION: A deep learning-based method for predicting novel drug-target interactions. *Analytical Biochemistry*, 610:113978, 12 2020.

[65] M.A. Ganaie, Minghui Hu, A.K. Malik, M Tanveer, and P.N. Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 10 2022.

[66] Ammar Mohammed and Rania Kora. A comprehensive review on ensemble deep learning: Opportunities and challenges. *Journal of King Saud University - Computer and Information Sciences*, 35(2):757–774, 2 2023.

[67] Frankie J. Fan and Yun Shi. Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction. *Bioorganic & Medicinal Chemistry*, 72:117003, 10 2022.

[68] Mikhail Volkov, Joseph-André Turk, Nicolas Drizard, Nicolas Martin, Brice Hoffmann, Yann Gaston-Mathé, and Didier Rognan. On the Frustration to Predict Binding Affinities from Protein–Ligand Structures with Deep Neural Networks. *Journal of Medicinal Chemistry*, 65(11):7946–7958, 6 2022.

[69] Manuel S. Sellner, Markus A. Lill, and Martin Smieško. Quality Matters: Deep Learning-Based Analysis of Protein-Ligand Interactions with Focus on Avoiding Bias. *bioRxiv*, page 2023.11.13.566916, 11 2023.

[70] Django. https://www.djangoproject.com/.

[71] Noel M. O'Boyle, Michael Banck, Craig A. James, Chris Morley, Tim Vandermeersch, and Geoffrey R. Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 12 2011.

[72] LLC Schrödinger. Schrödinger Release 2021-1: Ligprep, 2021.

[73] Alexander G. Dossetter, Edward J. Griffen, and Andrew G. Leach. Matched Molecular Pair Analysis in drug discovery. *Drug Discovery Today*, 18(15-16):724–731, 8 2013.

[74] Emanuel S. R. Ehmki and Matthias Rarey. Exploring Structure-Activity Relationships with Three-Dimensional Matched Molecular Pairs-A Review. *ChemMedChem*, 13(6):482–489, 3 2018.

[75] Andrew Dalke, Jérôme Hert, and Christian Kramer. mmpdb: An Open-Source Matched Molecular Pair Platform for Large Multiproperty Data Sets. *Journal of Chemical Information and Modeling*, 58(5):902–910, 5 2018.

[76] Ziyi Yang, Shaohua Shi, Li Fu, Aiping Lu, Tingjun Hou, and Dongsheng Cao. Matched Molecular Pair Analysis in Drug Discovery: Methods and Recent Applications. *Journal of Medicinal Chemistry*, 66(7):4361–4377, 4 2023.

[77] David S. Wishart, Yannick D. Feunang, An C. Guo, Elvis J. Lo, Ana Marcu, Jason R. Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour, Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, DIana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 1 2018.

[78] André Fischer, Martin Smieško, Manuel Sellner, and Markus A. Lill. Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. *Journal of Medicinal Chemistry*, 64(5):2489–2500, 3 2021.

[79] Ragy R. Girgis, Jared X. Van Snellenberg, Andrew Glass, Lawrence S. Kegeles, Judy L. Thompson, Melanie Wall, Raymond Y. Cho, Cameron S. Carter, Mark Slifstein, Anissa Abi-Dargham, and Jeffrey A. Lieberman. A proof-of-concept, randomized controlled trial of DAR-0100A, a dopamine-1 receptor agonist, for cognitive enhancement in schizophrenia. *Journal of Psychopharmacology*, 30(5):428–435, 5 2016.

[80] Michelle Landry, Daniel Lévesque, and Thérèse Di Paolo. Estrogenic Properties of Raloxifene, but Not Tamoxifen, on D<sub>2</sub> and D<sub>3</sub> Dopamine Receptors in the Rat Forebrain. *Neuroendocrinology*, 76(4):214–222, 10 2002.

[81] T W Weickert, D. Weinberg, R. Lenroot, S. V. Catts, R. Wells, A. Vercammen, M. O'Donnell, C. Galletly, D. Liu, R. Balzan, B. Short, D. Pellen, J. Curtis, V. J. Carr, J. Kulkarni, P. R. Schofield, and C. S. Weickert. Adjunctive raloxifene treatment improves attention and memory in men and women with schizophrenia. *Molecular Psychiatry*, 20(6):685–694, 6 2015.

[82] Karen J. Kieser, Dong Wook Kim, Kathryn E. Carlson, Benita S. Katzenellenbogen, and John A. Katzenellenbogen. Characterization of the Pharmacophore Properties of Novel Selective Estrogen Receptor Downregulators (SERDs). *Journal of Medicinal Chemistry*, 53(8):3320–3329, 4 2010.

[83] Chu Tang, Changhao Li, Silong Zhang, Zhiye Hu, Jun Wu, Chune Dong, Jian Huang, and Hai-Bing Zhou. Novel Bioactive Hybrid Compound Dual Targeting Estrogen Receptor and Histone Deacetylase for the Treatment of Breast Cancer. *Journal of Medicinal Chemistry*, 58(11):4550–4572, 6 2015.

[84] Rodrigo Mendoza-Sanchez, David Cotnoir-White, Justyna Kulpa, Isabel Jutras, Joshua Pottel, Nicolas Moitessier, Sylvie Mader, and James L. Gleason. Design, synthesis and evaluation of antiestrogen and histone deacetylase inhibitor molecular hybrids. *Bioorganic & Medicinal Chemistry*, 23(24):7597–7606, 12 2015.

[85] Alex Bateman, Maria-Jesus Martin, Sandra Orchard, Michele Magrane, Shadab Ahmad, Emanuele Alpi, Emily H. Bowler-Barnett, Ramona Britto, Hema Bye-A-Jee, Austra Cukura, Paul Denny, Tunca Dogan, ThankGod Ebenezer, Jun Fan, Penelope Garmiri, Leonardo Jose da Costa Gonzales, Emma Hatton-Ellis, Abdulrahman Hussein, Alexandr Ignatchenko, Giuseppe Insana, Rizwan Ishtiaq, Vishal Joshi, Dushyanth Jyothi, Swaathi Kandasaamy, Antonia Lock, Aurelien Luciani, Marija Lu-

garic, Jie Luo, Yvonne Lussi, Alistair MacDougall, Fabio Madeira, Mahdi Mahmoudy, Alok Mishra, Katie Moulang, Andrew Nightingale, Sangya Pundir, Guoying Qi, Shriya Raj, Pedro Raposo, Daniel L. Rice, Rabie Saidi, Rafael Santos, Elena Speretta, James Stephenson, Prabhat Totoo, Edward Turner, Nidhi Tyagi, Preethi Vasudev, Kate Warner, Xavier Watkins, Rossana Zaru, Hermann Zellner, Alan J. Bridge, Lucila Aimo, Ghislaine Argoud-Puy, Andrea H. Auchincloss, Kristian B. Axelsen, Parit Bansal, Delphine Baratin, Teresa M. Batista Neto, Marie-Claude Blatter, Jerven T. Bolleman, Emmanuel Boutet, Lionel Breuza, Blanca Cabrera Gil, Cristina Casals-Casas, Kamal Chikh Echioukh, Elisabeth Coudert, Beatrice Cuche, Edouard de Castro, Anne Estreicher, Maria L. Famiglietti, Marc Feuermann, Elisabeth Gasteiger, Pascale Gaudet, Sebastien Gehant, Vivienne Gerritsen, Arnaud Gos, Nadine Gruaz, Chantal Hulo, Nevila Hyka-Nouspikel, Florence Jungo, Arnaud Kerhornou, Philippe Le Mercier, Damien Lieberherr, Patrick Masson, Anne Morgat, Venkatesh Muthukrishnan, Salvo Paesano, Ivo Pedruzzi, Sandrine Pilbout, Lucille Pourcel, Sylvain Poux, Monica Pozzato, Manuela Pruess, Nicole Redaschi, Catherine Rivoire, Christian J A Sigrist, Karin Sonesson, Shyamala Sundaram, Cathy H. Wu, Cecilia N. Arighi, Leslie Arminski, Chuming Chen, Yongxing Chen, Hongzhan Huang, Kati Laiho, Peter McGarvey, Darren A. Natale, Karen Ross, C. R. Vinayaka, Qinghua Wang, Yuqi Wang, and Jian Zhang. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 1 2023.

[86] Helen M. Berman. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 1 2000.

[87] LLC Schrödinger. Schrödinger Release 2021-1: Maestro, 2021.

[88] G. Madhavi Sastry, Matvey Adzhigirey, Tyler Day, Ramakrishna Annabhimoju, and

Woody Sherman. Protein and ligand preparation: parameters, protocols, and influence on virtual screening enrichments. *Journal of Computer-Aided Molecular Design*, 27(3):221–234, 3 2013.

[89] John E. Ladbury. Just add water! The effect of water on the specificity of protein-ligand binding sites and its potential application to drug design. *Chemistry & Biology*, 3(12):973–980, 12 1996.

[90] Manuela Maurer and Chris Oostenbrink. Water in protein hydration and ligand recognition. *Journal of Molecular Recognition*, 32(12):e2810, 12 2019.

[91] Joel Wahl and Martin Smieško. Thermodynamic Insight into the Effects of Water Displacement and Rearrangement upon Ligand Modifications using Molecular Dynamics Simulations. *ChemMedChem*, 13(13):1325–1335, 7 2018.

[92] Amr H. Mahmoud, Matthew R. Masters, Ying Yang, and Markus A. Lill. Elucidating the multiple roles of hydration for accurate protein-ligand binding prediction via deep learning. *Communications Chemistry*, 3(1):19, 2 2020.

[93] Zhe Wang, Huiyong Sun, Xiaojun Yao, Dan Li, Lei Xu, Youyong Li, Sheng Tian, and Tingjun Hou. Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: the prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18):12964–12975, 5 2016.

[94] Andrew T. McNutt, Paul Francoeur, Rishal Aggarwal, Tomohide Masuda, Rocco Meli, Matthew Ragoza, Jocelyn Sunseri, and David Ryan Koes. GNINA 1.0: molecular docking with deep learning. *Journal of Cheminformatics*, 13(1):43, 12 2021.

[95] Sunghwan Kim, Jie Chen, Tiejun Cheng, Asta Gindulyte, Jia He, Siqian He, Qingliang Li, Benjamin A. Shoemaker, Paul A. Thiessen, Bo Yu, Leonid Zaslavsky, Jian Zhang, and Evan E. Bolton. PubChem 2023 update. *Nucleic Acids Research*, 51(D1):D1373–D1380, 1 2023.

[96] Jason J. Han. FDA Modernization Act 2.0 allows for alternatives to animal testing. *Artificial Organs*, 47(3):449–450, 3 2023.

The chemical nature of a composed molecule depends on the nature and quantity of its elementary constituents and on its chemical structure.

Alexander Mikhaylovich Butlerov

4

A Transformative Approach to Molecular Similarity Search in Drug Discovery

The advantage of the structure-based modeling approach used in PanScreen is its interpretability, as well as its potential to predict activity cliffs. However, ligand-based methods also have their merit, especially because of their generally lower computational cost.

With a growing PanScreen portfolio of dozens to hundreds of off-targets, the computational cost of evaluating all implemented off-targets will get increasingly high. Thus, it would prove beneficial to have a way of prioritizing off-targets to assess those with the highest probability of interaction first. This could be achieved by comparing a query molecule to a set of compounds known to bind to a given target using different similarity metrics. The target with the most similar known binders to the query will then be screened first. This would allow to focus on off-targets with the highest chance of interaction with the query molecule.

Therefore, our goal was to develop a method that can perform an efficient similarity screening in large databases of compounds using different similarity metrics. The following article was published in the Journal of Cheminformatics in 2023.[1] It contains a proof-of-concept study showing that it is possible to conserve molecular similarities in the form of Euclidean distances in latent space. Therefore, the computational cost of calculating the similarity between two molecules becomes the same as calculating the Euclidean distance between two points in high-dimensional space. Although tested only with 2D similarities, the article hypothesizes that this method can be used for any kind of molecular similarity, including complex alignment-based three-dimensional metrics.

Efficient Virtual High-Content Screening Using a Distance-Aware Transformer Model

Manuel S. Sellner[1]
manuel.sellner@unibas.ch

Amr H. Mahmoud[1]
amr.abdallah@unibas.ch

Markus A. Lill[1]*
markus.lill@unibas.ch

[1] Computational Pharmacy, Department of Pharmaceutical Sciences, University of Basel
* Corresponding author

February 08, 2023

4.1 Abstract

Molecular similarity search is an often-used method in drug discovery, especially in virtual screening studies. While simple one- or two-dimensional similarity metrics can be applied to search databases containing billions of molecules in a reasonable amount of time, this is not the case for complex three-dimensional methods. In this work, we trained a transformer model to autoencode tokenized SMILES strings using a custom loss function developed to conserve similarities in latent space. This allows the direct sampling of molecules in the generated latent space based on their Euclidian distance. Reducing the similarity between molecules to their Euclidian distance in latent space allows the model to perform independent of the similarity metric it was trained on. While we test the method here using 2D similarity as proof-of-concept study, the algorithm will enable also high-content screening with time-consuming 3D similarity metrics. We show that the presence of a specific loss func-

tion for similarity conservation greatly improved the model's ability to predict highly similar molecules. When applying the model to a database containing 1.5 billion molecules, our model managed to reduce the relevant search space by 5 orders of magnitude. We also show that our model was able to generalize adequately when trained on a relatively small dataset of representative structures. The herein presented method thereby provides new means of substantially reducing the relevant search space in virtual screening approaches, thus highly increasing their throughput. Additionally, the distance awareness of the model causes the efficiency of this method to be independent of the underlying similarity metric.

4.2 Introduction

4.2.1 Molecular Similarity Search

The mean financial burden of researching and developing a new drug has been estimated to exceed 1 billion US dollars[2]. Resource, cost, and time efficient methods of finding new drug molecules are therefore imperative for reducing the cost and duration of drug development. Using computer-based methods can help reach this goal.

A well-known concept in drug development is that similar molecules exhibit similar properties and activity profiles[3,4]. This can enable researchers to find novel hits by comparing them with known active substances, which is the main principle behind similarity search in drug development. Similarities between compounds can be determined by different strategies, from simple descriptor-based comparisons over 2D fingerprints to detailed 3D measures such as shape-based or field-based similarities dependent on alignment of the molecules to be compared. To calculate similarities between molecules for large-scale similarity search, typically molecular fingerprints are utilized and computed. These fingerprints encode chemical properties and usually consist of binary vectors. While traditional molecular fingerprints were mainly rule-based (e.g. based on the presence of substructures or atom-pairs[5,6]), data driven fingerprints (e.g. learned by machine learning models) became more prominent in recent years[7]. Various metrics like the Tanimoto or Dice coefficient, or the Tversky index can be used to compute similarities based on these binary fingerprints[4].

There is a large variety of molecular fingerprints, ranging from simple fragment-based 2D methods to complex 3D approaches[3,8]. 2D based fingerprints can easily be applied to virtual screenings of multi-million compound databases (up to several billion)[9,10]. While this is possible in a relatively short period of time due to their low complexity, more complicated 3D similarity measures such as shape screening and similarity based on field points are realistically only feasible to use on smaller datasets of several hundred thousands up to a few million compounds[11,12].

Here, we present a different approach to the problem of high-content similarity screening combining transformer-based autoencoder models, similarity-based latent space shaping, and direct sampling in the reduced latent space representation. In this current proof-of-concept study presented here, we demonstrate the feasibility of the approach using 2D fingerprint similarities. We show that our approach can capture molecular similarities very well in latent space. The performance of the presented model is, however, independent of the used similarity metric. This allows researchers to train a model on highly complex 3D similarity metrics and thus perform high-content screening using metrics that otherwise would not be feasible to apply to a large set of compounds. Since the presented problem falls under the domain of distance metric learning[13,14], we show how to overcome this obstacle by implementing a custom loss function specifically designed to map similarities to Euclidian distances.

4.2.2 Related Work

Since the goal of this project is to group similar samples closer together in latent space while pushing dissimilar samples further apart, it shares similarities with contrastive learning approaches[15,16]. Contrastive learning has been widely used in visual learning with great success[17–19]. Recently, it has also been applied to molecular data, not only in a supervised but also in a self- or unsupervised fashion[20–22]. Self-supervised methods have the advantage that they do not rely on the explicit labeling of positive (similar) and negative (dissimilar) samples. When it comes to molecular data, self-supervision is feasible in 2D space by slightly altering substructures of molecules to obtain positive samples. However, when moving to 3D representations, altering substructures may lead to large differences in the 3D conformation of a molecule, where it is not guaranteed that the newly generated structure is still similar to the original. Furthermore, our approach differs from contrastive learning by providing a continuous measure of similarities to allow for a ranking of molecules according to their similarity to a template.

The use of deep learning models to create latent space embedding of molecules is not novel and has been used for several years now[23,24]. However, to our knowledge, this is the first time that the generated latent space was explicitly shaped in a way that allows the direct conservation of molecular similarities without having to rely on the discrimination of the data into different classes and without losing the direct scalability to higher dimensional representations.

A well established approach of learning chemical properties of molecules is by using so called autoencoders[25–28]. An autoencoder is a model that attempts to encode its input into latent space and decodes it again while minimizing the difference between the input and the decoded output. The latent space can be considered a reduced representation of the underlying structures of the chemicals in the dataset. Herein, we make use of an autoencoder in order to learn similarities of molecules. Honda et al. previously used a transformer model to generate molecular fingerprints from SMILES strings using a simple reconstruction loss function[25]. Bjerrum et al. found that mapping enumerated to canonical SMILES improves the conservation of similarities in latent space[26].

As mentioned before, conserving similarities in latent space is not only of high relevance in drug discovery but also in other fields such as image recognition. Schroff et al.[29] proposed a loss function called triplet loss (Equation 4.1) which can be used to map related images to similar regions in latent space while increasing the distance between dissimilar images:

$$L(A, P, N) = max(||f(A) - f(P)|| - ||f(A) - f(N)|| + m, 0) \tag{4.1}$$

This loss function relies on the definition of an anchor (A), a positive (i.e. similar) sample (P), and a negative (i.e. dissimilar) sample (N) and is therefore well suited for data with discrete labels. $f(\cdot)$ describes the coordinates of a compound in latent space, $|| \cdot ||$ the L2-norm, and m the hyperparameter specifying a margin to separate similar from non-similar molecules.

In this work, we follow the approach of Honda et al. and use a transformer model to autoencode SMILES strings to generate fingerprints suitable for similarity calculations[25]. We then use the generated latent space encodings for similarity search based on Euclidian distances. In order to improve the similarity conservation in latent space, we compare a model based only on a reconstruction loss with models trained on additional loss terms to specifically learn similarities. Since the triplet loss function in Equation 4.1 requires discrete labels, working with similarities requires the definition of a similarity threshold separating similar molecules from dissimilar ones. As such a separation is highly ambiguous for diverse sets of molecules, we developed a novel loss function which we call the similarity loss function. The similarity loss function can be used to work with continuous data, rendering it well-suited for working with similarities.

The herein presented models are therefore intended to estimate similarities based on Euclidian distances in latent space, allowing the subsequent use of exhaustive similarity search on a drastically reduced search space. We also show that a model trained on a small dataset is able to generalize to huge compound libraries containing highly diverse structures.

4.3 Methods

4.3.1 Model Architecture

In recent years, transformer-based models witnessed great success in various areas such as natural language processing, speech recognition, object detection, and more[30–34]. In this work, we follow the initial transformer model architecture proposed by Vaswani et al.[35]. Figure 4.1 shows a representation of the implemented model architecture. To encode simple SMILES representations of molecules, we first tokenized the strings, embedded them and added a positional encoding. An example of a tokenized SMILES string can be found in Figure A4.3. The positional encoding is done using a set of sine and cosine functions of varying frequencies as indicated in Equation 4.2 where *pos* refers to the position of the token in the sequence, d is the size of the embedding, and i is the dimension of the embedding. In this study, we set $d = 256$.

$$\begin{aligned} PE(pos, 2i) &= sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \\ PE(pos, 2i+1) &= cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \end{aligned} \tag{4.2}$$

The pre-processed data are then passed to a transformer encoder consisting of four layers. Each layer contains a multi-head attention layer. In this model, we used four heads per attention layer. To compute the attention, we follow the original article where attention is defined

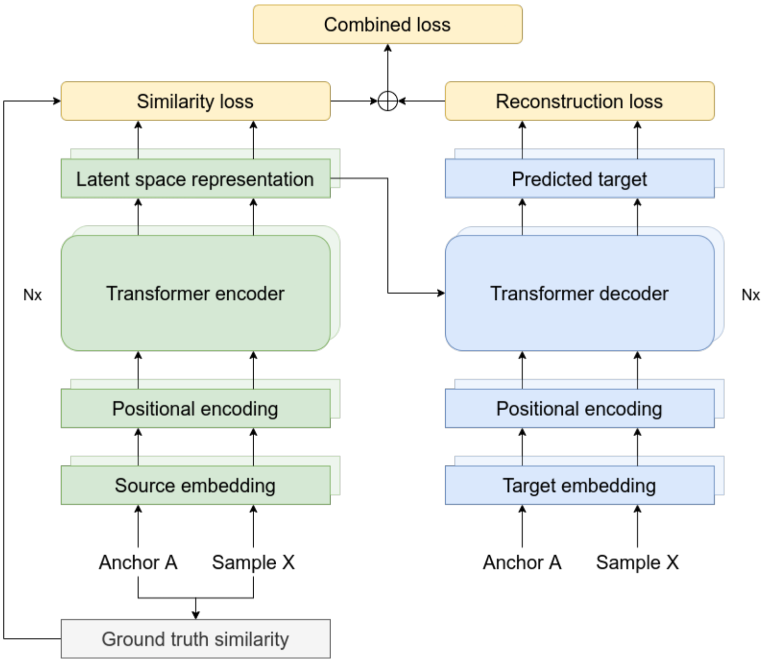


Figure 4.1: Architecture of the used transformer model. Encoder and decoder layers are constructed following the original publication of the transformer model by Vaswani et al.[35]. To help conserve similarities in latent space, a special loss function denoted as "similarity loss" is added to the reconstruction loss.

as shown in Equation 4.3 where Q, K, and V are matrices containing the queries, keys, and values, respectively, and d_k is the dimensionality of the keys[35].

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{4.3}$$

This encoder computes a latent space representation of the input. To obtain a single vector representation for each source molecule, we average over all tokens in the sequence. For the decoder part, we feed the tokenized target SMILES to an embedding layer and add a positional encoding the same way it was done for the encoder part. Note that since we are working with an autoencoder, the source and target represent the same SMILES string while the target is right shifted. This means that the matrices containing the queries, keys, and values

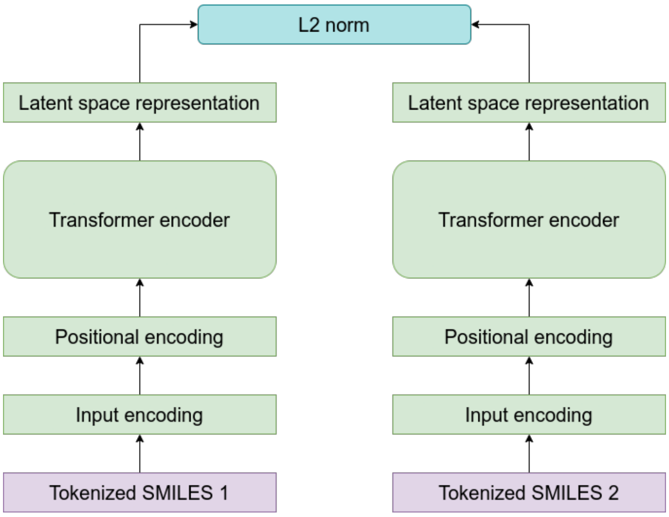


Figure 4.2: Predicting similarities between two molecules. The L2 norm is used to calculate the distance in latent space based on tokenized SMILES strings.

(Equation 4.3) all contain the same information consisting of the tokenized SMILES strings. The queries and keys are used to calculate attention weights which represent the importance of each element in the SMILES string. These attention weights can then be used to compute a weighted sum of the values. The transformer decoder layers combine the predicted latent space representation of the source with the attention weights and masked target embeddings, and subsequently predict the target sequence.

In a regular transformer model, this prediction is then used to calculate the reconstruction loss usually in form of a cross entropy loss which is used to train the model. Here, we develop and test novel loss functions to conserve similarities in the produced latent space. When applying the model to predict similarities, the decoder part of the model will not be used. Similarities are calculated based solely on the latent space representation of the query molecules; the L2 norm is used to calculate the distance between two molecules in latent space (Figure 4.2). In praxis, a perfect correlation between latent space distance and ground truth similarity metric cannot be expected. Therefore, the purpose of this model is to obtain high enrichment in predicted, similar compounds to reduce the relevant search space by a significant degree. This will drastically increase the efficiency of virtual screening.

4.3.2 Similarity Conservation in Latent Space

When using a transformer model to auto-encode SMILES strings, the used loss function commonly only consists of a reconstruction term, e.g. in form of a cross entropy loss. While this may be sufficient to conserve similarities in latent space for small datasets, the model does not specifically learn relationships between molecules. The triplet loss function introduced in the previous section can be used to separate labelled samples in latent space. Since the herein presented work uses continuous data, a similarity threshold has to be defined with the intention of distinguishing between similar and dissimilar compounds. The determination of such a threshold is ambiguous and may differ between systems and their active molecules.

To better deal with the continuous nature of our data, we developed a novel loss function which we call the similarity loss (Equation 4.4).

$$L(A,X) = \left|a \cdot \|(1 - sim(A,X))\| - \|f(A) - f(X)\|\right| \tag{4.4}$$

The similarity loss depends on an anchor (A) sample much like in the triplet loss function. However, it does not have to rely on the determination of positive and negative (i.e. similar and dissimilar) samples. Instead, it compares each anchor in a batch with all other samples (X) in the same batch. Since most similarity metrics $sim(\cdot,\cdot)$ range from 0 to 1 (0 being completely different and 1 being identical), $1 - sim(\cdot,\cdot)$ can be used to convert the similarity to a relative distance. The loss function is therefore trying to set the Euclidian distance in latent space equal to the relative distance in data space. In this study we used the Tanimoto coefficient calculated based on Morgan fingerprints as similarity metric. However, the described loss function is agnostic of the used similarity metric as long as its values are in the range $[0, 1]$. In order to spread the embedded samples in latent space, we included a scaling factor a to the term describing the relative distance in data space. The complete loss function consists of the sum of reconstruction loss (here we use a cross entropy loss) and our similarity loss:

$$L(A, X) = \left| a \cdot \|(1 - sim(A, X))\| - \|f(A) - f(X)\| \right| - \sum_{I \in \{A,X\}} \sum_{i=1}^{n_I} \sum_{c} t_{i,c} \cdot log(\hat{p}_{i,c}) \tag{4.5}$$

where $t_{i,c}$ is the label of a token i, $\hat{p}_{i,c}$ is the predicted probability for class c for token i, and n_I is the number of tokens for compound I. More information about the training of the model such as the selection of anchors during the batch generation can be found in the Appendix section A2.1.

In the following subsections, we compare the performance of the presented loss functions in order to determine their suitability to conserve similarities in latent space.

4.4 Results and Discussion

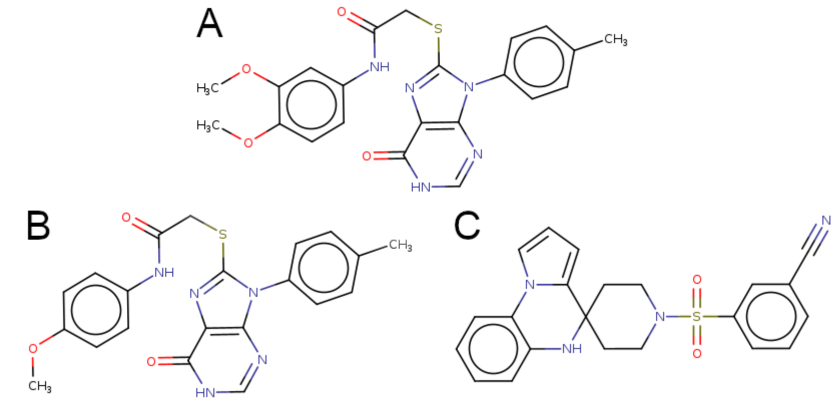
4.4.1 Initial Tests Using a Small Dataset

For a comparison of the three loss functions, the model was trained on a small dataset containing 10,000 compounds (see Appendix for details). The three models were trained using the reconstruction loss of SMILES strings (vanilla transformer), reconstruction plus triplet loss function, and reconstruction plus our newly developed similarity loss function. To compare the performance of the three models, we predicted the distances between a set of 100 randomly chosen reference compounds from the validation set and all other compounds in the dataset and compared them to the respective ground truth similarities. Based on these calculations, we computed the area under the receiver operating characteristics curve (AUROC) using different similarity thresholds to distinguish similar from dissimilar compounds. To avoid bias from the high number of dissimilar compounds leading to increased AUROC values, we only included compounds with a mimimum similarity of 0.40 to the individual reference compounds in this analysis. As shown in Table 4.1, although there were overlapping error bands, the model trained with our similarity loss function in addition to the reconstruction loss clearly outperformed the other two models. The AUROC values were above 0.90 for all tested similarity thresholds except the lowest two. For all three methods, we observed an increase in AUROC values with increasing similarity threshold. This is likely due to a negative correlation between the true positive rate and the total number of positives in a dataset.

Table 4.1: AUROC values for the different models trained on a small dataset of 10,000 compounds. While the vanilla transformer model was trained using only a reconstruction loss function, the other two models were trained with an additional loss term to specifically enforce the conservation of ground truth similarities in the latent space.

Similarity threshold	Vanilla transformer	Triplet loss	Similarity loss
0.45	0.68 ± 0.17	0.73 ± 0.17	0.82 ± 0.18
0.50	0.69 ± 0.18	0.75 ± 0.16	0.86 ± 0.17
0.55	0.75 ± 0.18	0.80 ± 0.15	0.92 ± 0.08
0.60	0.76 ± 0.18	0.81 ± 0.15	0.91 ± 0.11
0.65	0.80 ± 0.17	0.85 ± 0.13	0.94 ± 0.09
0.70	0.84 ± 0.18	0.89 ± 0.12	0.96 ± 0.07
0.75	0.87 ± 0.16	0.91 ± 0.12	0.97 ± 0.07
0.80	0.90 ± 0.14	0.94 ± 0.09	0.98 ± 0.07
0.85	0.92 ± 0.14	0.96 ± 0.08	0.98 ± 0.07
0.90	0.94 ± 0.14	0.98 ± 0.05	0.98 ± 0.08
0.95	0.97 ± 0.09	0.99 ± 0.04	1.00 ± 0.01

The vanilla model often failed to distinguish between similar and dissimilar compounds based on the Euclidian distances in latent space. The predicted distances are all very similar which likely caused a blurring in latent space, rendering it difficult to accurately distinguish between similar and dissimilar samples. While the model trained with an additional triplet loss was often able to map similar compounds closer to the reference than dissimilar compounds, it also generated a very dense latent space in which small errors can lead to incorrect



Ground truth similarity:	0.81
Approximated similarity similarity loss:	0.72
Approximated similarity triplet loss:	0.90
Approximated similarity vanilla model:	0.83

Ground truth similarity:	0.09
Approximated similarity similarity loss:	0.14
Approximated similarity triplet loss:	0.82
Approximated similarity vanilla model:	0.85

Figure 4.3: Similarity conservation in latent space. A) 2D structure of a randomly chosen reference compound. B) 2D structure of a molecule similar to the reference. Similarity was defined as having a Tanimoto coefficient above 0.8. The distances to the reference in latent space are shown for the individual models. C) 2D structure of a dissimilar molecule. Dissimilarity was defined as having a Tanimoto coefficient below 0.3. Latent space distances to the reference are shown for the individual models.

predictions. By including our custom similarity loss, the model not only learned to correctly distinguish between similar and dissimilar molecules most of the times, it also spread out the generated latent space much more, making a separation between molecules much clearer.

Figure 4.3 highlights the differences between the three models on a randomly selected example. Compound B is highly similar to compound A, whereas compound C does not share a high similarity with A. Scaling the latent space distance d_{ij} between two molecules i and j to the range $[0, 1]$ and translating them into similarities s_{ij}^{LS}, allows for a comparison of ground truth and predicted similarities in latent space:

$$s_{ij}^{LS} \approx 1 - \frac{d_{ij}}{d_{max}}, \tag{4.6}$$

where d_{max} is maximum distance between any two molecules in latent space.

By applying this formula to the compounds in Figure 4.3, we obtain approximated similarities between A and B of 0.724, 0.899, and 0.825, and between A and C of 0.139, 0.821, and 0.852 using the similarity loss model, the triplet loss model, and the vanilla model, respectively. This shows that the similarity loss model is clearly better at discriminating between similar and dissimilar molecules.

While the vanilla transformer model has no additional information about the similarity between molecules, the triplet loss function learns to group similar molecules together based on a similarity threshold. In contrast, the similarity loss function directly maps similarities to Euclidian distances and thereby, a superiority in this specific task was expected.

Based on these results, we expected the model with the additional similarity loss function to perform best, followed by the model with the triplet loss. Since the vanilla model did not have the ability of explicitly learning to couple similarities with latent space distances, we expected it to perform worst in the similarity-based virtual screening tasks.

4.4.2 Scale-up Using the ZINC Database

Training of the models was subsequently upscaled using a large dataset of around 500,000 molecules (see Appendix for details on the dataset generation). To test the optimized model, we chose a diverse set of 10 reference compounds and screened the whole downloadable ZINC database (around 1.5 billion SMILES) against each reference compound[36]. The 10 reference compounds were randomly selected from the complete ZINC database while ensuring some degree of structural diversity and making sure that the compounds were neither part of the training nor the validation set. An overview of all 10 reference compounds can be found in Figure A4.4. The goal of these models was not to achieve a perfect correlation with calculated 2D similarities but to reduce the search space to a manageable size for subsequent exhaustive similarity search. We therefore checked for each reference compound how many of the 10 most similar database entries (determined using an exhaustive search) can be found

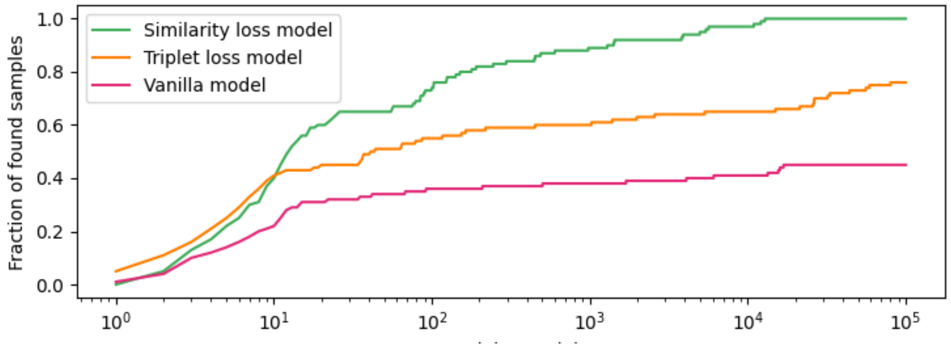


Figure 4.4: Comparison of reproduction abilities of the models with and without similarity loss function. The lines represent the normalized amount of the 10 most similar compounds within the top N closest samples in latent space for 10 reference compounds.

within the N closest samples according to each model (Figure 4.4).

The model trained with the similarity loss function proved to be effective in reproducing the top 10 most similar compounds within the 15,000 closest samples in latent space for all investigated reference compounds. This corresponds to a reduction of the search space by 5 orders of magnitude. In comparison, the vanilla model (i.e. without similarity loss function) only managed to identify 45% of all similar compounds within the top 100,000 predictions. With a identification rate of 75%, the model trained with the triplet loss was better than the vanilla model while still being worse than the model with similarity loss. To give further insights into the performance differences between the individual models, we selected three structurally different compounds from the 10 reference molecules. The first reference (**reference1**) is a large peptide with a molecular weight of more than 2000 g/mol (PubChem CID 44335764). The second (**reference2**) is a highly cyclized compound (PubChem CID 44605611) and the third (**reference3**) is a potent 5HT1B receptor antagonist (PubChem CID 44405730).

The first "ranking" analysis (Figure 4.5, middle column) shows the models' potential to correctly identify and rank the 100,000 most similar compounds from the ZINC database.

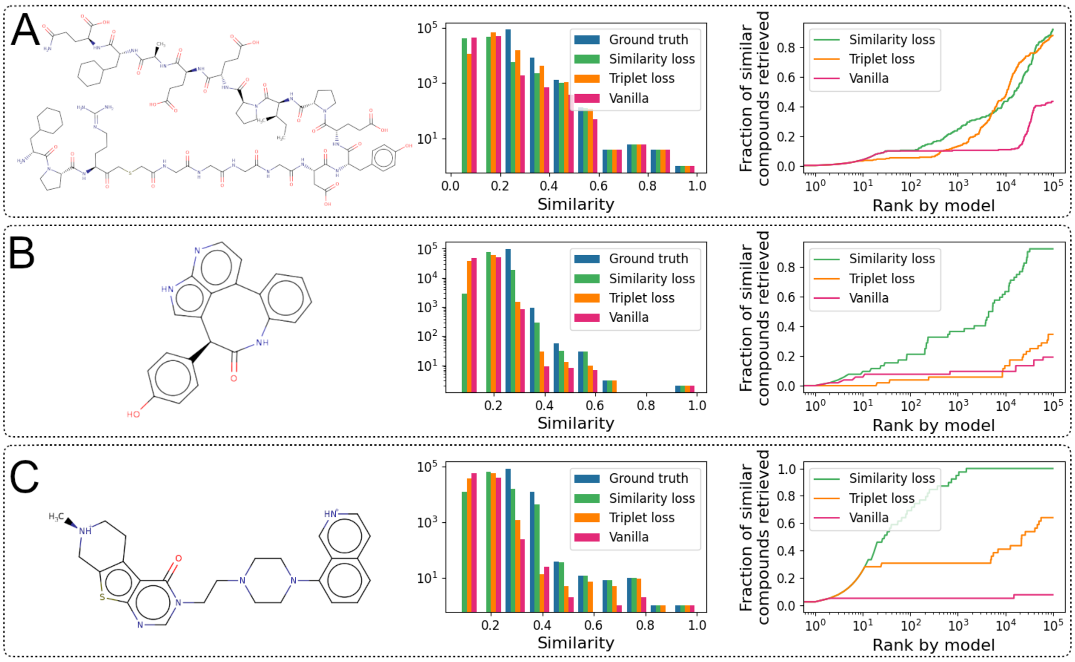


Figure 4.5: Similarity reproduction abilities. Left: 2D structure of the respective reference compound. Middle: Histogram of similarities (calculated using the exact method) of the 100,000 closest molecules to the reference in latent space (“ranking” task). Right: Reproduction of fairly similar compounds to the reference where a threshold of 0.5 was chosen to distinguish between similar and dissimilar compounds (“hit identification” task). A) analysis of the performance using a very large reference compound. B) performance with a smaller, cyclized reference compound. C) performance using a more linear compound with heterocycles.

The right column in Figure 4.5 analyses the models’ performance in identifying similar compounds to the reference (at a similarity threshold of 0.5). This analysis we name “hit identification” in the subsequent paragraphs. In general, the vanilla transformer was capable to identify similar compounds to large reference molecules such as **reference1**, but had significant difficulties for small substances, e.g. **reference3**. The same was true for the triplet loss model although the reproduction performance for the small substances was better compared to the vanilla model (Figure 4.5).

In detail, the analysis showed that all three models performed very well for **reference1** (Figure 4.5A), with the triplet loss model being slightly better at reproducing the similarity distribution of the exact metric than the other two models. In the "hit identification" task, with approximately the first 100 predictions, all models performed similarly. For the compounds ranked lower in predicted similarity to the reference, the similarity and triplet loss models started to clearly outperform the vanilla model. Within 100,000 top-ranked compounds, the similarity and triplet loss models were able to reproduce around 90% of the similar compounds whereas the vanilla model only managed to find around 40%.

For **reference2** (Figure 4.5B) and **reference3** (Figure 4.5C), the similarity loss model clearly outperformed the other two models in both "ranking" and "hit identification" tasks. For **reference2**, the similarity loss model, triplet loss model, and vanilla model were able to identify 90%, 33%, and 18% of the similar compounds, respectively. The largest difference was seen for **reference3**, where the similarity loss could identify all similar compounds within the top 2000 predictions while the vanilla model could only find around 7% of the similar compounds within the first 100,000 predictions. The triplet loss model was able to find 63% of the most similar compounds, thus performing much better than the vanilla model but still much worse than the model trained with the similarity loss. The comparatively good performance of the vanilla and triplet loss model for **reference1** is likely due to the relatively low number of very large molecules in the data set, placing those molecules in a well-separated location in latent space. The model trained on the similarity loss however performed well in all three cases, proving the advantage of the additional loss term.

Exclusion of Scaling Factor in Loss Function

To study the importance of the scaling factor in the similarity loss function (Equation 4.4), we trained an additional model with a scaling factor of 1, thus disabling its effect. Using the same analyses as previously discussed revealed a drop in accuracy compared to using larger scaling factors, although it still performs better than the vanilla model (Figure A4.5). These findings have likely to do with the fact that a well structured latent space that is not too densely packed may be important for a good reproduction performance.

Finding a good value for the scaling factor is not trivial and this hyperparameter has to be tuned during training. In our tests, we found a value of 20 to work well for the initial analyses with a smaller dataset. However, when moving to a larger set, we found that decreasing the scaling factor to 10 further improves the performance of the model.

4.5 Conclusion

In this work, we developed models for similarity-based high-content screening with the aim to translate pairwise similarities in data space to Euclidian distances in latent space. This will facilitate efficient similarity searches independent of similarity metrics. We could show that the use of a loss function specifically designed to conserve molecular similarities in latent space greatly improved the accuracy of the model. By training a transformer autoencoder using a novel similarity loss function, it was possible to obtain a model that could be successfully used for similarity search against a database of more than 1 billion compounds. We demonstrated that our model was able to generalize from a comparatively small dataset, making it possible to learn highly complex similarity metrics that could otherwise not be applied to large datasets. While the presented model did not obtain a perfect correlation to the underlying ground truth similarity metric, it can be used to substantially reduce the available search space by five orders of magnitude. Such a drastic reduction of search space allows for subsequent use of exhaustive classical screening methods.

Here, we provide a proof of concept showing the possibility of generating a model for similarity search that is unaware of the underlying similarity metric, thereby uncoupling its efficiency from the chosen method. For future adaptation of the method to 3D similarities, we will explore whether SMILES representations are sufficient as input or representations such as 3D graphs are necessary to allow the model to effectively learn 3D information. The proposed loss function for latent space shaping, however, will be not affected by this potential architecture change, as it is agnostic of the specific similarity metric.

Availability of data and materials

The code used to train the model and screen the database can be found on GitHub (`https://github.com/mmodbasel/HighContentScreening`).

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

A.H.M. and M.A.L. conceived the presented idea. M.S.S. implemented the idea, performed the computations and analysis. A.H.M. and M.A.L. supervised the findings of this work. All authors discussed the results. M.S.S. wrote the manuscript with input of all authors.

4.6 Appendix

A1. Additional Results and Discussion

To further investigate the reproduction abilities of the model with and without similarity loss, we analyzed the distribution of molecular weights of the 100,000 molecules predicted to be closest to the reference (Figure A4.1).

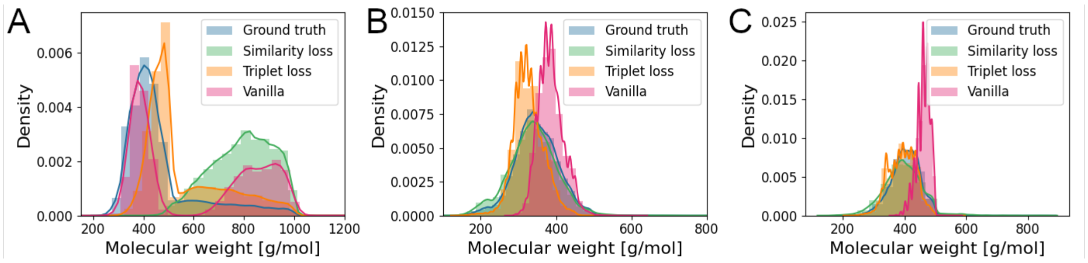


Figure A4.1: Reproduction of molecular weights. The histograms show the distribution of molecular weights of the 100,000 most similar compounds to **reference1** (A), **reference2** (B), and **reference3** (C) calculated using either the exact similarity metric, the model with similarity loss, or the vanilla transformer model.

The data show that all three models are well able to reproduce the molecular weight distribution of the 100,000 most similar compounds to **reference1** while the triplet loss model outperforms the other two models. This effect is the most pronounced at the lower end of the scale where the vanilla and triplet loss models are able to reproduce more of the low molecular weight compounds than the similarity loss model. More detailed analysis of this phenomenon revealed that these low molecular weight compounds are all highly dissimilar to the reference compound. When only including compounds with a similarity to **reference1** of 0.3 or more, these compounds disappeared and the similarity loss model showed a better overlap with the ground truth. Still, the triplet loss model showed a slightly better reproduction of the molecular weights than the other two models (Figure A4.2). The sampling of very dissimilar molecules may be due to the fact that the vanilla and triplet loss models generated a much denser latent space, leading to a generally lower distance between the very high molecular weight compounds and the molecules with lower molecular weight. While this benefits the two models for **reference1**, it decreases their performance for **reference2** and **reference3** (Figure A4.1 B & C). In these examples, the model with similarity loss is generally better able to reproduce the distribution of molecular weights from the underlying (exact) similarity metric. Here, the vanilla transformer model is likely suffering because there are a lot of molecules in the screened data set that have a similar molecular weight to the two reference compounds. This causes the model to over-sample these compounds in the densely packed latent space. In these cases, the sparser latent space generated by the similarity loss may prevent such an over-sampling.

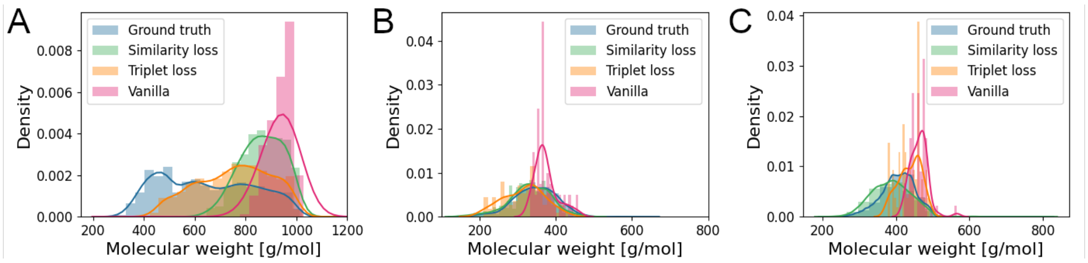


Figure A4.2: Reproduction of molecular weights. The histograms show the distribution of molecular weights of the 100,000 most similar compounds to **reference1** (A), **reference2** (B), and **reference3** (C) calculated using either the exact similarity metric, the model with similarity loss, or the vanilla transformer model. Only compounds with a similarity of at least 0.3 are considered.

For **reference 2**, the triplet loss model performed similarly to the vanilla model (Figure A4.1B). However, for **reference 3** it appears that the triplet loss model has about the same performance as the similarity loss model (Figure A4.1C). Here, it must be noted that while the 100,000 sampled compounds have a similar distribution of the molecular weight, only 96 had a similarity of at least 0.3 to the reference compound (compared to 8,236 for the similarity loss model, Figure A4.2C). Thus, while in some cases the triplet loss model is able to nicely reproduce compounds with a similar molecular weight compared to the ground truth, it still lacks the ability to find compounds with high structural similarity to the reference.

A2. Additional Materials and Methods

The following section will describe the detailed neural network architecture, its hyperparameters, and the datasets used to train and test the model.

A2.1 SMILES Transformer

Our model uses a transformer architecture as described in the publication by[35]. It was implemented in PyTorch using their integration of the Transformer module. The vocabulary was generated using tokenized SMILES strings that were used as input and encoded into 256 dimensional latent space. Our model consisted of 4 encoder and decoder layers with attention layers containing 4 heads. All models were trained using an Adam optimizer with a learning rate of 10^{-4} and 128 samples per batch. Since it was not possible to further increase the batch size due to memory limitations, we accumulated the gradients over 4 batches.

In order to determine the ground truth similarities, we calculated the Tanimoto coefficients based on 1024 bit Morgan fingerprints implemented in RDKit with a radius of 2. To conserve similarities in latent space, it is imperative that during training, each batch contains at least one similar compound to each sample (and for the triplet loss also at least one dissimilar compound). For the model trained on the similarity loss, we first randomly assigned compounds to a batch which act as anchor. To guarantee that similar compounds exist for each of those reference compounds, the algorithm randomly selected 3 of the 100 most similar compounds to the reference which were added to the batch. For the model with the triplet loss, we randomly selected 64 anchors per batch and for each chose a random compound with a Tanimoto similarity to the anchor of at least 0.6. It was assumed, that due to the intrinsic diversity of the dataset, for each anchor in a batch, there will always be a negative sample present. We defined negative samples as any compound with a similarity of less than 0.4 to the anchor.

The scaling factor a required by the similarity loss function was set to 20.0 in the initial tests on a small dataset and was later decreased to 10.0 for the scaled up training. The margin m for the triplet loss function was set to 1.0 for the comparison of the loss functions as well as for the scaled up model. These values were determined based on the retrospective analysis of the performance of each trained model.

Training a model with the similarity loss and the hyperparameters described above for 1000 epochs took roughly 9 days on a single GTX 1080 Ti.

A2.1.1 Datasets

During an initial test phase, we used a randomly selected subset of 10,000 SMILES extracted from the natural compounds dataset obtained from the ZINC database. The dataset was randomly split into a training (80%) and validation (20%) set. The validation set was used to compare the performances of three different loss functions. In the upscaling experiments, we randomly selected 0.03% of the compounds in each tranche downloaded from the ZINC database, leading to a dataset consisting of approx. 500,000 compounds. Following the method of the initial test, the dataset was randomly split into a training and validiation set using a 80/20 split. For testing the optimized model, the whole ZINC database was used which consisted of around 1,458,000,000 compounds at the time of testing.

For reproducibility, all used SMILES strings were converted to their canonical form using openbabel prior to training and testing.

A2.2 Similarity Search

Once obtained, the distance aware SMILES embeddings were used to efficiently calculate distances (i.e. similarities) in embedding space. Facebook's faiss was utilized for this task using a FlatL2 index to calculate Euclidian distances in latent space. Faiss allows the construction and search of several types of indexes with various degrees of approximation.

The search was performed on pre-calculated latent space embeddings of the whole ZINC database. Searching 94 reference compounds against the complete database took roughly

2.75 hours on a machine with 64GB RAM that was equipped with an HDD. Around 65% of the computation time was needed to read the pre-computed embeddings from disk. By using either a server with solid state drives or more memory, the computational cost could therefore be significantly decreased. Searching the same database using RDKit's BulkTanimotoSimilarity function (with pre-computed fingerprints) on the same machine required around 3.40 hours for a single reference compound.

A3. Additional Figures

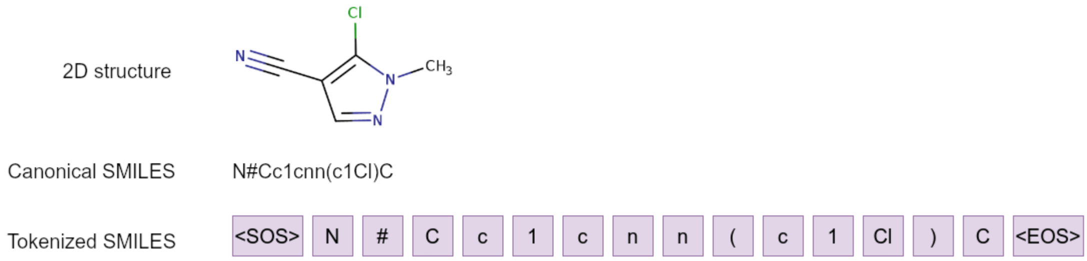


Figure A4.3: Example of SMILES tokenization. The 2D structure of a molecule, its SMILES representation, and the tokenized SMILES are shown. "<SOS>" and "<EOS>" represent labels specifying the start and the end of the sequence, respectively.

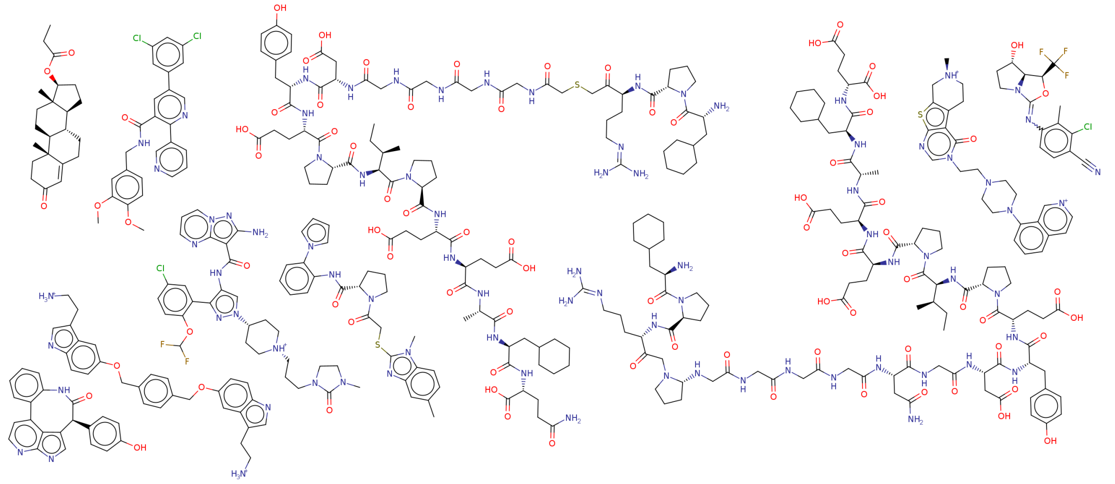


Figure A4.4: All reference compounds used for the assessment of the reproduction ability.

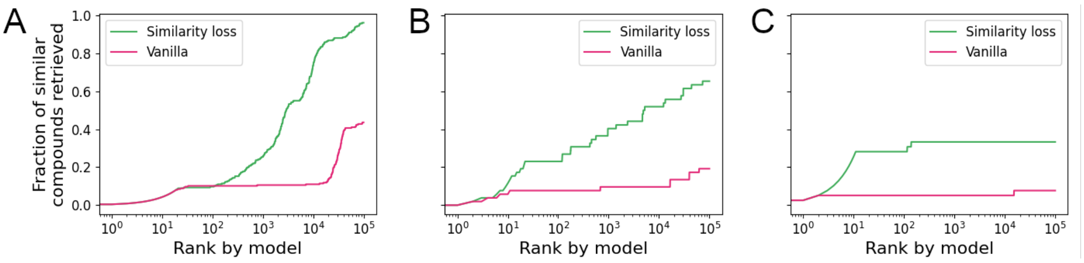


Figure A4.5: Performance of the model trained with the similarity loss scaling factor set to 1 for the “hit identification” task. The data for **reference1** (A), **reference2** (B), and **reference3** (C) are shown.

References

[1] Manuel S. Sellner, Amr H. Mahmoud, and Markus A. Lill. Efficient virtual high-content screening using a distance-aware transformer model. *Journal of Cheminformatics*, 15:18, 2 2023.

[2] Olivier J. Wouters, Martin McKee, and Jeroen Luyten. Estimated Research and Development Investment Needed to Bring a New Medicine to Market, 2009-2018. *JAMA*, 323(9):844, mar 2020.

[3] Ashutosh Kumar and Kam Y. J. Zhang. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Frontiers in Chemistry*, 6(JUL):315, jul 2018.

[4] Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, 11(2):137–148, feb 2016.

[5] Mahendra Awale and Jean-Louis Reymond. Atom Pair 2D-Fingerprints Perceive 3D-Molecular Shape and Pharmacophores for Very Fast Virtual Screening of ZINC and GDB-17. *Journal of Chemical Information and Modeling*, 54(7):1892–1907, jul 2014.

[6] Alice Capecchi, Daniel Probst, and Jean-Louis Reymond. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *Journal of Cheminformatics*, 12(1):43, dec 2020.

[7] Bulat Zagidullin, Ziyan Wang, Yuanfang Guan, Esa Pitkänen, and Jing Tang. Comparative analysis of molecular fingerprints in prediction of drug combination effects. *Briefings in Bioinformatics*, 22(6):1–15, nov 2021.

[8] Seth D. Axen, Xi-Ping Huang, Elena L. Cáceres, Leo Gendelev, Bryan L. Roth, and Michael J. Keiser. A Simple Representation of Three-Dimensional Molecular Structure. *Journal of Medicinal Chemistry*, 60(17):7393–7409, sep 2017.

[9] André Fischer, Manuel Sellner, Santhosh Neranjan, Martin Smieško, and Markus A. Lill. Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. *International Journal of Molecular Sciences*, 21(10):3626, may 2020.

[10] Adrià Cereto-Massagué, María José Ojeda, Cristina Valls, Miquel Mulero, Santiago Garcia-Vallvé, and Gerard Pujadas. Molecular fingerprint similarity search in virtual screening. *Methods*, 71(C):58–63, jan 2015.

[11] Fabien Fontaine, Evan Bolton, Yulia Borodina, and Stephen H. Bryant. Fast 3D shape screening of large chemical databases through alignment-recycling. *Chemistry Central Journal*, 1(1):12, dec 2007.

[12] Ya Chen, Neann Mathai, and Johannes Kirchmair. Scope of 3D Shape-Based Approaches in Predicting the Macromolecular Targets of Structurally Complex Small Molecules Including Natural Products and Macrocyclic Ligands. *Journal of Chemical Information and Modeling*, 60(6):2858–2875, jun 2020.

[13] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle Loss: A Unified Perspective of Pair Similarity Optimization. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6397–6406, feb 2020.

[14] Juan Luis Suárez-Díaz, Salvador García, and Francisco Herrera. A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms, Experimental Anal-

ysis, Prospects and Challenges (with Appendices on Mathematical Background and Detailed Algorithms Explanation). *ArXiv*, dec 2018.

[15] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A Survey on Contrastive Self-Supervised Learning. *Technologies*, 9(1):2, dec 2020.

[16] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9, pages 297–304. JMLR Workshop and Conference Proceedings, mar 2010.

[17] R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. Learning deep representations by mutual information estimation and maximization. *7th International Conference on Learning Representations, ICLR 2019*, aug 2018.

[18] Ishan Misra and Laurens van der Maaten. Self-Supervised Learning of Pretext-Invariant Representations. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6706–6716, dec 2019.

[19] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *37th International Conference on Machine Learning, ICML 2020*, PartF168147:1575–1585, feb 2020.

[20] Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying WEI, Wenbing Huang, and Junzhou Huang. Self-Supervised Graph Transformer on Large-Scale Molecular Data. *Advances in Neural Information Processing Systems*, 33:12559–12571, 2020.

[21] Daiki Koge, Naoaki Ono, Ming Huang, Md. Altaf-Ul-Amin, and Shigehiko Kanaya. Embedding of Molecular Structure Using Molecular Hypergraph Variational Autoencoder with Metric Learning. *Molecular Informatics*, 40(2):2000203, feb 2021.

[22] Sheng Wang, Yuzhi Guo, Yuhong Wang, Hongmao Sun, and Junzhou Huang. SMILES-BERT: Large Scale Unsupervised Pre-Training for Molecular Property Prediction. In *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pages 429–436, New York, NY, USA, sep 2019. ACM.

[23] Robin Winter, Floriane Montanari, Frank Noé, and Djork-Arné Clevert. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical Science*, 10(6):1692–1701, feb 2019.

[24] Rafael Gómez-Bombarelli, Jennifer N. Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D. Hirzel, Ryan P. Adams, and Alán Aspuru-Guzik. Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Central Science*, 4(2):268–276, feb 2018.

[25] Shion Honda, Shoi Shi, and Hiroki R. Ueda. SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery. *ArXiv*, nov 2019.

[26] Esben Bjerrum and Boris Sattarov. Improving Chemical Autoencoder Latent Space and Molecular De Novo Generation Diversity with Heteroencoders. *Biomolecules*, 8(4):131, oct 2018.

[27] Seung Hwan Hong, Seongok Ryu, Jaechang Lim, and Woo Youn Kim. Molecular Gen-

erative Model Based on an Adversarially Regularized Autoencoder. *Journal of Chemical Information and Modeling*, 60(1):29–36, jan 2020.

[28] Chaochao Yan, Sheng Wang, Jinyu Yang, Tingyang Xu, and Junzhou Huang. Rebalancing Variational Autoencoder Loss for Molecule Sequence Generation. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, volume 20, pages 1–7, New York, NY, USA, sep 2020. ACM.

[29] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June, pages 815–823. IEEE, jun 2015.

[30] Ishan Misra, Rohit Girdhar, and Armand Joulin. An End-to-End Transformer Model for 3D Object Detection. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2886–2897, sep 2021.

[31] Yangyang Shi, Yongqiang Wang, Chunyang Wu, Ching-Feng Yeh, Julian Chan, Frank Zhang, Duc Le, and Mike Seltzer. Emformer: Efficient Memory Transformer Based Acoustic Model For Low Latency Streaming Speech Recognition. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021-June:6783–6787, oct 2020.

[32] Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. ParsBERT: Transformer-based Model for Persian Language Understanding. *Neural Processing Letters*, 53(6):3831–3847, may 2020.

[33] Mohammad A. Hannan, Dickson N. T. How, M. S. Hossain Lipu, Muhamad Mansor, Pin Jern Ker, Zhao Y. Dong, Khairul S. M. Sahari, Sieh K. Tiong, Kashem. M. Muttaqi,

T. M. Indra Mahlia, and Frede Blaabjerg. Deep learning approach towards accurate state of charge estimation for lithium-ion batteries using self-supervised transformer model. *Scientific Reports*, 11(1):19541, dec 2021.

[34] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186, oct 2018.

[35] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-Decem, pages 5999–6009. Neural information processing systems foundation, jun 2017.

[36] Teague Sterling and John J. Irwin. ZINC 15 – Ligand Discovery for Everyone. *Journal of Chemical Information and Modeling*, 55(11):2324–2337, nov 2015.

The only way of discovering the limits of the possible is to venture a little way past them to the impossible.

Arthur C. Clarke

5

Going Beyond 2D: Transformer-Based Ligand Screening in 3D Space

In the previous chapter, we presented a proof-of-concept study showing that molecular similarities can be conserved in a latent space generated by a Transformer model. We hypothesized that this is possible independent of the underlying similarity metric, but we showed results only for 2D similarities. The following article was published as a preprint on BioRxiv and proves our previous claims by using an alignment-based 3D similarity measure.[1]

Enhancing Ligand-Based Virtual Screening with 3D Shape Similarity via a Distance-Aware Transformer Model

Manuel S. Sellner[1,2]
manuel.sellner@unibas.ch

Amr H. Mahmoud[1,2]
amr.abdallah@unibas.ch

Markus A. Lill[1,2*]
markus.lill@unibas.ch

[1] Computational Pharmacy, Department of Pharmaceutical Sciences, University of Basel
[2] SIB Swiss Institute of Bioinformatics
[*] Corresponding author

November 17, 2023

5.1 Abstract

Following the assumption that chemically similar molecules exhibit similar biologcial properties, ligand-based virtual screening can be a valuable starting point in drug discovery projects. While 2D-based similarity metrics generally focus on similar scaffolds or substructures, 3D-based methods can capture the shape of a molecule, allowing for the identification of compounds with different scaffolds. We recently published a proof-of-concept study which demonstrated how a Transformer model can be adapted to preserve 2D similarities in latent space in the form of Euclidean distances. In this work, we extend this research and prove that the approach can be adapted to 3D similarities. We use pharmacophore-based shape similarity as 3D similarity measure. We show that the model is able to enrich the predicted most similar

hits with compounds with different scaffolds that are indeed similar in 3D space. Whereas classical pharmacophore- or shape-based 3D similarity methods rely on expensive alignment processes, in our approach, we identify similar compounds directly by the Euclidean distances in latent space. This enables for the first time the 3D screening of ultra-large databases with high efficiency.

5.2 Introduction

Virtual screenings of ultra-large compound libraries are receiving increasing attention in the scientific community[2–4]. While some virtual screening strategies rely on structure-based approaches[5], most use ligand-based methods due to their simplicity and computational efficiency[6–9]. In this work, we explore a novel method to accelerate 3D ligand-based ultra-large virtual screening using a Transformer-based deep neural network.

5.2.1 Similarity Search

Ligand-based similarity methods are widely used because of their computational efficiency, which is orders of magnitude faster than classical structure-based virtual screening methods. Ligand-based similarity concepts assume that chemically similar molecules exhibit similar biological activity[10,11]. Common 2D similarity search consists of the extraction and comparison of molecular features. Usually, these features are stored in binary vectors (fingerprints) which can be easily compared using, e.g., the Tversky index, Tanimoto, or Dice coefficient. Due to the simplicity of this approach, these methods are usually fast enough to allow the screening of more than a billion compounds in a matter of hours.

On the other hand, 3D similarity methods are usually more computationally demanding. While alignment-free methods may still be computationally feasible[12], the usually more accurate alignment-based methods, such as pharmacophore or shape screening, come with an increased computational cost[13]. Nevertheless, it has been shown that even these computationally demanding methods can be used to screen billions of compounds with the right hardware. For example, Michino et al. screened approximately 1.12 billion compounds in 19.5 hours using 216 GPUs[14]. Given that not everyone has access to such significant computational resources, we believe it is essential to accelerate these accurate yet comparatively slow ligand-based screening methods.

5.2.2 Representation Learning

In the field of similarity search, a recurring issue is the maximization of the information content in a molecular fingerprint[15,16]. Since in this work we focus on the use of deep neural networks, this problem falls into the domain of representation learning. Representation learning aims to transform raw high-dimensional data into a reduced set of features that can be used to optimally represent the data and enable their use in downstream tasks[17,18]. Representation learning has been used in various fields such as language processing[19], time series[20], optimization of industrial processes[21], investigation of biological sensorimotor integration[22], and molecular property prediction[23].

A common approach is contrastive representation learning, in which similar samples are trained to be close together in embedding space, while dissimilar samples should be farther apart[24]. Thus, in contrastive learning, input samples are compared to each other. This allows for the use of unsupervised learning as long as input samples can be compared with a defined similarity metric. One of the earliest contrastive loss functions was developed by Chopra et al. and is used to cluster samples of the same class in a similar location in embedding space[25]. Other important loss functions used in contrastive representation learning include the triplet loss[26], lifted structured loss[27], N-pair loss[28], and noise contrastive estimation (NCE) loss[29].

Generative representation learning is another important category of representation learning[24]. In generative representation learning, a model is trained to generate new samples (or reconstruct samples from the input). The concept is that, for a model to generate realistic samples, it must learn the fundamental structure of the data.

5.2.3 Previous Work

In a proof-of-concept study, we recently demonstrated that it is possible to train a deep neural network model to create a similarity conserving latent space[30]. We demonstrated that the latent space can be shaped in such a way that allows to use the Euclidean distance between embedded molecules as a measure of their similarity. This was done using a combination of generative and contrastive representation learning. The utilized Transformer-based model reconstructed SMILES strings that were given as input. At the same time, it used a custom similarity loss for contrastive learning of continuous molecular similarities (see Equation 5.2).

Using this model, it was possible to reduce the search space for virtual screening by several orders of magnitude. For simplicity, we used a simple 2D similarity measure based on Morgan fingerprints. However, due to the low computational cost of calculating 2D similarities, training such a model does not give a significant benefit over directly using the underlying similarity metric. Here, we extend this work and adapt the model to computationally expensive alignment-based 3D similarity metrics, which results in significant improvements in efficiency compared to other 3D similarity methods.

5.2.4 Challenges Going From 2D to 3D

To utilize 3D similarity metrics, the architecture of the model needs to be modified to allow for 3D structural data as input instead of 1D SMILES strings. Here, we represent molecules in the form of graphs, where atoms are nodes, and bonds are edges. This approach also allows to include 3D distance information as part of the edge featurization. A detailed description of this process can be found in Section 5.4.3.

Arguably, the biggest challenge is the computational cost of complex 3D similarity calculations. For the model trained on 2D similarities in our previous proof-of-concept study, it was possible to either calculate all pairwise similarities in the training set before training or using online learning by calculating the similarities on the fly during training. When using alignment-based 3D similarity metrics, it is not feasible to calculate all pairwise similarities for a large dataset. Also, online learning would be simply too slow. One way to overcome this problem is to use active learning methods. Active learning is a technique to sample from unlabeled data and choose new samples to annotate and add to the training set based on a certain algorithm in order to maximize the model's improvement[31,32]. There are several algorithms (acquisition functions) that are often used in active learning. Regardless of the specific algorithm, their goal is always to select the best data to learn from in order to boost the model's performance as efficiently as possible. The use of active learning therefore allows to start training on a small training set which is iteratively grown based on the selection of the implemented algorithm(s). In our case, this has the advantage that only a small portion of the data has to be annotated (i.e. similarities have to be calculated) before the training. Each active learning cycle only adds new samples that are beneficial for the model's training, thus making the whole training process more efficient.

One commonly used active learning acquisition function is called query by committee (QBC). This algorithm employs a committee of models (so-called students). New samples for annotation are selected based on the maximum disagreement in prediction between the student models[33–35]. Therefore, a normal QBC algorithm requires multiple models to be trained in parallel. Since this comes with an additional computational cost, a committee can also be simulated by using the same model with activated dropout for the predictions. This method is also called query by dropout committee[36].

Another popular active learning algorithm is expected model change maximization (EMC). In this algorithm, the gradient of the loss with respect to an input sample is used to estimate the expected change of the model when learning from the sample[37,38]. In practice, a set of unlabeled samples is passed through the model. For each sample, the gradient of the loss with respect to the input is calculated and the samples with the largest gradient are chosen for annotation. Since it is necessary to calculate the loss for unlabeled samples, this method cannot be used for loss functions that require labels.

5.2.5 Our Contribution

In this work, we extend our previous proof-of-concept study using 2D similarities to the use of 3D similarity metrics for efficient high-content virtual screening. We present the necessary modifications to the model architecture and the training process to enable the training on computationally expensive alignment-based similarity metrics; here shape screening implemented within the Schrödinger software suite. We also show that our model is indeed capable of conserving 3D similarities in latent space and that it can be used to efficiently identify compounds with similar 3D features. In our opinion, such a model can be very valuable in the early stages of hit identification, where the main focus is the reduction of the search space.

5.3 Results and Discussion

There are several performance criteria that our model must meet. First, since the intended use for this model is ligand-based virtual screening, it should be able to actually predict similar compounds (according to the underlying similarity metric) within the top-ranked predictions. Second, the model should actually capture 3D features and not rely solely on 2D similarities. This means that the model should be able to identify similar compounds with different chemical scaffolds. Finally, the model should prove its usefulness in a "real-world example" such as successfully reproducing known binders to a given target protein based on a reference molecule.

5.3.1 General Analysis

With this general analysis, we tested the model's ability to find similar compounds and capture 3D information. We did this by screening several query molecules against a database of structurally diverse compounds and comparing the identified hits with the hits from the pharmacophore-based 3D shape screening. This test had two desired outcomes: 1) the model's predictions correlate with the baseline similarities and the model is able to identify a high percentage of the top-ranked hits according to the baseline similarity method. 2) the top-ranked predictions have a high shape overlap with the query molecules.

To construct the dataset for the screening, we randomly selected a subset of approximately 50,000 compounds from the ZINC database[39]. Only compounds that were not part of the training set were selected. We then clustered the compounds using the Butina algorithm implemented in RDKit based on Tanimoto similarities based on Morgan fingerprints[40]. For clustering, we used a similarity cutoff of 0.7. In total, there were 31,856 clusters, of which only 5,234 contained more than one compound. This shows that the compounds had high structural diversity. We used the centroids of the 10 largest clusters as reference compounds for our analysis. This ensured that the screening set contained compounds with 2D structures similar to those of the reference compounds. For these reference compounds, we generated a single 3D conformer using Schrödinger's LigPrep[41]. To create a dataset to screen, we took up to 10 compounds from the created clusters until we had a set of 10,000 compounds. For the selected 10,000 compounds, we then created up to 5 conformers each using Schrödinger's ConfGen[42,43]. This resulted in a total of 49,495 structures to screen. To create our baseline, we used Schrödinger's pharmacophore-based GPU shape screening tool to screen the 10 selected references against the created dataset[13,44]. Our model was used to screen the same 10 query molecules against the same dataset. Because there were multiple conformations per compound, the best score (highest similarity or shortest distance in latent space) was used for both methods. Table 5.1 shows an overview of the model's performance. The mean Pearson correlation coefficient (PCC) was -0.73 ± 0.13. Note that the correlation should be negative because the model predicts distances and not similarities. Thus, the smaller the predicted distance, the higher the estimated similarity. Figure 5.1 shows the correlation between the predicted distances and the calculated shape similarities for A) ZINC001763434742 (one of the best performing queries) and B) ZINC001183157671 (one of the worst performing queries). For ZINC001763434742, the true similarities are mostly in the range from 0.15 to 0.65 while the similarities for ZINC001183157671 are mainly in a comparable small range from 0.15 to 0.5. Thus, there seem to be no compounds that are highly similar to the query in

Table 5.1: Performance analysis for 10 query molecules screened against approximately 10,000 compounds from the ZINC database. "Mean similarity top 100" shows the mean shape similarity of the top 100 calculated hits according to the baseline method. "Mean similarity top 100 pred" shows the mean predicted similarity of the top 100 predicted hits according to the model. Since the model predicts distances and not similarities, the predicted similarity $\hat{s}$ was calculated as $\hat{s} = 1 - \frac{d}{a}$ where a is the scaling factor as in Equation 5.2 and d is the predicted latent space distance.

Query	PCC	Precision top 100	Precision top 1000	Mean similarity top 100	Mean similarity top 100 pred
ZINC000570771518	-0.78	0.17	0.47	0.54	0.46
ZINC000950159323	-0.77	0.30	0.47	0.56	0.49
ZINC000954430177	-0.54	0.16	0.33	0.54	0.44
ZINC000970035445	-0.79	0.33	0.53	0.57	0.51
ZINC001183157671	-0.64	0.14	0.35	0.44	0.36
ZINC001281147597	-0.75	0.25	0.49	0.54	0.48
ZINC001368797027	-0.78	0.20	0.44	0.49	0.43
ZINC001711902206	-0.84	0.34	0.69	0.49	0.45
ZINC001740566933	-0.88	0.30	0.76	0.54	0.50
ZINC001763434742	-0.87	0.58	0.75	0.60	0.57

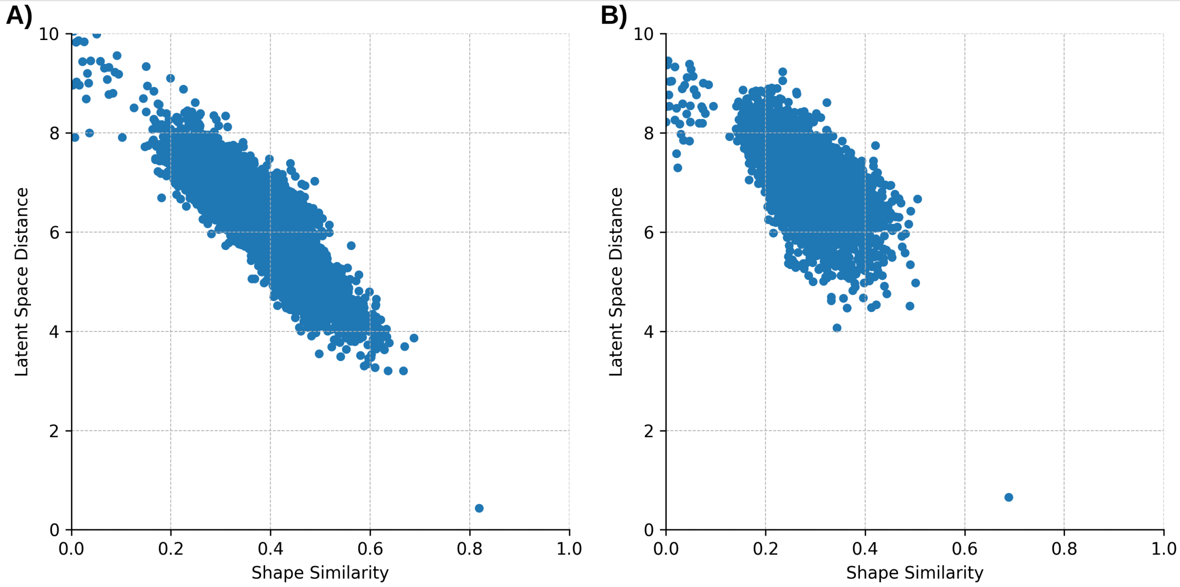


Figure 5.1: Correlation between predicted distance and calculated shape similarity for two query compounds. **A)** ZINC001763434742, $R^2 = 0.76$. **B)** ZINC001183157671, $R^2 = 0.40$

1B). This small range in similarity values contributes to the rather low correlation coefficient.

Using the latent space distance d and the scaling factor a used to train the model (see Equation 5.2), it is possible to approximate the similarity $\hat{s}$ to $\hat{s} = 1 - \frac{d}{a}$. We used this equation to calculate the similarity for the top 100 predictions and compare their mean with the mean similarity of the top 100 hits from the baseline. Ideally, these two means are the same, as this would indicate that the model was able to reproduce the true similarity values. According to Table 5.1, the difference between these values ranged from 0.03 to 0.10, indicating that the performance depends on the chosen query compound. It can also be seen that it correlates nicely with the PCC. The precision shows the fraction of the top N predicted hits that are actually among the top N according to the baseline method. For $N = 100$, these values were generally quite low, indicating that the model was not very good at reproducing the top 100 hits among the top 100 predictions. However, the values seem to correlate with the overall performance for the specific queries, as indicated by the PCC. With some exceptions, these

Table 5.2: Screening performance for 10 query molecules screened against approximately 10,000 compounds from the ZINC database. The area under the receiver operating characteristics curve (AUROC) and the enrichment factors (EF) at 1, 2, 5, and 10 percent are shown.

Query	AUROC	EF 1%	EF 2%	EF 5%	EF 10%
ZINC000570771518	0.92	17.0	14.0	10.0	6.4
ZINC000950159323	0.95	29.0	21.5	13.4	8.0
ZINC000954430177	0.88	15.0	11.0	9.0	6.1
ZINC000970035445	0.96	33.0	26.0	15.0	8.8
ZINC001183157671	0.86	14.0	13.0	8.0	5.2
ZINC001281147597	0.94	25.0	18.5	13.0	7.9
ZINC001368797027	0.92	20.0	15.0	10.6	7.0
ZINC001711902206	0.96	34.0	25.5	14.4	9.4
ZINC001740566933	0.98	30.0	27.0	17.0	9.9
ZINC001763434742	0.99	58.0	38.5	19.6	10.0

values are better for queries with a higher similarity to the top hits (represented by the mean similarity of the top 100 hits). This indicates that the model may generally perform better when there are compounds in a database that are highly similar to the query. Since reproducing the top 100 hits is a very difficult task and we did not expect the model to actually excel at it, we also calculated the precision for the top 1000 hits. There, the performance is generally higher, but varies greatly between the different queries, and again seems to correlate quite nicely with the PCC. On average, 53% ± 15% of the top 1000 predicted compounds were actually among the top 1000 hits according to shape screening.

To assess the screening performance of the model in more detail, we calculated the receiver operating characteristics (ROC) curves and enrichment factors (EF) for the 10 queries in Table 5.1. To calculate the ROC curves, we defined the first 100 hits from the shape screening as active and the rest as inactive. The goal of the model should be to accurately replicate the 100 hits as early as possible. The results of this analysis are shown in Table 5.2 and detailed ROC curves and reproduction plots can be found in the Supporting Information in Figures A5.1-A5.10. The average area under the ROC curve was 0.93 ± 0.04, indicating a very good screening performance. Also, the mean 1% EF was 27.5 ± 13.0. This means that on average the model was able to reproduce 27.5 times more active compounds than random selection when considering only the 1% top ranked predictions.

To get a full picture of the performance of the model, it is important to analyze examples in which the model predicted very wrong values. Therefore, we picked examples that had a short predicted distance while having a low calculated shape similarity. Figure 5.2 shows one of the cases where there is a large discrepancy between the predicted and calculated rank of the compound. Although the model ranked this compound 3680 ranks too high, the shape overlap with the query molecule seems to be high. In this case, the calculated similarity may be reduced due to few matching pharmacophores. This would indicate that the model has problems learning the pharmacophore information while being able to capture the 3D shape information well. Indeed, while the pharmacophore-based shape similarity between the two molecules was 0.35, the shape-only similarity (without pharmacophore information) was 0.63.

Another example in which the model overestimated a compound by 1571 ranks is given in Figure 5.3. It can again be seen that the overlap between the two compounds is rather high. In this example, the 2D similarity between the two compounds is low (0.35) and it can be seen

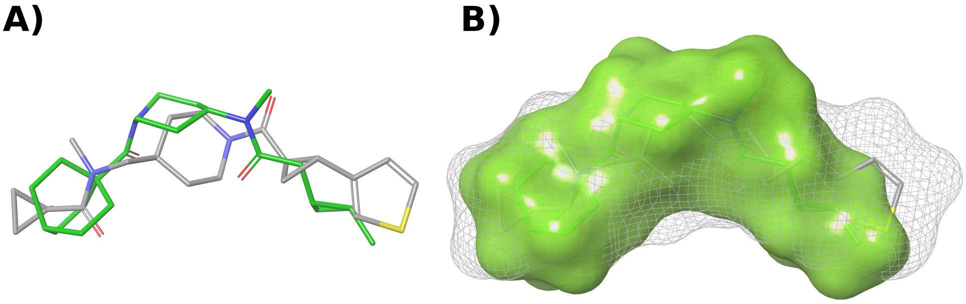


Figure 5.2: Shape overlap between ZINC000954430177 (query, green, solid surface) and ZINC001497961236 (grey, mesh surface) which was ranked top 8 by the model and rank 3688 by the baseline.

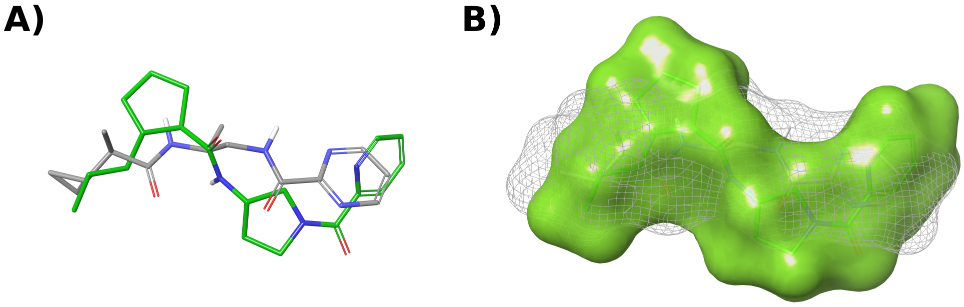


Figure 5.3: Shape overlap between ZINC000570771518 (query, green, solid surface) and ZINC001386435162 (grey, mesh surface) which was ranked top 14 by the model and rank 1585 by the baseline.

that the predicted similar compound has a scaffold different from the query. Instinctively, one may say that this is actually a good hit, but nevertheless the model did not reproduce the correct rank as calculated by the baseline. Comparing the pharmacophore-based shape similarity (0.41) with the shape-only similarity (0.63) shows again that the discrepancy of the ranks was caused by the model's inability to capture the pharmacophore information while nicely reproducing the 3D shape overlap.

There are, however, also instances where the model very nicely reproduced hits, even if the 2D structure was very different from the query. One such example is shown in Figure 5.4. In

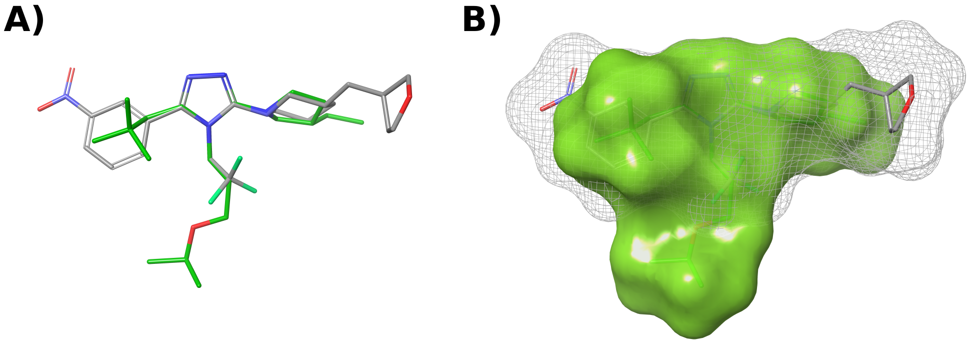


Figure 5.4: Shape overlap between ZINC001763434742 (query, green, solid surface) and ZINC001556038399 (grey, mesh surface) which was ranked top 34 by the model and rank 76 by the baseline.

this case, the shape overlap appears to be worse than in the previous examples, even though this compound was highly rated by the baseline. The shape similarities with and without pharmacophore information were very similar at 0.58 and 0.59, respectively. This indicates that the score was mainly influenced by the 3D shape and that the pharmacophore information did not contribute much. Since our model performed very well on this example, it again suggests that the model is good at finding compounds with a similar shape, but not as good at capturing pharmacophore information.

In these experiments, we could show that the model is able to produce hits with a high overlap with the query molecules in 3D space. However, there may still be some shortcomings in certain cases in reproducing the exact similarity metric, especially if the similarity goes beyond "simple" 3D overlap. Nevertheless, the model proved to be able to capture 3D similarities independent of the 2D structure of the molecules.

5.3.2 Real-World Examples

To simulate a real-world example of a possible screening study, we selected 2 co-crystallized ligands as queries to screen the drugs contained in the Drugbank[45]. In a first trial, we selected raloxifene (co-crystallized to the estrogen receptor (ER) in PDB ID 1ERR). This compound has not been seen by the model before. The precision of the top 100 predictions was 53%, which is much higher than for most of the examples in Subsection 5.3.1. However, the precision of the top 1000 predictions was slightly lower at 49%. The PCC of all predicted distances with calculated similarities was -0.77. Interestingly, the model was able to reproduce the top 10 most similar compounds according to the baseline within the top 47 ranked hits. Since we wanted to know if the model can be used to find other compounds that modulate the ER, we investigated the top 10 predictions. Table 5.3 shows the results of the analysis. All 10 predicted most similar compounds have literature confirmation of ER modulating activity.

Table 5.3: Predited top 10 most similar compounds to raloxifene. The first hit is omitted from the table because it is raloxifene itself.

Drugbank ID	Predicted rank	Name	ER activity
DB05414	2	Pipendoxifene	Yes[46]
DB06401	3	Bazedoxifene	Yes[47]
DB06249	4	Arzoxifene	Yes[48]
DB16080	5	Acolbifene	Yes[49]
DB03742	6	Compound 4-D	Yes[50]
DB07352	7	Apigenin	Yes[51]
DB01645	8	Genistein	Yes[52,53]
DB15464	9	Urolithin A	Yes[54]
DB13182	10	Daidzein	Yes[55]

While this is a very promising result, one also needs to keep in mind that the Drugbank is a biased database in that it contains not only mostly drug-like molecules, but also many known ER modulators.

Figure 5.5 shows the 2D structures of raloxifene and the 9 most similar compounds predicted by the model. It also contains the 3D structures of selected top-ranked compounds aligned with raloxifene. We further calculated the 2D similarity between the reference and the hits. While the 5 top ranked compounds were structurally quite similar (2D similarity between 0.44 and 0.66), the next 4 hits had scaffolds very different from raloxifene. They consisted of 3 isoflavonoids and 1 benzo-coumarin. The 3D alignment shows that even the compounds with different scaffolds share similar features, which could reproduce the binding mode of raloxifene. This analysis clearly shows that the model is able to find structurally dissimilar compounds by learning 3D features.

Next, we wanted to test a compound with a different chemical scaffold than raloxifene. We decided to screen tetrahydrogestrinone against the drugs contained in the Drugbank. Like before, this compound has not been seen by the model before. The 10 highest ranked compounds predicted by our model are shown in Table 5.4. The precision of the top 100 predictions was 72%, which is exceptionally high. We believe that this is because steroidal

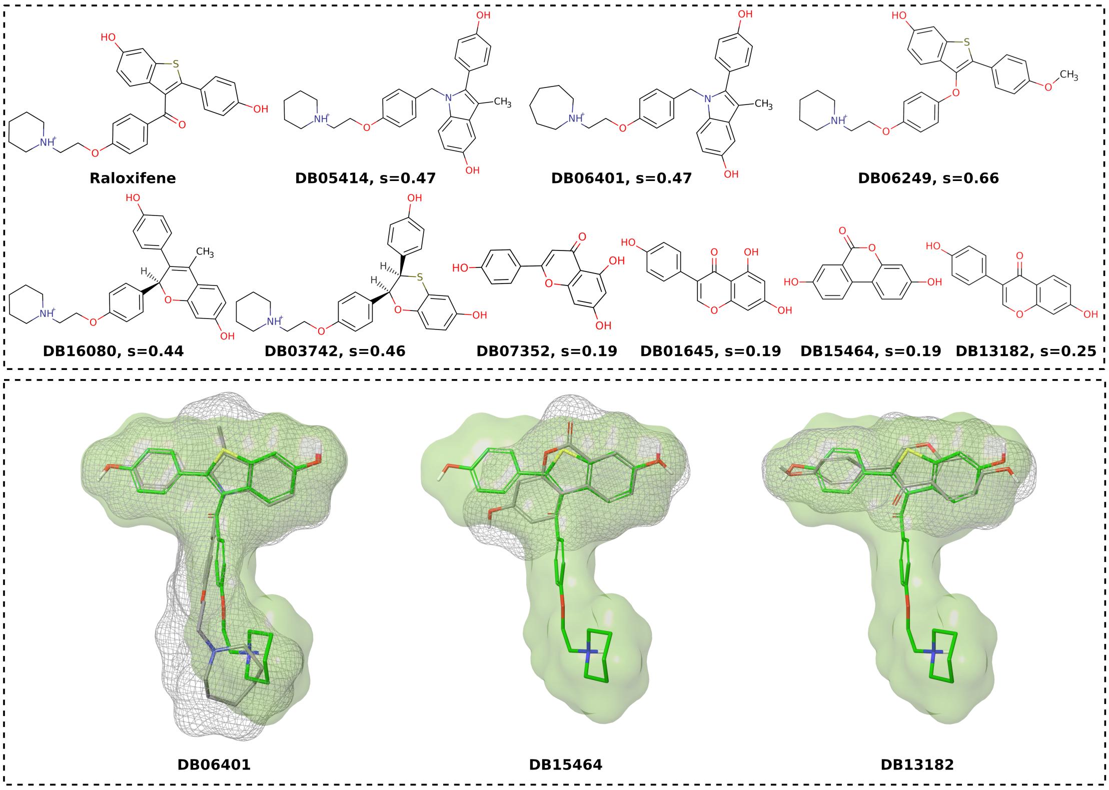


Figure 5.5: 2D and 3D structures of the 9 predicted most similar compounds to raloxifene. Top: 2D structures with 2D similarity s. Bottom: 3D alignment of raloxifene (green stick representation with green solid surface) with 3 of the top ranked compounds with diverse scaffolds (gray stick representation with mesh surface).

Table 5.4: Predited top 10 most similar compounds to tetrahydrogestrinone. The first hit is omitted from the table because it is tetrahydrogestrinone itself.

Drugbank ID	Predicted rank	Name	AR activity
DB11619	2	Gestrinone	Yes[56]
DB11372	3	Altrenogest	Yes[57]
DB02998	4	Metribolone	Yes[58]
DB13563	5	Norgestrienone	Yes[59]
DB11551	6	Trenbolone	Yes[60]
DB06730	7	Gestodene	Yes[61]
DB09389	8	Norgestrel	Yes[62]
DB13602	9	Promegestone	No[63]
DB00367	10	Levonorgestrel	Yes[64]

compounds have a very unique shape and therefore it may be easier for the model to find similar compounds. The precision of the first 1000 hits was equally high with 69% and the correlation between all predicted distances and the calculated shape similarities was -0.82.

The model was able to reproduce the top 10 hits from shape screening within the top 27 predictions. Eight of the 9 hits listed in Table 5.4 have been shown in the literature to have androgenic or anti-androgenic activity. Nevertheless, we still acknowledge the fact that the Drugbank may be a biased database in that molecules with androgenic or anti-androgenic properties are overrepresented compared to other databases.

Since steroidal compounds have such a distinct chemical structure (and shape), we wanted to see if the model can also find non-steroidal compounds as easily as shape screening. This would show that the model indeed learns from the provided 3D information instead of relying on 2D similarity. The first non-steroidal compound we identified in the baseline hits was borneol at rank 123. Despite the fact that borneol has a very different 2D structure than steroids, the model predicted this compound to be at rank 95. This underlines the model's ability to capture 3D similarities in latent space.

Based on these examples, we believe that our model is indeed suitable for use in real-world virtual screening applications. Although it is not able to perfectly reproduce the similarities found in the chosen baseline, its predictions are reasonable and useful in finding compounds with similar 3D features.

5.3.3 Investigation of Computational Cost

To assess whether our method actually allows screening (ultra) large databases at reduced computational cost, we screened 27 query compounds against databases of different sizes. We chose to screen 27 query compounds to be able to directly compare the results with those of Michino et al.[14] (as introduced in Section 5.2.1). We employed faiss, a tool created by Meta, to scan the databases[65]. Faiss allows to create searchable indexes from vectors. Several indexes with varying levels of accuracy and speed are available. We chose IndexFlatL2, an index that computes the exact squared L2 norm between the queries and all elements in the database.

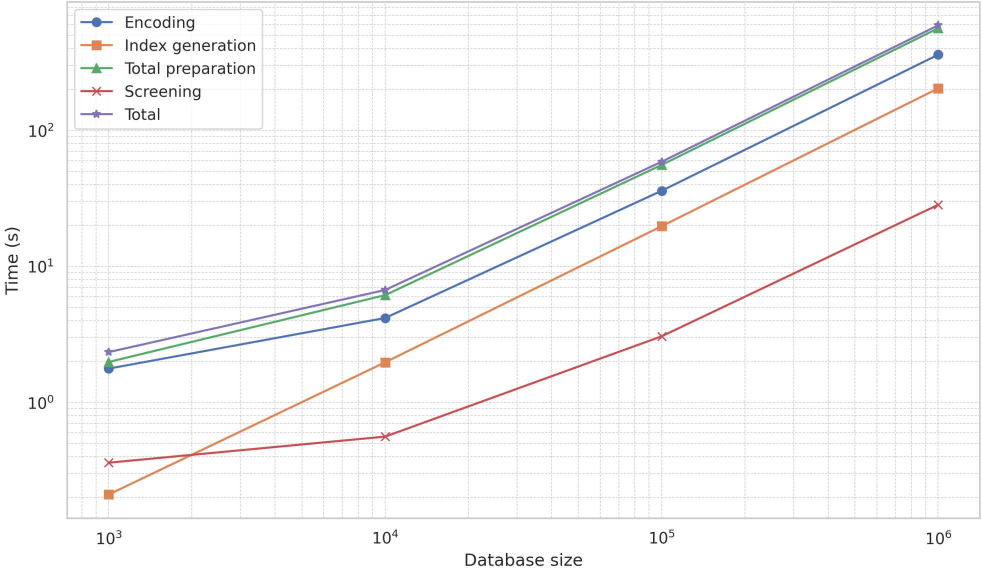


Figure 5.6: Times in seconds to screen 27 query molecules against databases of different sizes. Displayed are the time to encode the database into latent space (blue), the time to generate the IndexFlatL2 index in faiss (orange), the total preparation time (encoding plus index generation; green), the time to actually screen the database (red), and the total time (purple).

This index is the most accurate, but also the slowest. Thus, it would be possible to further increase the performance of the screening by choosing other more approximating indexes.

We tested databases containing 1k, 10k, 100k, and 1M compounds for screening. A screening consists of 3 steps: encoding into latent space, generation of the index, and screening of the index. Encoding into latent space is the most time-consuming task followed by the index generation (cf. Figure 5.6). However, these two tasks only need to be completed once, and the encoded data and index created can be reused for any future screenings. Figure 5.6 shows that the screening times required increase linearly with the size of the database. This allows for easy extrapolation to larger databases. Therefore, screening 27 queries against 1 billion conformers would require roughly 7.8 hours on a single CPU core. Michino et al. screened approximately 1.12 billion compounds with 10 conformers each (i.e. 11.2 billion conformers) in 19.5 hours using 216 GPUs. Extrapolating our tests to 11.2 billion conformers would result in approximately 87 hours or 3.6 days on a single CPU core. When parallelized to 8 cores (which is very reasonable for standard consumer-grade computers), the screening could be completed within 10.9 hours. If more powerful hardware is available, for example a CPU with 64 cores, the screening could be performed in only 82 minutes. We deliberately ran all experiments on a regular desktop computer with one RTX 2080 Super GPU. The GPU was used only to encode the conformers into latent space. Thus, we show that this technology enables the screening of ultra-large databases without the need for expensive hardware.

Another advantage of this method is that it allows parallel search of multiple queries. It took only an additional 2.35 seconds to screen 100 queries instead of 27 against a database containing 1 million conformers. Thus, the gain in speed over the classical shape screening increases with the number of query compounds.

5.4 Methods

5.4.1 Dataset Preparation

Like in our proof-of-concept study, we used a randomly selected subset from the ZINC database containing around 500k compounds to train our model[30]. From this dataset, we selected the 25k molecules that best cover the chemical space of the complete dataset using the Kennard-Stone algorithm[66]. From this narrowed-down subset, we again applied the Kennard-Stone algorithm to isolate the most diverse 5k compounds for use as our validation set. The remaining 20k compounds were used as the initial training set. The initially unused 475k compounds were split into an external test set, consisting of 10k compounds, and a pool of molecules that were used as an unlabeled set to sample from during the active learning cycles.

5.4.2 Data Handling

In this work, the model was trained to encode 3D molecules into latent space and decode them to SELFIES[67,68]. In order to enable the model to learn from 3D structures, the first step was to calculate atom features to be used as nodes. This was done using RDKit. For each atom in a molecule, we encoded the atom type, the atom degree (i.e. the number of neighbors), the number of connected hydrogen atoms, the implicit valence, the hybridization, and the aromaticity in a vector that could later be used in a learnable embedding.

The second step in encoding 3D information was to create edges between nodes in a way that conserved the 3D topology. This was done by passing the Euclidean distance matrix of a molecule through an exponential decay function and combining the result with the adjacency matrix. This is described in Equation 5.1.

$$E = \max(A, \exp(-D)) \tag{5.1}$$

Where A is the adjacency- and D the Euclidean distance matrix of a molecule. This approach allows 3D information to be encoded in a translation and rotation invariant way while also preserving information about atom connectivity.

Given the vast amount of data and the expense of training, we chose to train the model in this initial 3D-enabled version with only one conformation per molecule. However, we think that including multiple conformations could enhance the model's capacity to learn 3D similarities.

The SELFIES used in this work were converted from canonical SMILES which were created using Openbabel version 3.0.0[69]. Each SELFIES that was passed through the model was tokenized based on its individual components. We decided to use SELFIES instead of

SMILES strings due to their robustness.

As our baseline similarity method, we used Schrödinger's pharmacophore-based shape similarity shipped with their 2023-2 release[44]. Unless otherwise stated, all molecules were processed with LigPrep prior to shape screening. We chose to generate protonation states at pH 7.4 and for each molecule, we created one 3D conformation using the OPLS4 force field.

5.4.3 Model Architecture

The architecture of the model had to be only minimally adapted from our proof-of-concept study. Since the 3D- and connectivity information of the molecules are fully encoded by their edges, no positional encoding is needed in the encoder. In fact, removing the positional encoding is required for permutation equivariance because the order of the nodes does not matter, and thus a positional encoding would give incorrect information. The rest of the model is still based on the original implementation of the Transformer model by Vaswani et al.[70]. In the previous study, a masked mean was used to combine the nodes to generate a latent vector, whereas this work utilizes a weighted sum pooling technique. In this method, the weights are calculated using two linear layers with a tanh activation between them.

To train the model, we used the same combination of the reconstruction (i.e. cross-entropy) loss and our custom similarity loss as in the proof-of-concept study. Equation 5.2 shows this loss function in detail where A is an anchor sample, X is some other sample in the mini batch, $sim(\cdot,\cdot)$ is a similarity function (in this case Schrödinger's pharmacophore-based shape similarity), and $f(\cdot)$ is the encoder of the model, encoding a molecule into latent space. To decrease the density of the latent space, the scaling factor a is used. This scaling factor was set to 10 in this work.

$$L(A,X) = \left| a \cdot \|(1 - sim(A,X))\| - \|f(A) - f(X)\| \right| \tag{5.2}$$

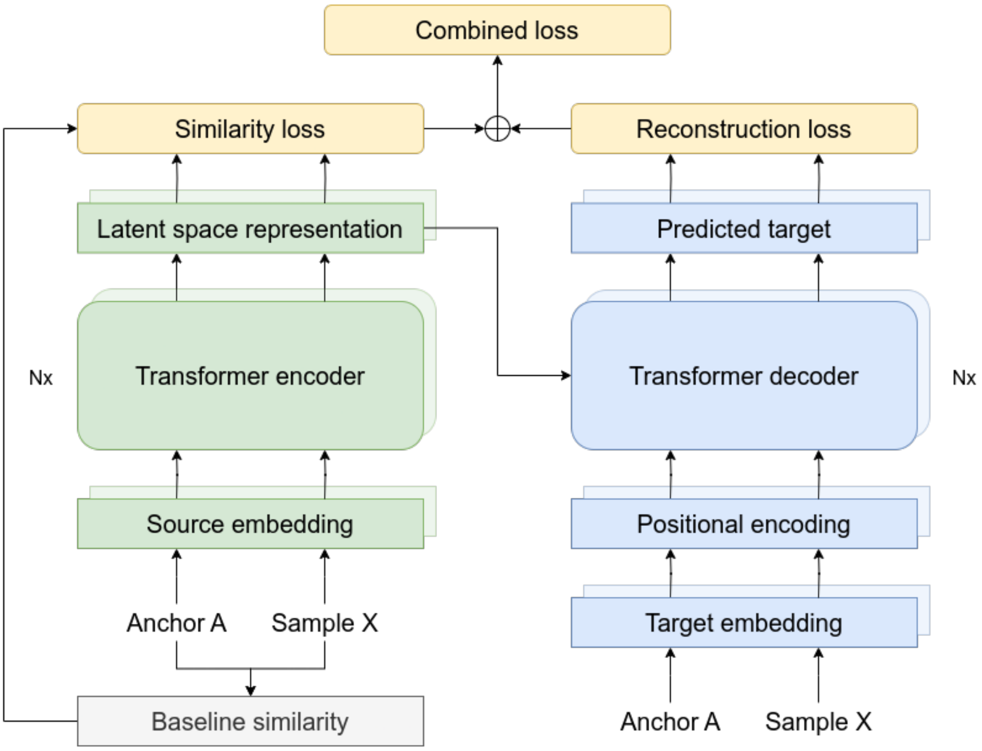


Figure 5.7: Architecture of the model used in this study. This adaptation of the original Transformer implementation enables the encoding of molecular three-dimensional information that is invariant to translations and rotations. We used Schrödinger's pharmacophore-based shape similarity as baseline.

Like in our proof-of-concept study, we used scaled dot-product attention as described in Equation 5.3, where Q, K, and V are tensors containing the queries, keys, and values, and d_k is the dimensionality of the keys[70].

$$attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{5.3}$$

The complete model architecture is depicted in Figure 5.7. The model was trained with a batch size of 64, a latent space dimensionality of 252, a learning rate of $1e^{-4}$, and 4 encoder and decoder layers each. Each attention module consisted of 4 heads and the model was trained for a minimum of 200 and a maximum of 800 epochs per active learning cycle.

Because we used an active learning approach, the initial training set was much smaller compared to the training set used in the proof-of-concept study. However, from the two implemented loss terms, only the similarity loss depends on labeled data, whereas the reconstruction term is completely unsupervised. Therefore, for each mini batch, we created an additional mini batch containing random samples from the unlabeled dataset. This mini batch was used to train only the reconstruction. We chose this approach to improve the reconstruction abilities of the model while preventing overfitting on the small datasets.

Since calculating the 3D similarities is comparatively expensive, we precomputed pairwise similarities for our initial training and validation set. During training, we used a similarity sampler to ensure that each molecule in a mini batch contained at least one similar compound. This was accomplished by randomly sampling 3 compounds from the 100 most similar to a given anchor molecule and adding them to the mini batch. This step is imperative for the model to conserve similarities in latent space and has already been described in our proof-of-concept study.

5.4.4 Active Learning

In this work, we used an active learning approach using a combination of QBC and EMC acquisition functions. Calculating the EMC involves calculating the gradient of the loss with respect to the input. Thus, one needs to be able to calculate the loss of a sample. Since the point of active learning is to select samples from unlabeled data to be labeled, this approach cannot be applied to the similarity loss. However, it is possible to use EMC for the reconstruction loss. For each active learning cycle, we encoded and decoded all unlabeled samples and calculated the gradient of the reconstruction loss with respect to the input. The gradients were then normalized using the L2 norm and squared to ensure positivity. This is described in Equation 5.4 where $\nabla_x L$ is the gradient of the reconstruction loss L with respect to the input sample x.

$$\text{EMC} = \|\nabla_x L\|^2 \tag{5.4}$$

The EMC was mixed with QBC in order to also sample based on the similarity loss. To achieve this, we predicted the latent vector for each unlabeled sample 100 times using a dropout rate of 10%. We then calculated the mean over the variance of the predictions. This is shown in Equation 5.5 where P is a matrix containing N predictions and $\text{Var}(\cdot)$ is the variance.

$$\text{QBC} = \frac{1}{N} \sum_{i=1}^{N} \text{Var}(P_i) \tag{5.5}$$

The sampling of the unlabeled data was performed on the basis of the magnitude of the EMC and QBC values. Starting from the samples with the highest values, an equal number of samples were drawn based on the EMC and QBC values. This was done until a total of 5k unique molecules were sampled from the unlabeled dataset. These samples were then labeled by calculating pairwise similarities (including the existing labeled samples) and added to the training set. Thus, for each active learning cycle, the training set grew by 5k compounds. Because we calculated pairwise similarities, the time used to label the newly sampled compounds increased exponentially. The model in this work was trained for 5 active learning cycles, resulting in a final training set containing 45k compounds.

5.5 Conclusion

We previously demonstrated that a distance-aware transformer model can be used to preserve 2D similarities in latent space. We claimed that this method can be used independent of the underlying similarity metric, allowing to efficiently estimate highly complex 3D similarities.

In this work, we show how a slightly adapted model is capable of capturing such 3D similarities. We demonstrated that our model, which uses a translation and rotation invariant molecular representation, is able to recognize 3D features of molecules and identify molecules with similar shapes in a virtual screening context. Although the model cannot perfectly reproduce the underlying pharmacophore-based shape similarity, it is still capable of enriching the top hits with highly similar compounds. In fact, we believe that using a shape-only similarity metric would lead to much better performance because the model does not seem to be able to fully capture the pharmacophore information. Thus, for such special similarity metrics, the model might need to be further adapted to better reproduce the baseline similarity.

The approach described herein enables the use of the Euclidean distance in latent space as an approximation of computationally expensive 3D similarity metrics. It therefore allows researchers to run quick and efficient (pre)screenings on ultra-large databases using regular low-cost computer hardware.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of two RTX A5000 GPUs used for this research.

5.6 Appendix

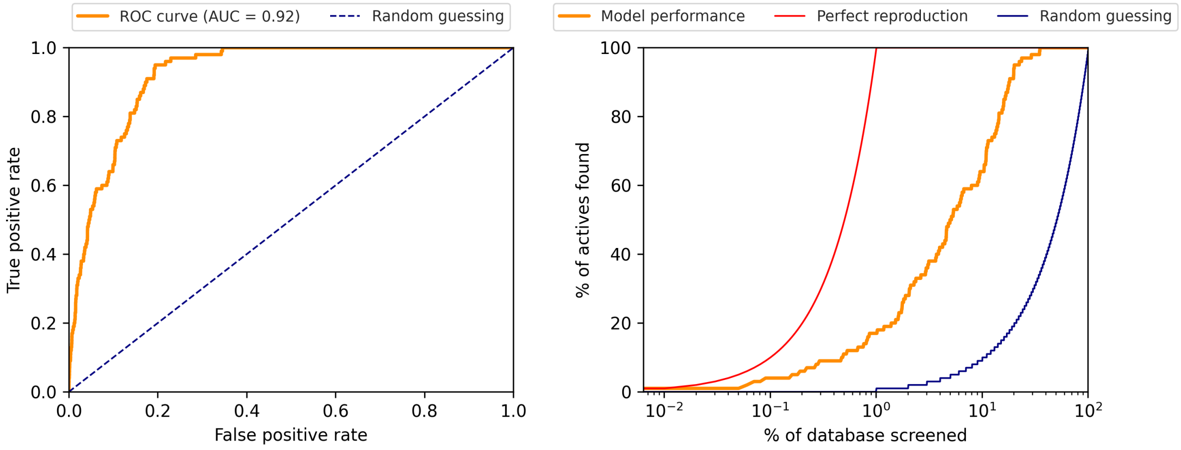


Figure A5.1: Screening performance for query ZINC000570771518. Left: ROC curve, right: reproduction performance.

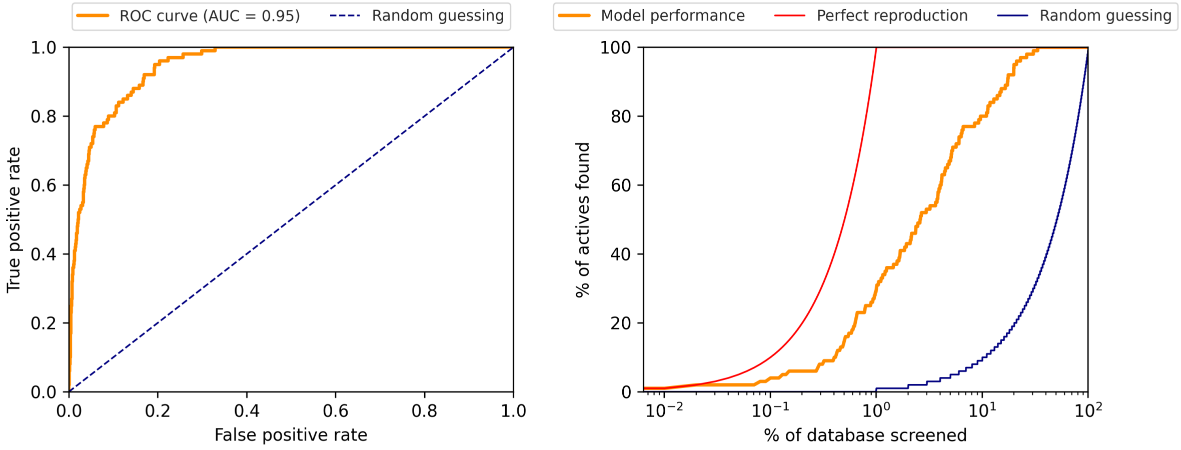


Figure A5.2: Screening performance for query ZINC000950159323. Left: ROC curve, right: reproduction performance.

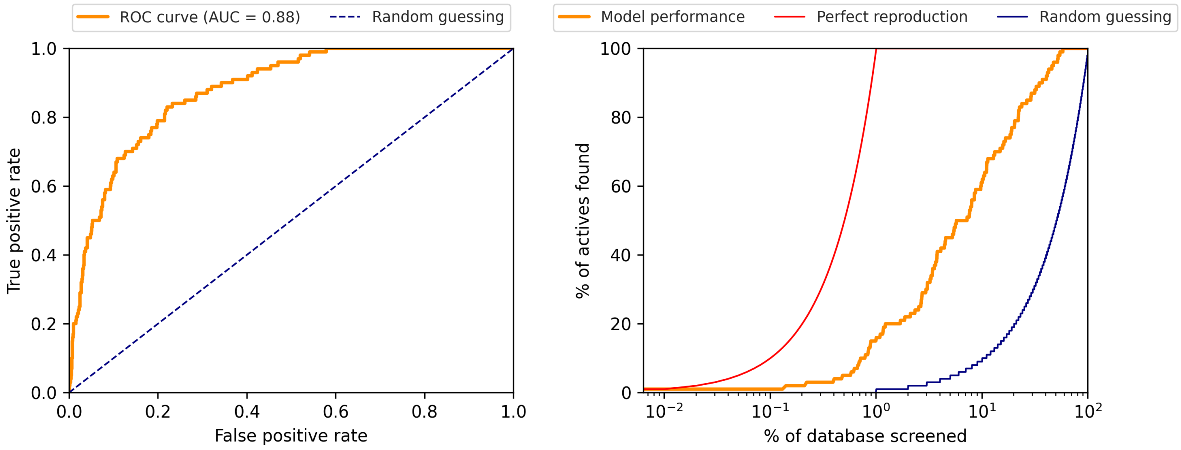


Figure A5.3: Screening performance for query ZINC000954430177. Left: ROC curve, right: reproduction performance.

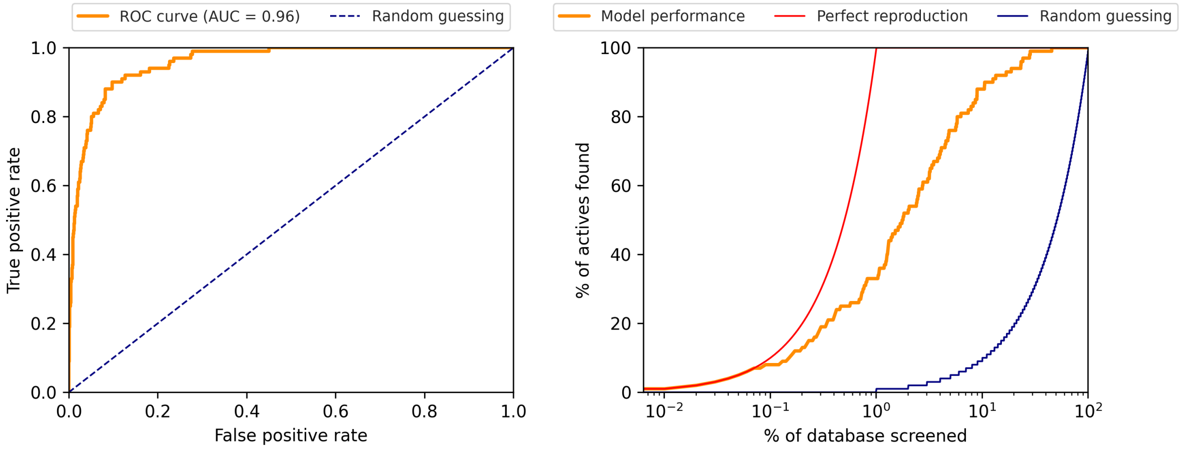


Figure A5.4: Screening performance for query ZINC000970035445. Left: ROC curve, right: reproduction performance.

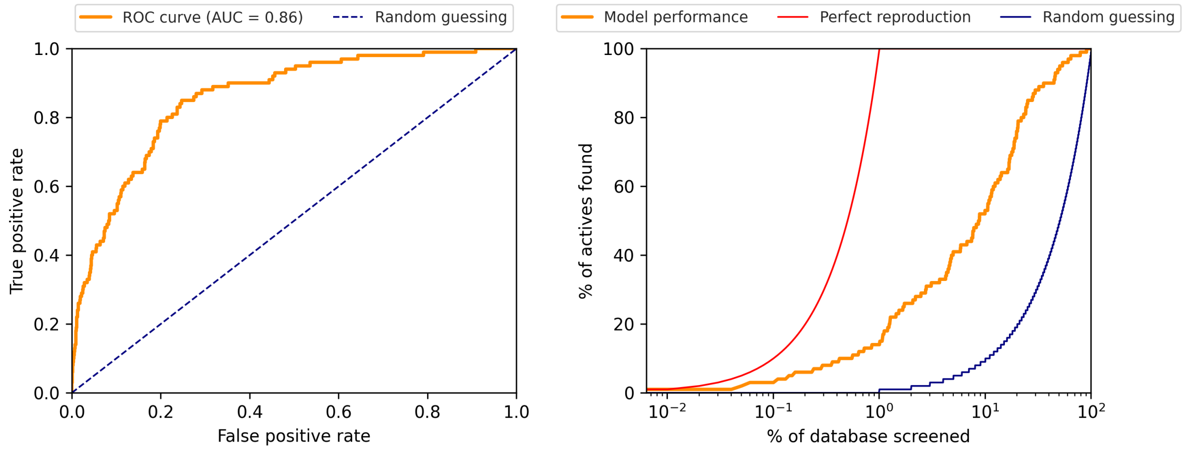


Figure A5.5: Screening performance for query ZINC001183157671. Left: ROC curve, right: reproduction performance.

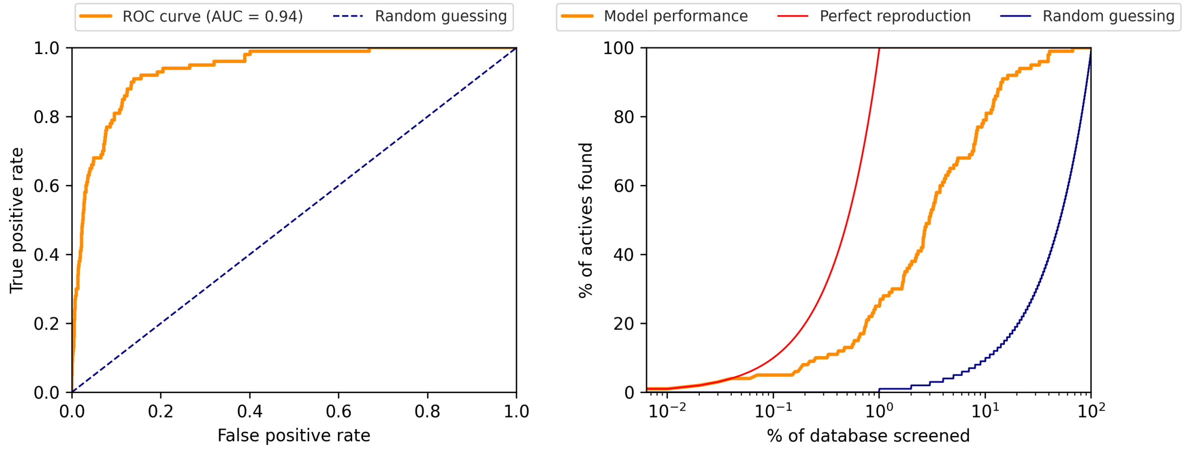


Figure A5.6: Screening performance for query ZINC001281147597. Left: ROC curve, right: reproduction performance.

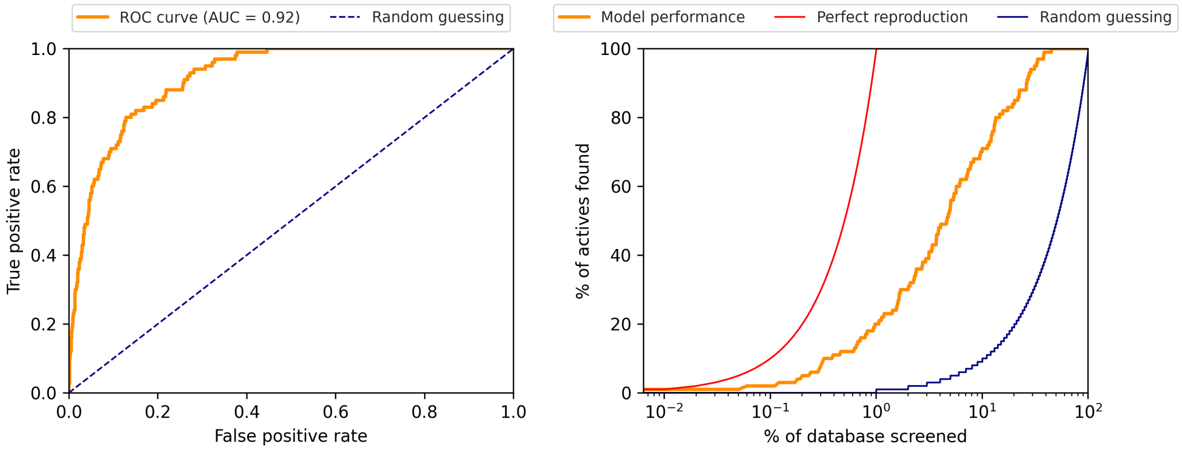


Figure A5.7: Screening performance for query ZINC001368797027. Left: ROC curve, right: reproduction performance.

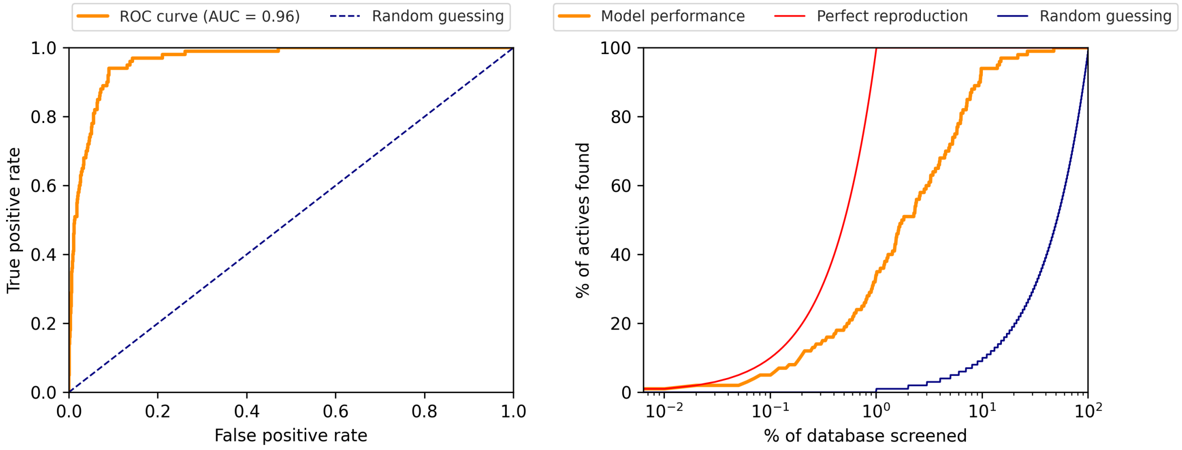


Figure A5.8: Screening performance for query ZINC001711902206. Left: ROC curve, right: reproduction performance.

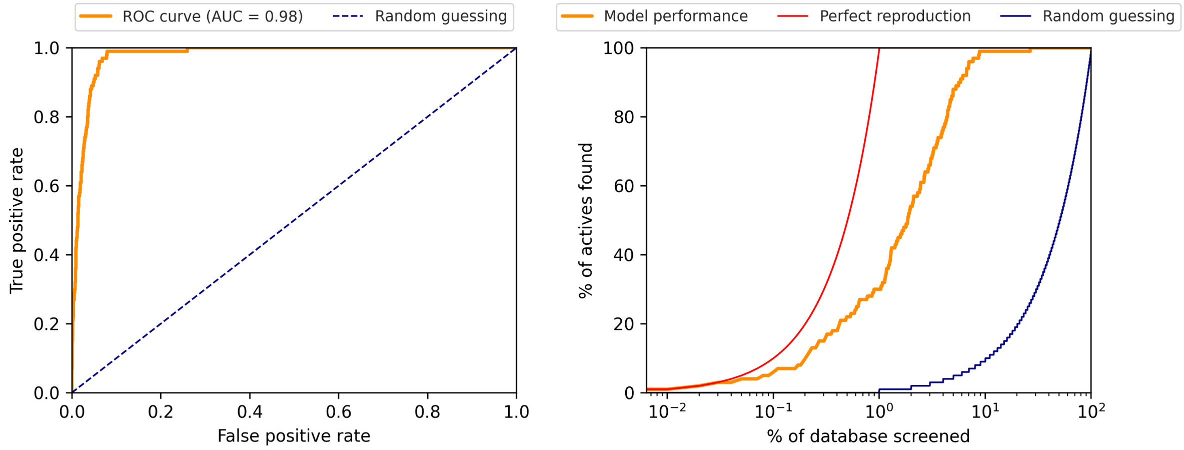


Figure A5.9: Screening performance for query ZINC001740566933. Left: ROC curve, right: reproduction performance.

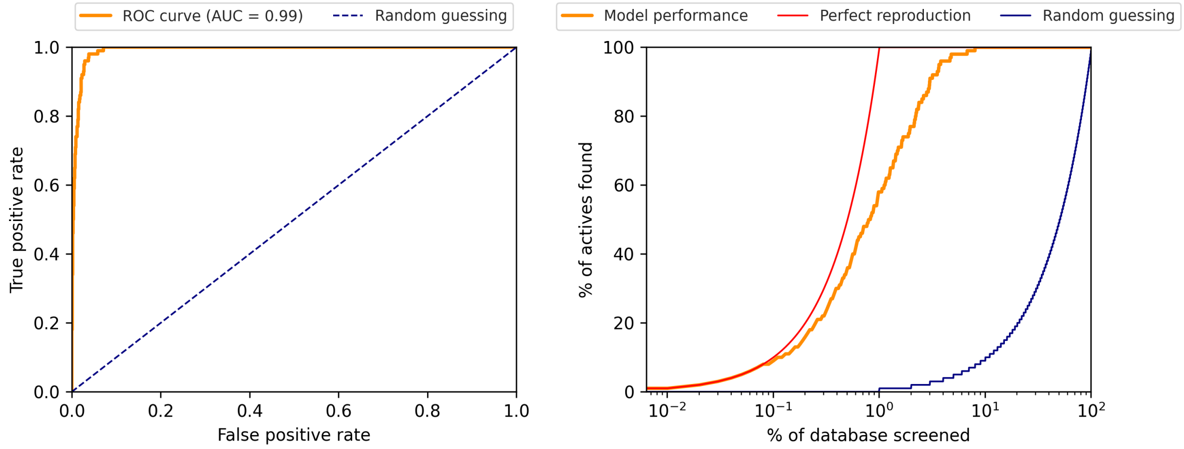


Figure A5.10: Screening performance for query ZINC001763434742. Left: ROC curve, right: reproduction performance.

References

[1] Manuel S. Sellner, Amr H. Mahmoud, and Markus A. Lill. Enhancing ligand-based virtual screening with 3d shape similarity via a distance-aware transformer model. *bioRxiv*, page 2023.11.17.567506, 11 2023.

[2] Andreas Luttens, Hjalmar Gullberg, Eldar Abdurakhmanov, Duy Duc Vo, Dario Akaberi, Vladimir O. Talibov, Natalia Nekhotiaeva, Laura Vangeel, Steven De Jonghe, Dirk Jochmans, Janina Krambrich, Ali Tas, Bo Lundgren, Ylva Gravenfors, Alexander J. Craig, Yoseph Atilaw, Anja Sandström, Lindon W. K. Moodie, Ake Lundkvist, Martijn J. van Hemert, Johan Neyts, Johan Lennerstrand, Jan Kihlberg, Kristian Sandberg, U. Helena Danielson, and Jens Carlsson. Ultralarge Virtual Screening Identifies SARS-CoV-2 Main Protease Inhibitors with Broad-Spectrum Activity against Coronaviruses. *Journal of the American Chemical Society*, 144(7):2905–2920, 2 2022.

[3] Pedro Andrade Bonilla, Cody L. Hoop, Karen Stefanisko, Sergey G. Tarasov, Sourav Sinha, Marc C. Nicklaus, and Nadya I. Tarasova. Virtual screening of ultra-large chemical libraries identifies cell-permeable small-molecule inhibitors of a "non-druggable" target, STAT3 N-terminal domain. *Frontiers in Oncology*, 13:1144153, 4 2023.

[4] Christoph Gorgulla, Krishna M. Padmanabha Das, Kendra E. Leigh, Marco Cespugli, Patrick D. Fischer, Zi-Fu Wang, Guilhem Tesseyre, Shreya Pandita, Alec Shnapir, Anthony Calderaio, Minko Gechev, Alexander Rose, Noam Lewis, Colin Hutcheson, Erez Yaffe, Roni Luxenburg, Henry D. Herce, Vedat Durmaz, Thanos D. Halazonetis, Konstantin Fackeldey, J.J. Patten, Alexander Chuprina, Igor Dziuba, Alla Plekhova, Yurii Moroz, Dmytro Radchenko, Olga Tarkhanova, Irina Yavnyuk, Christian Gruber, Ryan Yust, Dave Payne, Anders M. Näär, Mark N. Namchuk, Robert A.

Davey, Gerhard Wagner, Jamie Kinney, and Haribabu Arthanari. A multi-pronged approach targeting SARS-CoV-2 proteins using ultra-large virtual screening. *iScience*, 24(2):102021, 2 2021.

[5] Christoph Gorgulla, Andras Boeszoermenyi, Zi-Fu Wang, Patrick D. Fischer, Paul W. Coote, Krishna M. Padmanabha Das, Yehor S. Malets, Dmytro S. Radchenko, Yurii S. Moroz, David A. Scott, Konstantin Fackeldey, Moritz Hoffmann, Iryna Iavniuk, Gerhard Wagner, and Haribabu Arthanari. An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668, 4 2020.

[6] Timothy S. Chisholm, Mark Mackey, and Christopher A. Hunter. Discovery of High-Affinity Amyloid Ligands Using a Ligand-Based Virtual Screening Pipeline. *Journal of the American Chemical Society*, 145(29):15936–15950, 7 2023.

[7] Di Zhu, Sandra Johannsen, Tiziana Masini, Céline Simonin, Jörg Haupenthal, Boris Illarionov, Anastasia Andreas, Mahendra Awale, Robin M. Gierse, Tridia van der Laan, Ramon van der Vlag, Rita Nasti, Mael Poizat, Eric Buhler, Norbert Reiling, Rolf Müller, Markus Fischer, Jean-Louis Reymond, and Anna K. H. Hirsch. Discovery of novel drug-like antitubercular hits targeting the MEP pathway enzyme DXPS by strategic application of ligand-based virtual screening. *Chemical Science*, 13(36):10686–10698, 9 2022.

[8] Shizhen Zhao, Xinping Li, Wenjing Peng, Le Wang, Wenling Ye, Yang Zhao, Wenbo Yin, Wei-Dong Chen, Weiguo Li, and Yan-Dong Wang. Ligand-based pharmacophore modeling, virtual screening and biological evaluation to identify novel TGR5 agonists. *RSC Advances*, 11(16):9403–9409, 3 2021.

[9] Domingo Méndez-Álvarez, Maria F. Torres-Rojas, Edgar E. Lara-Ramirez, Laurence A.

Marchat, and Gildardo Rivera. Ligand-Based Virtual Screening, Molecular Docking, and Molecular Dynamic Simulations of New β-Estrogen Receptor Activators with Potential for Pharmacological Obesity Treatment. *Molecules*, 28(11):4389, 5 2023.

[10] Ingo Muegge and Prasenjit Mukherjee. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opinion on Drug Discovery*, 11(2):137–148, 2 2016.

[11] Ashutosh Kumar and Kam Y. J. Zhang. Advances in the Development of Shape Similarity Methods and Their Application in Drug Discovery. *Frontiers in Chemistry*, 6(JUL):383861, 7 2018.

[12] Matthew P. Seddon, David A. Cosgrove, Martin J. Packer, and Valerie J. Gillet. Alignment-Free Molecular Shape Comparison Using Spectral Geometry: The Framework. *Journal of Chemical Information and Modeling*, 59(1):98–116, 1 2019.

[13] G. Madhavi Sastry, Steven L. Dixon, and Woody Sherman. Rapid Shape-Based Ligand Alignment and Virtual Screening Method Based on Atom/Feature-Pair Similarities and Volume Overlap Scoring. *Journal of Chemical Information and Modeling*, 51(10):2455–2466, 10 2011.

[14] Mayako Michino, Alexandre Beautrait, Nicholas A. Boyles, Aparna Nadupalli, Alexey Dementiev, Shan Sun, John Ginn, Leigh Baxt, Robert Suto, Ruslana Bryk, Steven V. Jerome, David J. Huggins, and Jeremie Vendome. Shape-Based Virtual Screening of a Billion-Compound Library Identifies Mycobacterial Lipoamide Dehydrogenase Inhibitors. *ACS Bio & Med Chem Au*, 9 2023.

[15] Dagmar Stumpfe and Jürgen Bajorath. Similarity searching. *WIREs Computational Molecular Science*, 1(2):260–282, 3 2011.

[16] Oliver Laufkötter, Tomoyuki Miyao, and Jürgen Bajorath. Large-Scale Comparison of Alternative Similarity Search Strategies with Varying Chemical Information Contents. *ACS Omega*, 4(12):15304–15311, 9 2019.

[17] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 8 2013.

[18] Kourosh T. Baghaei, Amirreza Payandeh, Pooya Fayyazsanavi, Shahram Rahimi, Zhiqian Chen, and Somayeh Bakhtiari Ramezani. Deep representation learning: Fundamentals, Perspectives, Applications, and Open Challenges. *ArXiv*, 11 2022.

[19] Xiaodong Liu, Jianfeng Gao, Xiaodong He, Li Deng, Kevin Duh, and Ye-Yi Wang. Representation Learning Using Multi-Task Deep Neural Networks for Semantic Classification and Information Retrieval. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 912–921, Stroudsburg, PA, USA, 2015. Association for Computational Linguistics.

[20] Qianwen Meng, Hangwei Qian, Yong Liu, Yonghui Xu, Zhiqi Shen, and Lizhen Cui. Unsupervised Representation Learning for Time Series: A Review. *ArXiv*, 8 2023.

[21] Jinchuan Qian, Zhihuan Song, Yuan Yao, Zheren Zhu, and Xinmin Zhang. A review on autoencoder based representation learning for fault detection and diagnosis in industrial processes. *Chemometrics and Intelligent Laboratory Systems*, 231:104711, 12 2022.

[22] Ahmad Suhaimi, Amos W. H. Lim, Xin Wei Chia, Chunyue Li, and Hiroshi Makino.

Representation learning in the artificial and biological neural networks underlying sensorimotor integration. *Science Advances*, 8(22):984, 6 2022.

[23] Xiaomin Fang, Lihang Liu, Jieqiong Lei, Donglong He, Shanzhuo Zhang, Jingbo Zhou, Fan Wang, Hua Wu, and Haifeng Wang. Geometry-enhanced molecular representation learning for property prediction. *Nature Machine Intelligence*, 4(2):127–134, 2 2022.

[24] Phuc H. Le-Khac, Graham Healy, and Alan F. Smeaton. Contrastive Representation Learning: A Framework and Review. *IEEE Access*, 8:193907–193934, 10 2020.

[25] S. Chopra, R. Hadsell, and Y. LeCun. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[26] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 07-12-June-2015, pages 815–823. IEEE, 6 2015.

[27] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep Metric Learning via Lifted Structured Feature Embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2016-December, pages 4004–4012. IEEE, 6 2016.

[28] Kihyuk Sohn. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. *Advances in Neural Information Processing Systems*, 29, 2016.

[29] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Yee Whye Teh and Mike Titterington, editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 297–304, Chia Laguna Resort, Sardinia, Italy, 13–15 May 2010. PMLR.

[30] Manuel S. Sellner, Amr H. Mahmoud, and Markus A. Lill. Efficient virtual high-content screening using a distance-aware transformer model. *Journal of Cheminformatics*, 15(1):18, 2 2023.

[31] Samuel Budd, Emma C. Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71:102062, 7 2021.

[32] Sareer Ul Amin, Adnan Hussain, Bumsoo Kim, and Sanghyun Seo. Deep learning based active learning technique for data annotation and improve the overall performance of classification models. *Expert Systems with Applications*, 228:120391, 10 2023.

[33] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294, New York, NY, USA, 7 1992. ACM.

[34] Yoav Freund, H. Sebastian Seung, Eli Shamir, and Naftali Tishby. Selective Sampling Using the Query by Committee Algorithm. *Machine Learning*, 28(2-3):133–168, 1997.

[35] Hideitsu Hino and Shinto Eguchi. Active learning by query by committee with robust divergences. *Information Geometry*, 6(1):81–106, 6 2023.

[36] Melanie Ducoffe and Frederic Precioso. QBDC: Query by dropout committee for training deep supervised architecture. 11 2015.

[37] Wenbin Cai, Ya Zhang, and Jun Zhou. Maximizing Expected Model Change for Active Learning in Regression. In *2013 IEEE 13th International Conference on Data Mining*, pages 51–60. IEEE, 12 2013.

[38] Sung Ho Park and Seoung Bum Kim. Robust expected model change for active learning in regression. *Applied Intelligence*, 50(2):296–313, 2 2020.

[39] John J. Irwin, Khanh G. Tang, Jennifer Young, Chinzorig Dandarchuluun, Benjamin R. Wong, Munkhzul Khurelbaatar, Yurii S. Moroz, John Mayfield, and Roger A. Sayle. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073, 12 2020.

[40] rdkit/rdkit: 2023_09_1 (Q3 2023) Release Beta. https://zenodo.org/records/8413907.

[41] Schrödinger Release 2021-1. LigPrep, 2021.

[42] Schrödinger Release 2023-2. ConfGen, 2023.

[43] K. Shawn Watts, Pranav Dalal, Robert B. Murphy, Woody Sherman, Rich A. Friesner, and John C. Shelley. ConfGen: A Conformational Search Method for Efficient Generation of Bioactive Conformers. *Journal of Chemical Information and Modeling*, 50(4):534–546, 4 2010.

[44] Schrödinger Release 2023-2. Phase, 2023.

[45] David S Wishart, Yannick D Feunang, An C Guo, Elvis J Lo, Ana Marcu, Jason R Grant, Tanvir Sajed, Daniel Johnson, Carin Li, Zinat Sayeeda, Nazanin Assempour,

Ithayavani Iynkkaran, Yifeng Liu, Adam Maciejewski, Nicola Gale, Alex Wilson, Lucy Chin, Ryan Cummings, Diana Le, Allison Pon, Craig Knox, and Michael Wilson. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 1 2018.

[46] Wendy Shelly, Michael W. Draper, Venkatesh Krishnan, Mayme Wong, and Robert B. Jaffe. Selective Estrogen Receptor Modulators: An Update on Recent Clinical Findings. *Obstetrical & Gynecological Survey*, 63(3):163–181, 3 2008.

[47] Dennis J. Cada and Danial E. Baker. Conjugated Estrogens and Bazedoxifene. *Hospital Pharmacy*, 49(3):273–283, 3 2014.

[48] Michael B. Sporn. Arzoxifene: A Promising New Selective Estrogen Receptor Modulator for Clinical Chemoprevention of Breast Cancer. *Clinical Cancer Research*, 10(16):5313–5315, 8 2004.

[49] Ju Liu, Hong Liu, Richard B. van Breemen, Gregory R. J. Thatcher, and Judy L. Bolton. Bioactivation of the Selective Estrogen Receptor Modulator Acolbifene to Quinone Methides. *Chemical Research in Toxicology*, 18(2):174–182, 2 2005.

[50] Timothy A Blizzard, Frank DiNinno, Jerry D Morgan, Jane Y Wu, Helen Y Chen, Seongkon Kim, Wanda Chan, Elizabeth T Birzin, Yi Tien Yang, Lee-Yuh Pai, Zhoupeng Zhang, Edward C Hayes, Carolyn A DaSilva, Wei Tang, Susan P Rohrer, James M Schaeffer, and Milton L Hammond. Estrogen receptor ligands. Part 8: Dihydrobenzoxathiin SERAMs with heteroatom-substituted side chains. *Bioorganic & Medicinal Chemistry Letters*, 14(15):3865–3868, 8 2004.

[51] Thu Ha Pham, Yann Le Page, Frédéric Percevault, François Ferrière, Gilles Flouriot, and Farzad Pakdel. Apigenin, a Partial Antagonist of the Estrogen Receptor (ER), In-

hibits ER-Positive Breast Cancer Cell Proliferation through Akt/FOXM1 Signaling. *International Journal of Molecular Sciences*, 22(1):470, 1 2021.

[52] Thomas T.Y. Wang, Neeraja Sathyamoorthy, and James M. Phang. Molecular effects of genistein on estrogen receptor mediated pathways. *Carcinogenesis*, 17(2):271–275, 2 1996.

[53] Claudia Montani, Marialetizia Penza, Marija Jeremic, Giorgio Biasiotto, Gina La Sala, Massimo De Felici, Paolo Ciana, Adriana Maggi, and Diego Di Lorenzo. Genistein is an Efficient Estrogen in the Whole-Body throughout Mouse Development. *Toxicological Sciences*, 103(1):57–67, 5 2008.

[54] Wei Zhang, Jo-Hsin Chen, Irene Aguilera-Barrantes, Chung-Wai Shiau, Xiugui Sheng, Li-Shu Wang, Gary D. Stoner, and Yi-Wen Huang. Urolithin A suppresses the proliferation of endometrial cancer cells by mediating estrogen receptor-α-dependent gene expression. *Molecular Nutrition & Food Research*, 60(11):2387–2395, 11 2016.

[55] Karen K. L. Chan, Michelle K. Y. Siu, Yu-xin Jiang, Jing-jing Wang, Thomas H. Y. Leung, and Hextan Y. S. Ngan. Estrogen receptor modulators genistein, daidzein and ERB-041 inhibit cell migration, invasion, proliferation and sphere formation via modulation of FAK and PI3K/AKT signaling in ovarian cancer. *Cancer Cell International*, 18(1):65, 12 2018.

[56] Alison K. Death, Kristine C. Y. McGrath, Rymantas Kazlauskas, and David J. Handelsman. Tetrahydrogestrinone Is a Potent Androgen and Progestin. *The Journal of Clinical Endocrinology & Metabolism*, 89(5):2498–2500, 5 2004.

[57] Ashley Gillon, Emmie N.M. Ho, George H.M. Chan, Alexia Kauff, Gillian Hughes, Rachel A. Lund, Zoe Ashley, Terence S.M. Wan, and Alison K. Heather. Unravelling

androgens in sport: Altrenogest shows strong activation of the androgen receptor in a mammalian cell bioassay. *Drug Testing and Analysis*, 13(3):523–528, 3 2021.

[58] Claude Bonne and Jean-Pierre Raynaud. Methyltrienolone, a specific ligand for cellular androgen receptors. *Steroids*, 26(2):227–232, 8 1975.

[59] Henri Rozenbaum. Relationships between chemical structure and biological properties of progestogens. *American Journal of Obstetrics and Gynecology*, 142(6):719–724, 3 1982.

[60] Hua Tian, Rui Liu, Suqiu Zhang, Shuhui Wei, Wei Wang, and Shaoguo Ru. 17*β*-Trenbolone binds to androgen receptor, decreases number of primordial germ cells, modulates expression of genes related to sexual differentiation, and affects sexual differentiation in zebrafish (Danio rerio). *Science of The Total Environment*, 806:150959, 2 2022.

[61] H.J. Kloosterboer, C.A. Vonk-Noordegraaf, and E.W. Turpijn. Selectivity in progesterone and androgen receptor binding of progestagens used in oral contraceptives. *Contraception*, 38(3):325–332, 9 1988.

[62] Daniel L.J. Thorek, Anson T. Ku, Nicholas Mitsiades, Darren Veach, Philip A. Watson, Dipti Metha, Sven-Erik Strand, Sai Kiran Sharma, Jason S. Lewis, Diane S. Abou, Hans G. Lilja, Steven M. Larson, Michael R. McDevitt, and David Ulmert. Harnessing Androgen Receptor Pathway Activation for Targeted Alpha Particle Radioimmunotherapy of Breast Cancer. *Clinical Cancer Research*, 25(2):881–891, 1 2019.

[63] H Kuhl. Pharmacology of estrogens and progestogens: influence of different routes of administration. *Climacteric*, 8(sup1):3–63, 8 2005.

[64] Marie Shamseddin, Fabio De Martino, Céline Constantin, Valentina Scabia, Anne-Sophie Lancelot, Csaba Laszlo, Ayyakkannu Ayyannan, Laura Battista, Wassim Raffoul, Marie-Christine Gailloud-Matthieu, Philipp Bucher, Maryse Fiche, Giovanna Ambrosini, George Sflomos, and Cathrin Brisken. Contraceptive progestins with androgenic properties stimulate breast epithelial cell proliferation. *EMBO Molecular Medicine*, 13(7):e14314, 7 2021.

[65] facebookresearch/faiss: A library for efficient similarity search and clustering of dense vectors. https://github.com/facebookresearch/faiss.

[66] R. W. Kennard and L. A. Stone. Computer Aided Design of Experiments. *Technometrics*, 11(1):137–148, 2 1969.

[67] Mario Krenn, Florian Häse, AkshatKumar Nigam, Pascal Friederich, and Alan Aspuru-Guzik. Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 12 2020.

[68] Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. SELFIES and the future of molecular string representations. *Patterns*, 3(10):100588, 10 2022.

[69] Noel M O'Boyle, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch,

and Geoffrey R Hutchison. Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1):33, 12 2011.

[70] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017-December:5999–6009, 6 2017.

It is not that machines are going to replace chemists. It's that the chemists who use machines will replace those that don't.

Derek Lowe

6

Outlook and Conclusion

In this thesis, we cover ligand-based and structure-based methods and tools for drug development and the safety assessment of small molecules. However, in reflection of Albert Einstein's famous quote "The more I learn, the more I realize how much I don't know", we understand that the development of these tools is an ongoing process, and there is much to improve and refine.

First, we showed that PanScreen, our online platform for the automated screening of off-target liabilities, had very promising performance in our tests. The next step will be to validate these results using in vitro tests and to find out in which cases the platform works and in which cases it does not. This will help us to further improve the quality and robustness of the predictions.

A low-hanging fruit in improving the platform is the revision of predicting interaction fingerprints using po-sco. The currently implemented model was trained on the PDBbind dataset to predict binding affinities using the interaction information extracted by po-sco. However, since PanScreen applies this model to complexes containing docked ligand poses, training it on crystal structures is not the best approach. Thus, the model could be improved in three steps: 1) A better suited dataset needs to be used for training. This can involve creating a new dataset that contains high-quality crystal structures that have undergone a strict and standardized quality assessment and structural preparation. Otherwise, an existing dataset such as MISATO can be used.[1] In any case, the ligands in the dataset should be re-docked or cross-docked using various docking programs. These steps will help the model train on the same kind of data to which it will be applied. 2) Before attempting to predict binding affinities, the best pose, i.e. the pose with the most favorable interactions with the protein, should be identified. A very similar model architecture to what was presented in Chapter 2 could be used to re-rank ligand poses. 3) A model can be trained to predict the binding affinities for the best identified poses according to step 2), in the dataset of step 1). We expect that this approach will lead to much better results.

Another opportunity for improvement lies in the implementation of additional methods. Specifically, physics-informed neural networks such as PIGNet could be a valuable addition to the methods currently implemented in PanScreen.[2] Implementing docking programs that allow to account for induced fit effects such as DOLINA could help improve predictions in cases that are limited by the low conformational diversity of the implemented structure ensemble.[3]

In general, the use of deep Taylor decomposition could help improve the interpretability of the implemented deep neural networks.[4] This method can provide insight into the influence of a model's input components on its prediction. This could allow, for example, to analyze which interactions in a protein-ligand complex are the driving force for a high (or low) predicted binding affinity. Therefore, it could be possible to identify which moieties in a molecule are responsible for binding to a specific off-target, providing assistance in the development of safe drugs.

As some of the implemented off-targets are known to be subject to different modes of action, mainly agonism or antagonism, it would be beneficial to provide further insight into the consequences of a binding event. For some proteins with clearly distinct agonistic or antagonistic binding site conformations, this could be as easy as identifying to which conformation a small molecule binds the strongest. However, there are also much more complex situations, which is often the case in G-protein coupled receptors. For these proteins, there are also inverse agonists, partial agonists, and neutral antagonists.[5] This makes prediction of the exact mode of action much more difficult and will likely not be possible without much additional effort.

An advantage of computational tools such as PanScreen is that it is relatively easy to model mutated protein structures. Therefore, it could be interesting to model polymorphisms, especially for proteins such as CYPs. This would bring the platform one step closer to the development of personalized medicine. However, training models for specific isoforms of proteins could be difficult due to the lack of large isoform-specific datasets. Therefore, at least for the time being, information on polymorphism could be implemented by checking for intermolecular interactions with polymorphic hotspots, frequently mutated residues, and raising warning flags.

A limitation of the PanScreen platform is that small molecules are only docked to the orthosteric binding site of the implemented off-targets. Theoretically, it is also possible to dock to allosteric sites. However, allosteric binding sites are not always known and it is difficult to find binding affinity information for specific allosteric sites. The first problem could be overcome by using binding site prediction tools to identify potential allosteric sites for a given small molecule. Alternatively, novel methods such as RoseTTAFold All-Atom or Umol could be employed to model the protein around a small molecule and, if necessary, dock to the generated binding site conformation.[6,7] However, this does not solve the second problem. In the current approach, binding affinity information is needed to train the models. Therefore, if no such information is available for specific allosteric binding sites, no models can be trained. It would, however, be possible to make predictions based on generalized models, probably at the cost of lower accuracy.

Similarly, including models such as AlphaFold2, RoseTTAFold All-Atom, or Umol, could allow to generate possible protein and binding site conformations for any protein with known primary sequence.[6–8] It is therefore possible to implement a low-accuracy/high-coverage mode in which the protein structure for any protein in the Uniprot database can be generated on-the-fly. Due to the lack of protein-specific models, this would again require the use of generalized models, which likely have a lower accuracy.

PanScreen only predicts off-target binding and not adverse effects. Knowing whether a small molecule binding to an off-target actually causes side effects in humans is essential in chemical safety assessment. Once the portfolio of off-targets implemented in PanScreen grows in size, it could be possible to create interaction profiles encoding the predicted interaction strength between a small molecule and many off-targets. These interaction profiles could then be used to predict adverse effects, for example, using epidemiological databases such as SIDER.[9]

The predicted off-target interactions can also be seen as molecular initiating events in an

AOP. Therefore, another way of predicting actual side effects could be to identify AOPs beginning with a ligand binding to one of the implemented off-targets. Using, for example, information from the AOP wiki, it could then be possible to predict whether or not the key events leading to an adverse outcome will be triggered.[10] However, this approach would likely involve more than just the prediction of protein-ligand binding, and other methods would be required.

Further, the ability of a small molecule to bind to a protein becomes relevant only if the molecule can reach the protein in the human body. Hence, prediction of pharmacokinetics, for example, with PBPK methods, would help in the comprehensive assessment of a chemical's safety.

As already mentioned in Chapter 4, a model predicting 3D shape similarity could be used to prioritize the off-targets implemented in PanScreen. Furthermore, since ligand-based methods often work well in identifying potential binders (especially those with a similar chemical structure or shape), it would also be possible to integrate the predicted shape similarity to other molecules known to bind to a given off-target into the consensus models.

We showed that the prediction of similarities based on latent space distances could be applied to 2D similarities and 3D shape similarities. The next step toward increasing the usefulness of this technology could be to apply it to molecular field points.[11] This could further improve the ability of such a model to find molecules that bind to a given (off-)target.

In conclusion, this work contributes to the advancement of in silico tools in drug development and chemical safety assessment. However, the journey toward excellence is still ongoing, and the ideas presented in this chapter can provide a roadmap for future improvement of the developed technologies in the hope of one day providing a viable alternative to in vivo toxicity testing in animals.

References

[1] Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Stefan Kesselheim, Marie Piraud, Fabian J. Theis, Michael Sattler, and Grzegorz M. Popowicz. Misato - machine learning dataset of protein-ligand complexes for structure-based drug discovery. *bioRxiv*, page 2023.05.24.542082, 5 2023.

[2] Seokhyun Moon, Sang-Yeon Hwang, Jaechang Lim, and Woo Youn Kim. Pignet2: A versatile deep learning-based protein-ligand interaction prediction model for binding affinity scoring and virtual screening. 7 2023.

[3] Martin Smieško. Dolina – docking based on a local induced-fit algorithm: Application toward small-molecule binding to nuclear receptors. *Journal of Chemical Information and Modeling*, 53:1415–1423, 6 2013.

[4] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222, 5 2017.

[5] William I. Weis and Brian K. Kobilka. The molecular basis of g protein–coupled receptor activation. *Annual Review of Biochemistry*, 87:897–919, 6 2018.

[6] Rohith Krishna, Jue Wang, Woody Ahern, Pascal Sturmfels, Preetham Venkatesh, Indrek Kalvet, Gyu Rie Lee, Felix S Morey-Burrows, Ivan Anishchenko, Ian R Humphreys, Ryan McHugh, Dionne Vafeados, Xinting Li, George A Sutherland, Andrew Hitchcock, C Neil Hunter, Minkyung Baek, Frank DiMaio, and David Baker. Generalized biomolecular modeling and design with rosettafold all-atom. *bioRxiv*, page 2023.10.09.561603, 10 2023.

[7] Patrick Bryant, Atharva Kelkar, Andrea Guljas, Cecilia Clementi, and Frank Noé. Structure prediction of protein-ligand complexes from sequence information with umol. *bioRxiv*, page 2023.11.03.565471, 11 2023.

[8] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 8 2021.

[9] Michael Kuhn, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. The sider database of drugs and side effects. *Nucleic Acids Research*, 44:D1075–D1079, 1 2016.

[10] Aop-wiki. https://aopwiki.org/.

[11] Florian B Hinz, Amr H Mahmoud, and Markus A Lill. Prediction of molecular field points using se(3)-transformer model. *Machine Learning: Science and Technology*, 4:035016, 9 2023.

6.2 List of Publications

[1] **Sellner, M. S.**, Mahmoud, A. H., & Lill, M. A. (2023). Enhancing Ligand-Based Virtual Screening with 3D Shape Similarity via a Distance-Aware Transformer Model. BioRxiv, 2023.11.17.567506. https://doi.org/10.1101/2023.11.17.567506

[2] **Sellner, M. S.**, Lill, M. A., & Smiesko, M. (2023). PanScreen: A Comprehensive Approach to Off-Target Liability Assessment. BioRxiv, 2023.11.16.567496. https://doi.org/10.1101/2023.11.16.567496

[3] **Sellner, M. S.**, Lill, M. A., & Smieško, M. (2023). Quality Matters: Deep Learning-Based Analysis of Protein-Ligand Interactions with Focus on Avoiding Bias. BioRxiv, 2023.11.13.566916. https://doi.org/10.1101/2023.11.13.566916

[4] **Sellner, M. S.**, Mahmoud, A. H., & Lill, M. A. (2023). Efficient virtual high-content screening using a distance-aware transformer model. Journal of Cheminformatics, 15(1), 18. https://doi.org/10.1186/s13321-023-00686-z

[5] Inderbinen, S. G., Kley, M., Zogg, M., **Sellner, M.**, Fischer, A., Kędzierski, J., Boudon, S., Jetten, A. M., Smieško, M., & Odermatt, A. (2022). Activation of retinoic acid-related orphan receptor γ(t) by parabens and benzophenone UV-filters. Toxicology, 471, 153159. https://doi.org/10.1016/j.tox.2022.153159

[6] Papaj, K., Spychalska, P., Kapica, P., Fischer, A., Nowak, J., Bzówka, M., **Sellner, M.**, Lill, M. A., Smieško, M., & Góra, A. (2022). Evaluation of Xa inhibitors as potential inhibitors of the SARS-CoV-2 Mpro protease. PLOS ONE, 17(1), e0262482. https://doi.org/10.1371/journal.pone.0262482

[7] Fischer, A., Bardakci, F., **Sellner, M.**, Lill, M. A., & Smieško, M. (2023). Ligand pathways in estrogen-related receptors. Journal of Biomolecular Structure and Dynamics, 41(5), 1639–1648. https://doi.org/10.1080/07391102.2022.2027818

[8] Fischer, A., Smieško, M., **Sellner, M.**, & Lill, M. A. (2021). Decision Making in Structure-Based Drug Discovery: Visual Inspection of Docking Results. Journal of Medicinal Chemistry, 64(5), 2489–2500. https://doi.org/10.1021/acs.jmedchem.0c02227

[9] Fischer, A., **Sellner, M.**, Mitusińska, K., Bzówka, M., Lill, M. A., Góra, A., & Smieško, M. (2021). Computational Selectivity Assessment of Protease Inhibitors against SARS-CoV-2. International Journal of Molecular Sciences, 22(4), 2065. https://doi.org/10.3390/ijms22042065

[10] **Sellner, M.**, Fischer, A., Don, C. G., & Smieško, M. (2021). Conformational Landscape of Cytochrome P450 Reductase Interactions. International Journal of Molecular Sciences, 22(3), 1023. https://doi.org/10.3390/ijms22031023

[11] Fischer, A., **Sellner, M.**, Neranjan, S., Smieško, M., & Lill, M. A. (2020). Potential Inhibitors for Novel Coronavirus Protease Identified by Virtual Screening of 606 Million Compounds. International Journal of Molecular Sciences, 21(10), 3626. https://doi.org/10.3390/ijms21103626