CHAPTER I

Introduction

1. Introduction

1.1. Revolutionizing drug discovery: The power of computer-aided drug design

The process of discovering new active compounds is complex task and requires expertise in various fields, including chemistry and biology. The primary goals in drug development process are to ensure efficacy, safety, and overall benefits to human health [1, 2]. The impact of physicians and researchers in developing new drugs to treat various diseases holds massive importance due to its direct impact on the lives of patients [3]. The process of developing a medicine involves several stages, including target identification, hit design/discovery, hit-to-lead optimization, and preclinical and clinical research. However, this process is both time-consuming and costly, taking approximately 10 to 17 years and requiring an estimated investment of US$ 2.558 billion. Surprisingly, the success rate is only around 10%. Therefore, there is a need for new strategies to enhance success rates, reduce the time taken, and improve overall efficiency [4].

In 1990, first computer-guided techniques was used for the discovery of Viracept (first HIV protease inhibitor), which started a revolution in pharmaceutical field. So, scientists and pharmaceutical companies had grown their interest to discover or design new drugs through computational methods to reduce costs and increase their chances of success [5]. During the early days of computer aided drug design (CADD), different *in silico* methods were employed as computational power, becoming the essential foundation for any drug discovery campaign [5, 6] (**Figure 1.1**).

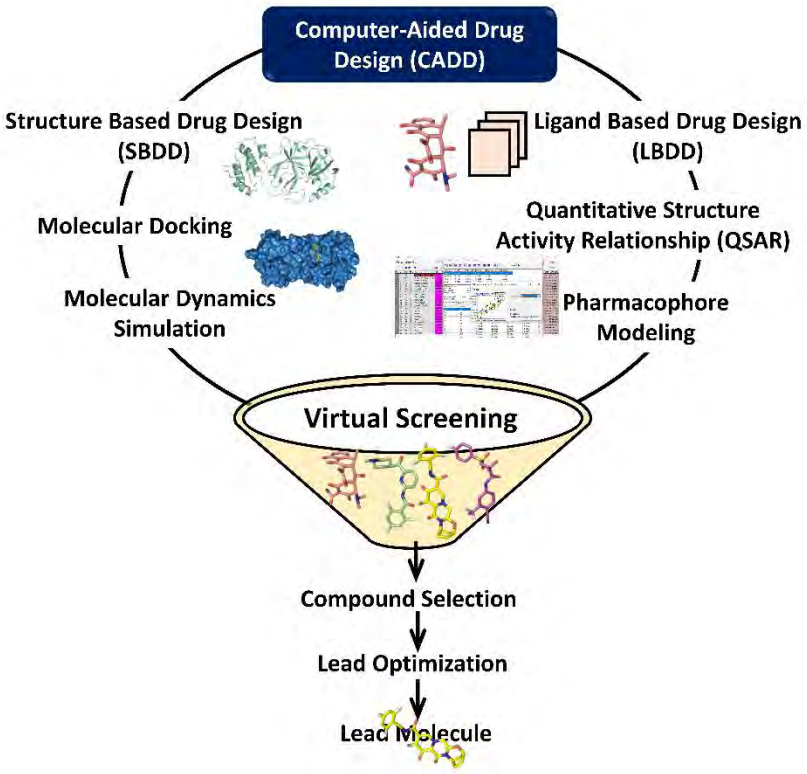

Figure 1.1. The main approaches in CADD methods that are currently employed.

Rational drug design begins with identifying biological targets relevant to diseases and aims to create new drugs with different modes of action. CADD have modernized the field by lowering the costs and improving success rates. Nowadays, many advanced technologies like high-throughput screening and artificial intelligence, are being employed to accelerate early-stage drug development. Drug repurposing is also gaining attention as a potential faster and more cost-effective approach [7-9].

1.2. Evolution and significance of CADD

CADD integrates important computational methods with traditional drug design and facilitate the overall discovery and design of new drugs. It involves useful computer algorithms, molecular modeling, and simulation techniques to understand the interactions between drug and its target. In early days, the evolution of CADD started when researchers used computer to model three-dimensional structures of molecule. Initially, it focused for generating static models to study the shape and electrostatic properties of molecule. In recent time, it has become more advanced with the developments in computing power and thus capable of explain complex drug discovery challenges. However, it has expanded enormous significance in pharmaceutical industry due to several reasons:

a) **Accelerated drug discovery:** CADD is most prominent to speed up the drug discovery process by reducing time and costs. Through virtual screening techniques, we can efficiently analyze large databases and focus on most promising molecules. This resources improves the efficiency of drug discovery procedures.
b) **Rational drug design:** CADD provides a rational and systematic approach to drug design. Researchers can gain insights into the binding mechanism of a drug with its target at molecular level. That's why, for better efficacy, it can help to optimize the drug like properties and reduce side effects. It can also be used to modify existing drugs and design new molecules with specific target interaction, thereby increasing the success rate of drug development.
c) **Reduced attrition rates:** The attrition rate is very high for drug development as many promising compounds are failed in later stages of clinical trial. CADD offers valuable aspects at the beginning level of drug discovery, thereby reduce the failure risk. Besides, it predicts absorption, distribution, metabolism, excretion properties (ADME) and also toxicity and helps to detect the problem initially before expensive experiments are conducted.
d) **Target identification and validation:** It has crucial role in target identification and validation of drug discovery. By considering the three-dimensional structure of target protein, it helps to identify the suitable binding site and design new molecules. This knowledge supports in understanding the interaction mechanism and develops target therapy.
e) **De novo drug design:** CADD has expanded the capability of de novo drug design process by screening existing compounds from relevant database. Researchers can now use computational algorithm to design new molecules with specific characteristics (improved potency or selectivity).

1.3. Principles and techniques of CADD

CADD is the strategy which plays a major role in the early stage of drug discovery and development. Based on the different basic principles, CADD can be roughly divided into two types: (a) Structure-based drug design (SBDD) and (b) Ligand-based drug design (LBDD). SBDD and LBDD have evolved and transformed modern drug discovery methods which allow hit identification and lead optimization into powerful tools that make a huge impact on modern drug discovery [10].

1.4. Aurora kinase

Aurora kinase (AURK) is a family of serine/threonine kinases that play important roles in cell division. They are involved in a variety of cellular processes, including chromosome segregation, centrosome maturation, and mitotic spindle assembly. The AURK family was first discovered in the early 1990s. The first member of the family, Aurora A, was identified in 1992 by researchers at the University of California, San Francisco.

1.4.1. Types of Aurora kinase

The Aurora kinase family consists of three subtypes: Aurora A, B, and C. Although they have different functions, they share a common feature: a highly conserved C-terminal domain and an N-terminal domain [11]. Aurora A is an important participant in the cell cycle, performing various roles. It controls mitotic entry, helps in centrosome maturation, facilitates spindle assembly, and aids in repairing damaged spindles [12]. Ajuba activates Aurora A during the S phase and at the G2/M transition [11]. The function of Aurora A relies on the phosphorylation of a specific threonine residue (T288) in its activation loop. In prometaphase, it contributes to spindle assembly and microtubule organization by phosphorylating TPX2 [13]. Besides its involvement in mitosis, Aurora A also participates in DNA damage response [14] and triggers epithelial-mesenchymal transition [15]. Inhibiting Aurora A can result in concerns with chromosome distribution, abnormal spindle structure, and aneuploidy [16].

Aurora B is a protein that localizes to the centromere and phosphorylates histone H3 at serine 10 during mitosis. It collaborates with other proteins to ensure proper kinetochore-microtubule attachments [11]. It supports the cell cycle and survival by phosphorylating p53 and promoting its degradation [17]. Inhibiting Aurora B disrupts the spindle checkpoint, can cause premature termination of cell division, and may lead to polyploidy. It also interferes with cytokinesis and chromosomal segregation by disrupting microtubule connections [18].

Aurora C, exclusively found in the testis, has a role similar to Aurora B in chromosomal segregation [19]. However, it is linked to chromosomal segregation abnormalities and can promote cancer cell survival and centrosome amplification [20].

1.4.2. Function of Aurora kinase

Aurora kinases play a role in various mitotic checkpoints, and their abnormal expression has been linked to tumor development.

a) **In mitosis:** Aurora A is crucial for centrosome duplication, maturation, and microtubule spindle formation [21]. It is activated through the Ran signaling pathway and promotes cell death and cell cycle arrest after mitosis [19]. Aurora B is essential for chromosome segregation and cytokinesis. It is activated by cofactors such as the chromosome passenger complex (CPC), borealin, survivin, and inner centromere protein (INCENP) [22]. Aurora B ensures proper attachment of the microtubule spindle to kinetochores and phosphorylates various substrates involved in chromosome condensation, spindle assembly, and stabilization. It also facilitates cytokinesis by phosphorylating kinesin family member 2A (KIF2A) [23]. Blocking Aurora B function leads to chromatin de-condensation and cell death. Aurora C shares similar functions to Aurora B and interacts with transforming, acidic coiled-coil containing protein 1 (TACC-1) during cytokinesis [24].

b) **In tumorigenesis:** AURKs were initially found in the breast cancer gene STK15 on chromosome 20q13. They are excessively produced in various types of human cancers, such as prostate, colorectal, ovarian, breast, neuroblastoma, and cervical cancer [25]. AURKs interact with several tumor suppressor genes, like p53, breast cancer type 1 (BRCA1), and BRCA2, and this interaction plays a crucial role in promoting the development of tumors [26].

Aurora A can add phosphate groups to p53 at specific sites (Ser215 and Ser315) [27], which reduces p53 ability to bind to DNA and decreases its transcriptional activity [28]. Aurora B can also hinder p53 activity by phosphorylating different sites (Ser183, Thr211, and Ser215) [17]. BRCA1, a tumor suppressor protein, is important for DNA repair, proper chromosome segregation during mitosis, and chromatin regulation. Its role in mitosis is specifically mediated through the tumor suppressor kinase checkpoint Chk2 [26].

1.5. Hepatitis C virus

Hepatitis C virus (HCV) is a highly common infectious agent that is making more and more people sick with liver issues and causing a major global health problem. While the current treatments can effectively help over 95% of people diagnosed with HCV, it is hard for many to get the medical care they need, because of non-availability of good vaccine [29]. Furthermore, acute HCV infections often have no symptoms, and around 70% of patients develop a chronic infection, which can lead to various liver issues, including cirrhosis and potentially hepatocellular carcinoma [30]. HCV can be divided into seven established genotypes and 67 subtypes. These genetic variations can result in different outlooks for hepatitis C and impact the choice of antiviral treatments [31].

1.5.1. Discovery of the hepatitis C virus

During the 1970s, there was a remarkable increase in hepatitis cases linked to blood transfusions, leading to post-transfusion hepatitis. However, researchers discovered that the known hepatotropic viruses, like hepatitis A and B, were not causing this new type of hepatitis. So, they called it non-A-non-B hepatitis (NANBH) [32]. This observation led to years of extensive research. Finally, in 1989, two papers were published in the same issue of the

journal Science, describing the isolation of viral cDNA clones [33]. This breakthrough allowed scientists to identify NANBH antibodies and ultimately led to the discovery of HCV [34].

1.5.2. Molecular biology of the hepatitis C virus

HCV belongs to the Flaviviridae family and is classified under the Hepacivirus genus. Its genetic material consists of a single-stranded RNA. The HCV genome is approximately 9.6 kb long and contains a large polyprotein precursor. This precursor is cleaved by both viral and host proteases to form structural (core, E1, and E2) and non-structural (p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B) proteins (**Figure 1.2**) [30]. The single-stranded RNA is enclosed in an icosahedral protein coat and enveloped in a lipid bilayer. Within this lipid envelope, highly glycosylated E1 and E2 glycoproteins are found [35, 36]. Newer HCV treatments have improved because of direct-acting antiviral (DAA) drugs. They mainly target NS proteins, which are essential for the virus to replicate, reducing its ability to proliferate [37].

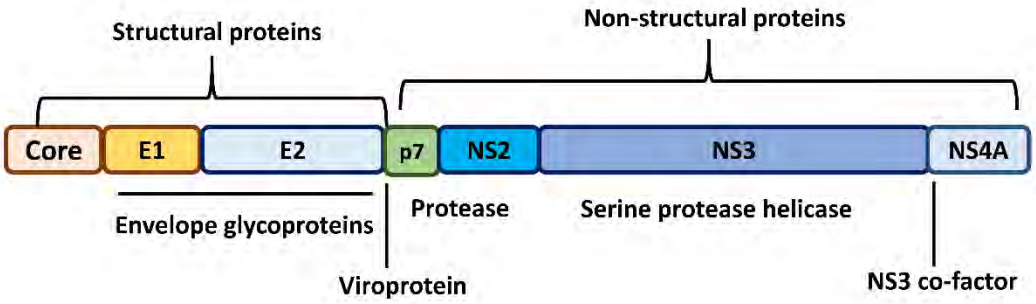

Figure 1.2. Schematic representation of HCV polyprotein referring to the individual proteins and their relevance for the HCV life cycle.

1.5.2.1. NS3/4A protease

The HCV NS3/4A protease is a complex consisting of NS3 and NS4A, and it plays multiple roles. In addition to its involvement in viral replication and assembly, it is essential for evading the antiviral immunity [38, 39]. One of its primary functions involves cleaving mitochondrial antiviral signaling protein (MAVS), a protein present in various cellular membranes. This cleavage disrupts signaling pathways, ultimately causing a reduction in the expression of interferons and interferon-stimulated genes. This effect has been observed in the liver of HCV-infected patients. NS3-4A also cleaves TIR domain-containing adaptor protein inducing interferon beta (TRIF), an adaptor protein for toll-like receptor 3 (TLR3), further suppressing interferon and interferon-stimulated gene (ISG) expression [40, 41]. This is significant because HCV infection is primarily detected by retinoic acid-inducible gene-I-like receptors (RLRs) and TLR3. NS3/4A has the way of affecting both pathways to potentially reduce excessive inflammation caused by HCV and slow down the immune cell-mediated defense. Besides impacting these pathways, NS3/4A also interferes with the innate immune response. When HCV activates RLRs or TLR3, it initiates interferon regulatory factor 3 (IRF3). NS3/4A binds to TANK-binding kinase 1 (TBK1), prevents it from activating IRF3 [42, 43]. As a result, it hinders the initiation of interferon expression [44]. NS3/4A

also has an impact on Riplet, which is an E3 ubiquitin ligase which plays a vital role in retinoic acid-inducible gene I (RIG-I) activation [45]. Vazquez et al. found that the NS4A Y16F residue is responsible for inhibiting Riplet, but this does not affect MAVS cleavage [46]. Besides, some reports indicate that HCV can inhibit the activation of NF-κB by TNF-α [47, 48]. Chen et al. established that HCV NS3 interacts with linear ubiquitin chain assembly complex (LUBAC). This complex is responsible for a unique type of protein modification called polyubiquitylation of the NF-κB essential modulator (NEMO), which is crucial for activating NF-κB. As a result, this interaction inhibits the expression of various inflammatory cytokines [49]. It has also been reported to facilitate the degradation of signal transducer and activator of transcription 1 (STAT1) [50].

1.5.3. Developed HCV therapeutics: Success and failure

For HCV drug development, considerable human resources and financial investments are employed to develop new antiviral drugs. Though a perfect antiviral therapy totally for HCV is not yet available, the FDA has approved and recommended many small-molecule drugs for individual or combination use, and others are in various stages of development, including clinical trials. These drugs primarily target specific viral components like NS3/4A, NS5A, NS5B, or the internal ribosome entry site (IRES) [51].

Danoprevir is a small-molecule macrocyclic peptidomimetic drug produced by Roche Pharmaceuticals. The HCV NS3/4A protease is competitively inhibited by this drug. The drug was introduced to the Chinese market in August 2018 after successfully completing phase III clinical studies. Clinical trials with the drug telaprevir (VX-950), which inhibits the NS3/4A protease, had limited success. It is being marketed by Janssen R&D Ireland all over the world [52].

1.6. Drug repurposing

The approval process for a new drug is lengthy and costly, taking an average of 10–15 years. This lengthy exploration process creates an opportunity for repurposing drugs (finding new uses) as another way to speed up drug development. Drug repurposing involves finding new purposes for drugs that are already available, and it is generally seen as an effective and cost-friendly method [53]. Repurposing, also called reprofiling, re-tasking, repositioning, or drug rescue, offers a cost-effective and quicker way to discover new disease treatments, especially when preclinical safety studies have been conducted [54]. When a drug is repurposed, it means using drugs that are already approved by agencies like the FDA, European Medicines Agency (EMA), Medicines and Healthcare products Regulatory Agency (MHRA), and others, for a different reason [53]. It can be really important for deciding the best treatments for patients with rare, complex, or long-lasting illnesses that have few treatment options [54]. Since it can make things faster, many pharmaceutical companies are using drug repurposing to redevelop drugs which already got approved by the FDA, including ones that didn't work before, into new treatments for different diseases [53]. The use of computational approaches has brought about new ways of thinking about repurposing drugs [54].

1.6.1. Significance of drug repurposing

New drugs must adhere to strict regulations before they can be introduced to the market. Discovering a new drug and making improvements to it demands a substantial investment, mainly due to the varied physical and chemical characteristics of the compounds involved, as well as the challenges in increasing production on a larger scale [55-57]. This limitation gives more power to pharmaceutical companies or academic institutions, enabling them to rapidly and effectively use already-approved drugs for a new purpose that has not been available to patients with that illness before. Usually, experimental substances that do not work as expected for a specific purpose can be a valuable beginning for repurposing and revitalizing them. These compounds could also be explored again for a different purpose, eventually becoming effective treatments, especially helpful for rare diseases that are challenging to diagnose, treat, and lack necessary resources. Repurposing drugs provides a faster and cost-effective approach to delivering effective treatments to patients, in contrast to the time-consuming and complex traditional methods of drug discovery [53].

In contrast, a repurposed molecule comes with all safety, preclinical, and effectiveness information readily accessible. This allows the researcher to make well-informed choices at every step of drug development [58]. Having prior knowledge about the safety, effectiveness, and proper usage of a drug significantly decreases development costs and speeds up the process. This ultimately makes it less challenging to successfully introduce a repositioned drug to the market [59].

It becomes a suitable alternative when unexpected medical situations like COVID-19 pandemic happen in March 2020 [60, 61]. Healthcare systems have faced unique challenges during these outbreaks. The urgent requirement for proper pharmacological therapies provides medical experts with limited time for new drug discovery. Finding new drugs is difficult and requires several years, from the initial idea to being available in the market [62]. As new virus variants have evolved more quickly than anticipated, the demand for vaccines and drugs has grown more urgent [63, 64].

1.7. COVID-19

The emergence of novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) led to a worldwide COVID-19 pandemic. Initially identified in Wuhan, China, during a SARS outbreak in December 2019, SARS-CoV-2 quickly spread, prompting the World Health Organization (WHO) to declare a pandemic in March 2020 [65]. Coronaviruses (CoVs) are part of the Coronaviridae family, a group of enveloped positive RNA viruses with single-stranded genetic material. This family is categorized into four subgroups: alpha (α), beta (β), gamma (γ), and delta (δ). Among these, the beta (β-CoV) category includes severe acute respiratory syndrome coronavirus (SARS-CoV), the coronavirus behind COVID-19 known as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), and Middle East respiratory syndrome coronavirus (MERS-CoV) [66]. These fatal viruses cause harm to the respiratory, liver, gastrointestinal, and central nervous systems in both humans and animals. The SARS-CoV-2 virus, accountable for COVID-19, exhibits greater pathogenicity compared to SARS-CoV and MERS-CoV, leading to fatal disease [54]. The main route of

transmission for SARS-CoV-2 involves direct contact with droplets and aerosols containing the respiratory virus emitted by infected individuals. Coughing, sneezing, and nasal discharge are significant means of spreading the contagion. However, the detection of SARS-CoV-2 genetic material in various organs indicates a wide-ranging tropism, not confined solely to the upper and lower respiratory tracts. This may partly be explained by the fact that the main cellular receptor for SARS-CoV-2, angiotensin-converting enzyme 2 (ACE-2), is expressed in several human tissues and organs [67]. Common indications of COVID-19 comprise fever, cough, fatigue, and breathing difficulties [68]. However, patients and individuals in the recovery phase have reported dysfunction in multiple organs, including the nervous system, heart, liver, kidneys, and more. The disease outcome is often fatal, primarily attributed to severe viral pneumonia symptoms that particularly affect older individuals and those with weakened immune systems. Furthermore, the presence of additional concurrent clinical conditions, like chronic cardiac disease and diabetes, is also considered crucial risk factors that influence the disease outcome [65].

1.7.1. Drug targets of COVID-19

Positive-sense RNA viruses with a genome size around 30 kb make up the human coronaviruses. These viruses contain two distinct sets of proteins: structural proteins (Spike (S), nucleocapsid (N), membrane (M), and envelope (E)) and non-structural proteins (nsp1 to nsp16), including essential elements like the RNA-dependent RNA polymerase (RdRp) (nsp12). Around two-thirds of the 50 end of the SARS-CoV-2 genome is responsible for encoding two overlapping polyproteins: pp1a and pp1ab12. These viral proteases break them down into 16 non-structural proteins (NSPs), which play a crucial role in viral replication and transcription. A group of structural proteins, which includes the nucleocapsid (N), spike (S) protein, membrane (M) protein, and envelope (E) protein, is encoded by four open reading frame (ORF) located at the 30 terminus of the viral genome. These proteins are involved in virion assembly and contribute to the inhibition of the host immune response [54].

1.7.2. Strategies in drug repurposing for COVID-19

In order to use repurpose drugs for treating various diseases, it is necessary to examine the structure of the organism that causes the disease. Researchers are utilizing computational tools and enhancing their understanding of virology and the way COVID-19 clinically, enabling them to explore a wider range of possible targets. Due to the urgent need for effective treatments against COVID-19, repurposing existing drugs presents an ideal approach to make progress in this direction. When repurposing, a wide range of techniques including virtual screening, pharmacophore modeling, and other computational and experimental methods are extensively used [54]. The process of repurposing drugs for COVID-19 involves three key stages prior to their potential development and marketing: (a) identifying potential candidate drugs, (b) evaluating the drug effects in preclinical models, and (c) assessing its efficacy through phase II clinical trials. The most critical of these three steps is the first one, which involves screening and identifying drugs with significant repurposing potential. As such, drug repositioning has two alternative and complementary approaches: an experiment-based approach and a theoretical or *in silico*-based approach, as shown in **Figure 1.3** [69, 70].

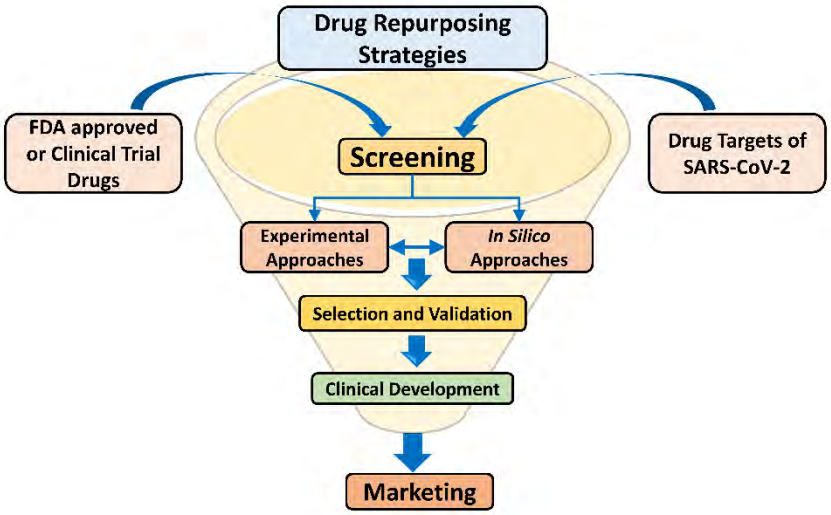

Figure 1.3. Two alternative and complementary drug repurposing approaches for COVID-19: one is an experiment based approach, and the other is a theoretical or *in silico*-based approach.

In contrast, *in silico* repositioning utilizes computational biology and bioinformatics or cheminformatics tools to virtually screen extensive drug/chemical libraries present in public databases. In this method, the identification of molecules with potential biological activity is accomplished through the molecular interaction between drug molecules and protein targets. Though experimental studies are useful for determining therapeutic efficacy, they can be time-consuming and produce very modest results. As a result, computer methods have helped to improve this procedure. They can look closely at how drugs work, predict new ways they might affect the body, and help develop drugs faster and cheaper. This is really important during the COVID-19 pandemic [71].

1.7.3. Drugs repurposed against SARS-CoV-2 drug target

Researchers are studying various groups of drugs to see if they can be useful against COVID-19. These groups include drugs like hydroxychloroquine, remdesivir, and chloroquine. These drugs have been used before to treat other viruses like SARS-CoV and MERS-CoV. A detailed list of these repurposed drugs and their targets for treating SARS-CoV2 is shown below [54].

1.7.3.1. Chloroquine and hydroxychloroquine

Chloroquine, a powerful antimalarial, has gained attention for its potential in treating COVID-19-related pneumonia due to its inflammation-reducing and immune-boosting properties [72-74]. An exciting alternative, hydroxychloroquine, a modified version of chloroquine, shares a similar action but is better tolerated for extended use in treating conditions like rheumatoid arthritis [75, 76]. While both have historically been used to combat malaria and chronic inflammatory diseases, they have also displayed broad antiviral effects against HIV, SARS-CoV, Marburg, Zika, Dengue, and Ebola viruses. The scientific community was intrigued by this fact, prompting them to explore these compounds for

treating COVID-19 [72-74]. Numerous clinical trials were conducted, but unfortunately, the results proved disappointing. These drugs did not show effectiveness in preventing or treating mild to moderate COVID-19 cases, nor did they reduce mortality rates compared to standard care. Worse, severe side effects emerged, including heart issues, liver or kidney damage, retinopathy, and low blood sugar. These concerns, combined with a lack of clear benefits, led the WHO to halt hydroxychloroquine trials for COVID-19 treatment [65, 77].

1.7.3.2. Baricitinib

Baricitinib, an approved small molecule inhibitor of Janus kinase 1 (JAK1) and JAK2, used for rheumatoid arthritis [78, 79], was repurposed using artificial intelligence to combat SARS-CoV-2. Its mechanism of action involves modulation of the cytokine storm and blocking virus entry into host cells [78]. Combined with remdesivir, it established enhanced efficacy in reducing recovery time for COVID-19 patients needing oxygen or ventilation [80]. FDA authorized emergency use with remdesivir in November 2020, later revising it in July 2021 for standalone use. The EMA in April 2021 began evaluating its extension for treating hospitalized COVID-19 patients who need supplemental oxygen [65].

1.7.3.3. Remdesivir

Remdesivir (Veklury TM) was originally developed by Gilead Sciences, Inc. (Foster City, CA, USA) as a potential anti-Hepatitis C virus candidate and was later repurposed and tested against Ebola and Marburg viruses in clinical trials. This prodrug was thought to be a potential promising candidate to be repurposed for COVID-19 treatment since almost the beginning of the pandemic. Phase III clinical trials were carried out in early 2020. However, the efficacy of remdesivir was, in several moments, questionable and surrounded in controversy, due to the lack of evidence of its efficacy in mortality reduction. Several important adverse side effects were observed, including hepatocellular toxicity, nausea, anaemia, kidney injury, hypotension, respiratory failure, and constipation [65, 81]. A multinational cohort study supported by Gilead Sciences showed clinical improvement for 68% of severe COVID-19 patients treated with compassionate use of remdesivir [82].

1.8. Monkeypox virus

Monkeypox virus (MPXV), belonging to the Orthopoxvirus genus of the Poxviridae family [83], is a zoonotic virus initially identified in 1958 during outbreaks of pox-like manifestations in macaque monkeys, subsequently named ‘monkeypox’ [84]. The first human case was documented in 1970 within an intensified smallpox (SPX) surveillance framework in the Democratic Republic of the Congo (DRC) [85]. Clinically, monkeypox closely resembles with traditional Variola (SPX) and Vaccinia viruses [86]. At present, there are no approved medications to address monkeypox (MPX), and while the smallpox vaccine can offer protection, its utilization is restricted due to safety concerns associated with its live viral nature. Consequently, preventing MPX relies on minimizing human contact with infected wildlife and averting the transmission of the virus from person to person [87, 88]. The transmission of monkeypox can occur through animal bites or direct contact with an animal’s

bodily fluids. It can also propagate through respiratory droplets during extended face-to-face interactions, direct exposure to bodily fluids of an infected person, or via contaminated objects carrying viral particles [89].

The symptoms of monkeypox includes fever within the range of 38.5°C to 40.5°C, accompanied by malaise, rash, and headaches. The incubation period for monkeypox spans from 7 to 17 days, with fever subsiding around 3 days after the onset of the rash [90]. The lesions are characterized by swelling, rigidity, and pain. Notably, lymphadenopathy, observed in monkeypox but not in smallpox, has been suggested to induce a more robust immune response compared to smallpox [91]. Depending on the available tests, monkeypox can be identified using viral culture, polymerase chain reaction (PCR), immunohistochemistry, such as enzyme-linked immunosorbent assay (ELISA), or electron microscopy [92]. Erez et al. (2019) investigation of a case revealed viral particles with size of 281 ± 18 nm × 220 ± 17 nm (n = 24) [89].

1.8.1. Drugs repurposed against monkeypox virus target

Discovering small molecules as potential inhibitors against the MPXV for those who are not vaccinated or have inadequate vaccine responses is urgently needed in the absence of specific treatment. The development of effective antiviral options against the MPXV may be sped up via drug repurposing techniques, greatly reducing the expense and time required to create a novel cure [93]. The US FDA has approved the use of tecovirimat, which is one of the potential treatments. Cidofovir or brincidofovir have also been proposed as potential treatments for MPXV, but its use has not been explored outside of laboratory models [94]. The use of other drugs has not been usually explored which indicates an open knowledge gap regarding MPXV treatment [89].

1.9. Human adenovirus

People are always being attacked by different new and old viruses. Some viruses like the SARS-CoV-2 or influenza A virus (infA) have caused large global pandemics. Others, like human adenoviruses (HADV), have mostly caused outbreaks in specific areas (**AV 1**). Adenoviruses (ADV) represent a diverse family of viruses with the capability to infect a wide variety of vertebrates, including humans [95]. Human Adenovirus (HADV) was initially discovered in the 1950s when it was first isolated from adenoid tissues [96, 97]. HADV displays a predictably unpredictable emergence pattern. Researcher have become more interested in dealing with the appearance and reappearance of HADV. This interest is closely linked with a better understanding of HADV epidemiology and its potential for zoonotic transmission. Also, the increasing interest in using HADV for vaccines and treatments demands ongoing research into how infections interact with immune responses. This has recently brought HADV outbreak into the research spotlight [98].

1.9.1. Pediatric Adenovirus or HADV infections in children

HADV is a significant pathogen responsible for causing community-acquired pneumonia (CAP). Pediatric adenovirus pneumonia is most commonly observed in children aged between 6 months and 5 years, comprising approximately 2%–5% of all respiratory illnesses

and 5%–10% of lower respiratory tract infections in children [99, 100]. Unlike other CAPs (Mycoplasma and Staphylococcus aureus pneumonia), early recognition of severe adenovirus pneumonia (SAP) is difficult due to nonspecific symptoms. This is because children have immature immune systems, particularly humoral immunity, which contributes to a high mortality rate and the likelihood of severe cases with long-term consequences [101].

Most HADV infections (species A and D) are mild or asymptomatic [102]. But, HADV types 1–7 accounts for childhood respiratory infections, sometimes, leading to hospitalization [103-105]. Among these, HADV-B3 and -B7 are common and associated with acute respiratory disease. Recent research shows that HADV-B7 is more pathogenic than HADV-B3, exhibiting more cytotoxicity in cell culture, robust replication, and higher inflammatory responses in mice, providing insights into its increased severity in humans [106].

1.9.2. Adenovirus characteristic

The ADV particle has two main parts: a core and an approximately 90 nm icosahedral capsid. The core contains the viral genome divided into early, intermediate, and late regions [107], along with core proteins IVa2, V, VII, Mu, Tp (terminal protein), and antiviral protein [108, 109] (**Figure 1.4**). In contrast, the icosahedral capsid has three major capsid proteins (fiber, penton base, and hexon) and four minor capsid proteins. Hexon has a serotype-specific antigen, and the penton base helps the virus bind to host cell receptors for endocytosis [110]. The fiber has species-specific antigenic determinants, and minor capsid proteins (IIIa, VIII, VI, IX) stabilize the capsid during replication [98, 111]. Inside the host cell nucleus, the viral genome duplicates and transcribes into mRNA, then translates into viral proteins in the cytoplasm. Ultimately, the target cell undergoes lysis, simultaneously leading to the release of newly formed adenovirus progeny [112].

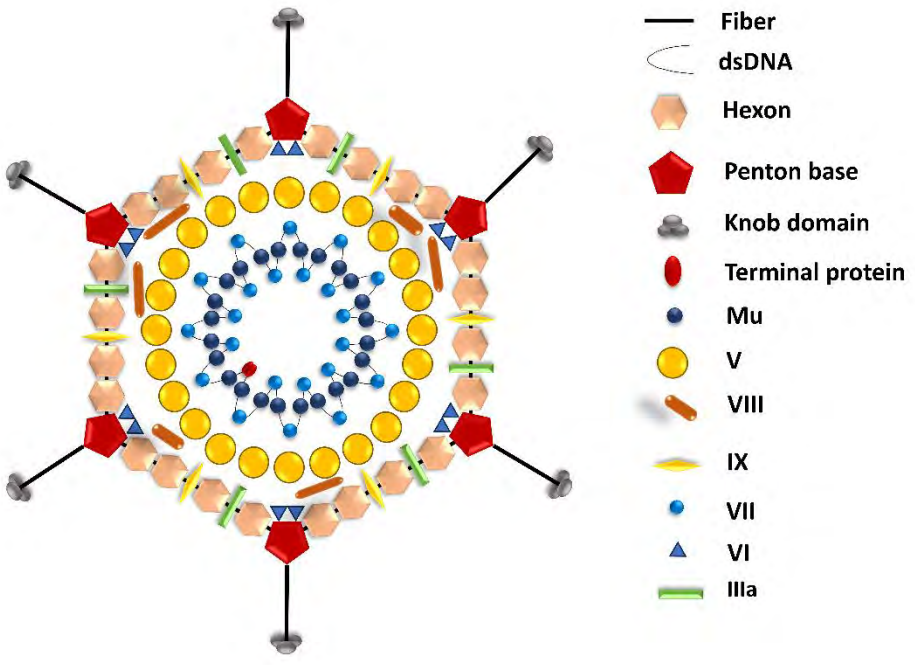

Figure 1.4. The icosahedral structure of adenovirus.

1.9.3. HADV transmission, incubation, and viral shedding

HADV infections normally spread through respiratory droplets, the fecal-oral route, and contact with contaminated surfaces [102]. There have been cases of direct eye inoculation, like an outbreak in an ophthalmology clinic due to contaminated eye drops [113]. The incubation period for HADV-related symptoms varies from 2 to 14 days [114]. After recovery, virus shedding can persist in the gastrointestinal and respiratory tracts, especially in immunocompromised individuals [115]. A recent study found that in immunocompetent patients with HADV-B55-positive respiratory infections, viral shedding continued for up to 52 days after fever onset [116]. In hospitalized children with severe HADV pneumonia, shedding lasted an average of 96.9 days for HADV-B7 and 51.4 days for HADV-B3 infections [117]. This lengthy shedding of HADV is notably longer than that of other respiratory viruses in healthy individuals, such as influenza (up to 18 days), respiratory syncytial virus (average 4 days), and rhinovirus (average 11 days) [116]. These findings have practical implications for infection control and surveillance strategies during HADV outbreaks, particularly in pediatric healthcare settings [98].

1.9.4. Clinical features of HADV infection

a) **Respiratory tract involvement:** HADV causes 5–10% of childhood infections and 1–7% of adult respiratory infections [118]. Symptoms include fever, cough, tonsillitis, and pharyngitis, with gastrointestinal (GI) symptoms in children [119]. In immunocompetent individuals, recovery usually takes about two weeks, resulting in type-specific immunity. However, infants, previously healthy adults or children can develop pneumonia [120]. Immunocompromised individuals may experience severe respiratory issues and transmission (10–30%), with a severe HADV pneumonia mortality rate up to 50% [121].

b) **Ocular involvement:** Ocular HADV infections include non-specific conjunctivitis, pharyngeal conjunctival fever, and epidemic keratoconjunctivitis (EKC) [122]. HADV-D is the primary species causing acute ocular infections and is linked to EKC outbreaks, which can lead to vision problems. HADV-B is mainly associated with conjunctivitis [123].

c) **GI infection:** HADV-F, specifically HADV-40 and HADV-41, mainly cause GI infections with symptoms like colitis, diarrhea, and hepatitis [124]. Though respiratory infections are common, HADV can also give GI symptoms and persists in GI tract of children [125].

d) **Urinary tract infection:** HADV can induce urinary tract infections (UTI) in patients having hematopoietic stem cell transplantation (HSCT) and solid organ transplantation (SOT), with symptoms like hematuria, dysuria, renal transplant issues, and hemorrhagic cystitis (HC) [126].

e) **Liver infection:** In a case report by Kawashima et al., elevated γ-glutamyltransferase (γ-GTP) levels could potentially signal HADV hepatitis. Subsequent studies by Onda et al. suggest that a γ-GTP increase more than two weeks before hepatitis onset could aid early diagnosis. ADV hepatitis is more common in children, the elderly, and immunocompromised adults than other hepatitis types, and it has a more acute onset [127].

1.9.5. Treatment of HADV infection

In medicinal practice, the main treatment usually involves general care, antiviral drugs, and respiratory support. Doctors need to carefully consider the symptoms in children and lab test results before deciding on the right treatment plan.

1.9.5.1. General treatment

The common treatments involve lowering the fever, isolation, supporting the functions of organs like heart, liver, and kidneys, and closely watching how these organs are doing. It is very important to isolate children with adenovirus pneumonia early to stop it from spreading to others.

1.9.5.2. Antiviral drugs

Whether to complement antiviral drugs remains controversial, but the European Respiratory Society (ERS) recommends their use in critically ill patients [128]. Generally used antiviral drugs include cidofovir, ribavirin, and acyclovir. Cidofovir, a nucleoside and phosphonate analogue, inhibits viral DNA polymerase and demonstrates broad antiviral activity *in vitro* against various viruses. If applied in the early stage of disease, clinical outcomes can be positively altered [129]. Moreover, its antiviral activity *in vivo* has been documented, but its nephrotoxicity and toxicity to bone marrow have limited its widespread use [114, 130]. Nevertheless, since heparin increases the possibility of bleeding, which also restricts its application in adenovirus pneumonia, more clinical trials should be carried out carefully to verify its anti-viral availability and security [112].

1.10. Quantitative structure-activity relationship (QSAR)

The primary goal of QSAR modeling is to establish a mathematical relationship between experimental activity (or any other attribute of interest) and structural characteristics calculated for a group of molecules. The approach depends on the ability to quantitatively express a molecule's structure and then establish a relationship between the quantitative values representing each structure and the experimental activity value [131, 132]. It is important to note that experimental activity values are generally referred to as the response or dependent variable. In contrast, quantitative representations of compounds are referred to as descriptors or independent variables.

The dependent variable falls into three main groups: biological activity (like anti-cholinergic or anti-bacterial effects), physical properties (such as melting point, lipophilicity, aqueous solubility, etc.), and toxicity (including both non-systemic effects like ecotoxicity and systemic effects like hepatotoxicity, cardiac toxicity, etc.). Based on these three types of variables, the activity relationship is known as QSAR, QSPR (quantitative structure-property relationship), and QSTR (quantitative structure-toxicity relationship). As the response variable is mainly determined experimentally, taking adequate precautions to reduce experimental mistakes is essential. One must verify the sources' reliability before collecting response data from external sources. Additionally, it is essential to remember that the stated response values must be determined using the same experimental technique when combining the data from various sources, such as databases and literature.

The descriptors or independent variables are mathematical representations of chemical structure. Descriptors give a solid basis for investigating molecule's structural and chemical diversity and play an essential part in predictive modeling analysis. The descriptors can be generated using both experimental and theoretical studies. Taking precautions to prevent potential mistakes while carrying out studies is essential because experimental descriptors are obtained through empirical analysis of the compounds. These descriptors mainly indicate physical parameters such as the octanol-water partition coefficient, melting, boiling point, pKa values, rate of reaction, molar refractivity, and so on. Theoretical descriptors are chemical features that can be computed using the most accurate mathematical algorithm. The introduction of theoretical descriptors provides the QSAR study with a dimensional perspective. Usually, parameters generated from higher dimensional (>2D) chemical concepts need an appropriate energy minimization approach. Currently, various theoretical descriptors are calculated utilizing multiple available resources [133-135].

Additionally, certain requirements must be fulfilled before the actual construction of QSAR models, including the following: (a) a sufficient number of compounds for adequate statistical degree of freedom; (b) a varied range of calculated dependent variables for regression models; or (c) a suitable distribution of compounds in each class (i.e., active and inactive) for classifying models; and (d) the use of the same experimental protocol to determine the dependent variables for all the molecules before modeling.

1.10.1. Brief historical background of QSAR

The concept of relating chemical characteristics to their behavioral traits can be traced back to Mendeleev's theory, where he utilized the 'rule of eight' to identify similar chemicals during the late 19th century. In 1863, Cros explored the connection between the toxicity of primary aliphatic alcohols and their aqueous solubility. Additionally, a significant step was taken by Crum-Brown and Fraser in 1868, as they introduced the first mathematical idea within QSAR. They proposed a potential mathematical correlation between the biological activity of various alkaloids and their molecular structure, exemplified by the following equation (**equation 1.1**):

$$\phi = f(C) \qquad \textbf{(1.1)}$$

where, the relationship between the chemical constitution C of a substance and its physiological effect ϕ within a biological system is illustrated as a function (f).

In 1874, Körner postulated that the variation in color of di-substituted benzenes is linked to their distinct chemical arrangement. Subsequently, in 1884, Mills discovered that the melting and boiling points of compounds exhibit a connection with their chemical structure. Around a decade later, in 1893, Richets documented an inverse correlation between the cytotoxicity and aqueous solubility of various organic compounds. Independently in 1899, Meyer and Overton proposed that the narcotic effect of a set of organic compounds is correlated to their partition coefficients (olive oil/water). This insight gained further support from Traube's findings (1904), which linked surface tension with chemical narcosis. Later, in 1912, Seidell determined solubility and partition coefficient as the defining factors of narcotic behaviour.

Then, in 1935, Hammett introduced a ground-breaking concept by establishing a relationship between the chemical reactivity and structural features of benzene derivatives, utilizing terms like rate constant and electronic constant (**equation 1.2 and 1.3**).

$$\log\left(\frac{K_X}{K_H}\right) = \rho\sigma_X \quad \textbf{(1.2)}$$

$$\log\left(\frac{k_X}{k_H}\right) = \rho\sigma_X \quad \textbf{(1.3)}$$

where, the symbols k_X and k_H represent the rate constants, while K_X and K_H denote the equilibrium constants for benzenes with substitutions and without substitutions. The Hammett constant σ_X is utilized to convey electronic information, and ρ stands for constant.

Hammett parameter 'σ' was initially expressed in terms of ionization constant, which enabled its correlation with linear free energy principles, as expressed by the Gibbs equation ($\Delta G^o = -RTlnK$). Consequently, Hammett's equation marks the inception of the linear free energy relationship (LFER) concept. Employing this concept, Taft (in 1952) introduced a steric aspect that allowed for the separate assessment of polar, steric, and resonance effects. This was achieved by investigating the acid and base catalyzed hydrolysis of aliphatic esters. Corwin Hansch is widely known as 'Father of QSAR' due to his remarkable impact on this field. In 1962, Hansch and Muir conducted a significant investigation into the relationships between molecular structure and biological activity within the context of plant growth regulators. This study emphasized the reliance of these relationships on Hammett constants and the degree of hydrophobicity. Leveraging the octanol/water partition system, they introduced a new hydrophobic parameter 'π'. This parameter helped to measure the partition coefficient of the substituent, using the contribution of entire molecule (**equation 1.4)**.

$$\pi_X = \log P_X - \log P_H \quad \textbf{(1.4)}$$

where, P_X and P_H are the partition coefficient of the derivative and parent molecule, respectively and π_X provides the substituent hydrophobicity. Fujita and Hansch followed this by combining Hammett's electronic constant (σ) and the hydrophobicity measure (π) into a single equation. Fujita had combined various free energy variables like π, logP, σ into a single equation for representing biological activity. This involved a logarithmic transformation of the concentration term to maintain the LFER technique (**equation 1.5)**.

$$\log\left(\frac{1}{C}\right) = k_1\pi + k_2\sigma + k_3 \quad \textbf{(1.5)}$$

This model was further modified by Hansch who gave a parabolic relationship between log(1/C) and logP (**equation 1.6 and 1.7**).

$$\log\left(\frac{1}{C}\right) = k_1(\log P) - k_2(\log P)^2 + k_3\sigma + k_4 \quad \textbf{(1.6)}$$

$$\log\left(\frac{1}{C}\right) = k_1\pi - k_2\,\pi^2 + k_3\sigma + k_4 \quad \textbf{(1.7)}$$

In 1976, Kubinyi introduced an enhanced bilinear model (**equation 1.8**) by modifying Hansch's parabolic correlation. The coefficients a, b, and c can be determined through

multiple linear regression analysis, while the non-linear term β needs derivation via a stepwise iterative method.

$$\log\left(\frac{1}{C}\right) = a\log P - b\log(\beta - P + 1) + c \quad \textbf{(1.8)}$$

Free-Wilson strategy in 1964 established a mathematical relationship between the biological activity of a set of closely related chemicals and the combined impact of the parent structure along with the influence of individual structural substituents (**equation 1.9**).

$$BA = \sum a_i x_i + \mu \quad \textbf{(1.9)}$$

where, BA represents the assumed biological activity, covering both the average contribution of the parent structure (μ) and the individual contributions of each substituent (a_i). The variable x_i is a Boolean indicator, signifies the presence (one) or absence (zero) of a particular structural fragment. Further, in 1971, Fujita and Ban introduced a log transformed of biological activity and offering a modified version (**equation 1.10**).

$$\log BA = \sum G_i X_i + \mu \quad \textbf{(1.10)}$$

Here, μ represents the parent moiety contribution, G_i stands for chemical fragments contribution, and X_i indicates whether G_i is present or not.

Initially, a few variables were used in the development of QSAR models. However, this approach was quickly expanded in order to include other parameters, such as quantum chemical and steric factors and thus several topological, connectivity, electrotopological state, and other variables were added. Topliss in 1972 said that incorporating a large number of variables into a model significantly increases the risk of chance correlations. Several researchers ignored such warnings, and as a result, the literature became overloaded with illogical models. Therefore, in 1973, C. Hansch and Unger established a set of five criteria that defined a good QSAR model.

The five guidelines comprised the following: i) selecting independent variables from a diverse range of possibilities; ii) validating the selected independent variables through statistical methods; iii) obeying to the principle of parsimony, favoring the simplest model when all other factors are equal; iv) to minimize the risk of coincidental correlations, it is recommended to ensure the presence of at least five to six data points per variable; v) trying to develop a qualitative model that holds physicochemical or biochemical relevance, and so on.

1.10.2. Objective of QSAR analysis

The primary objective of QSAR studies are as follows:

a) to estimate the biological activity for newly designed molecules, b) to provide direction for designing and developing new lead molecules with enhanced activity, c) to examine large databases involving various molecules, thereby reducing the substantial cost and time for initial experimental studies, d) to follow the basic guideline of REACH (Registration, Evaluation, and Authorization of Chemicals) [136] for animal safety by limiting the animals

used for primary screening, e) to fill in the data gaps by anticipating the behavior of compounds still awaiting experimental investigation using the limited experimental data (currently available). To attain the goals mentioned earlier, it is highly recommended for QSAR researchers to concentrate on executing correct procedures and to avoid mistakes, thereby enhancing the quality of QSAR modeling.

1.10.3. Classification of QSAR

QSAR methods are classified into various categories, determined by how descriptor values are calculated, as illustrated in **Figure 1.5**.

i) Zero-dimensional (0D) QSAR models are established using descriptors extracted from molecular formula data, such as molecular weight, number of atoms, and the sum of atomic properties.
ii) One-dimensional (1D) QSAR models link activity with global molecular properties like pKa, solubility, logP, and functional groups.
iii) Two-dimensional (2D) QSAR models analyze the relationship between activity and structural patterns, like connectivity indices and Wiener index.
iv) Three-dimensional (3D) QSAR considers molecular spatial orientation and associates activity with non-covalent interaction fields (steric and electrostatic fields) surrounding molecules. This includes methods like Comparative Molecular Field Analysis (COMFA) and Comparative Molecular Similarity Indices Analysis (COMSIA).
v) Four-dimensional (4D) QSAR incorporates an ensemble of ligand configurations in 3D-QSAR, representing molecules in distinct conformations, stereoisomers, orientations, tautomers, or protonation states.
vi) In five-dimensional (5D) QSAR, various induced-fit models are explicitly represented in 4D-QSAR.
vii) Six-dimensional (6D) QSAR further integrates different solvation models within 5D-QSAR.
viii) Finally, seven-dimensional (7D) QSAR comprises actual receptor or target-based receptor model data.

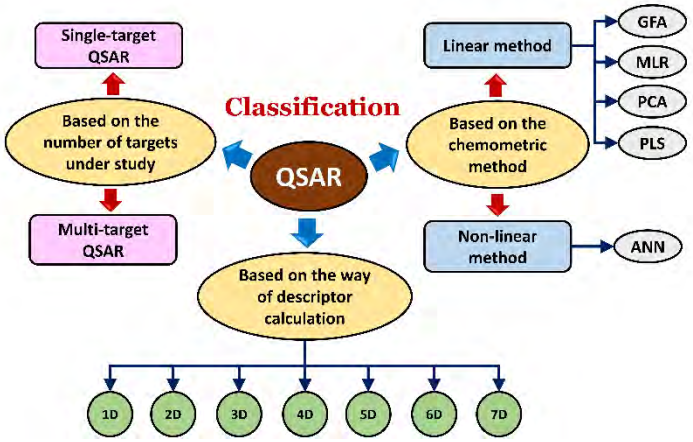

Figure 1.5. Classification of QSAR approaches.

Depending on how they work, QSAR methods are sorted into two categories: linear and non-linear. Multiple linear regression (MLR), principal component analysis (PCA), partial least-squares (PLS), and genetic function approximation (GFA) are fallen into linear methods. Although, non-linear methods also include both PLS and GFA techniques. Further advancement in the chemometric field has also generated several methods, which include non-linear regression and artificial neural networks (ANN), for building predictive models.

1.10.4. QSAR and the OECD guidelines

The Organisation for Economic Co-operation and Development (OECD) supports the use of QSAR technique as an effective tool for evaluating the chemicals safety. The members (countries) of this organisation have put together a set of rules that can be used for regulatory purposes. The 37th OECD Joint Meeting accepted the OECD Principles for validating QSAR Models for regulatory purposes in 2004. The OECD released the "Guidance document on the validation of (quantitative) structure-activity relationships [(Q)SAR] models" in 2007 to guide researchers in how to assess exact QSAR models under OECD standards. The OECD QSAR project aims to enhance the use of QSAR modeling by governments and industry to improve chemical safety assessment [137]. It comprises the QSAR toolbox, the principles for validating developed models, and a guidance document. The OECD has chosen the following guidelines to make using QSAR modeling easier for regulatory purposes.

i) Principle 1: A defined endpoint
ii) Principle 2: An unambiguous algorithm
iii) Principle 3: A defined domain of applicability
iv) Principle 4: Appropriate measures of goodness-of-fit, robustness and predictivity
v) Principle 5: A mechanistic interpretation, if possible

i) OECD principle 1: A defined endpoint

The idea behind this is to ensure that the endpoint predicted by the model is as transparent as possible since the endpoint can be expected by different methods and under other conditions. The development of QSAR models is expected to be developed using homogeneous datasets with information collected from a single protocol's generated response. While compiling biological activity data from various sources, care should be taken to thoroughly verify the experimental procedures, data quality, concentration unit, etc. It should also be considered that all of the used compounds have similar mechanisms of action. The term 'endpoint' may have different meanings depending on the scientific or regulatory background. A specific biological impact can designate an endpoint, but it can also be known through testing procedures that follow regulatory guidelines (such as the OECD Test Guidelines). Overall, an appropriate endpoint can be obtained by combining various test procedures to achieve a specific response.

ii) OECD principle 2: An unambiguous algorithm

This principle aims to guarantee clarity in outlining the model algorithm. This encompasses the methodologies applied during data pre-processing, the separation of the dataset, and the

feature selection process. As a result, this guideline aims to enhance transparency in constructing the model, making it replicable by others and elucidating its capacity to achieve the outcomes. Regarding determining regulatory endpoints, the OECD guidelines prescribe a clear framework for utilizing QSAR models tailored to specific classes of chemicals, general QSAR models for common responses, and the calculation and implementation of chemical-specific descriptors using an appropriate modeling algorithm. The OECD outlines the necessary components for maintenance methodological transparency while developing QSAR models.

a) A collection of chemical compounds accompanied by their respective endpoint and descriptor data.
b) Transparent presentation of the process of computing descriptors and their measurement methodology.
c) Explanation of the training and test sets, accompanied by a well-founded rationale for excluding any outlier observations, if applicable.
d) Elaboration on the mathematical models that illustrate the correlation between endpoints and descriptors, including the extracted chemical insights.
e) Inclusion of statistical parameters to assess the prediction's credibility.

The OECD-recommended linear modeling algorithms commonly used encompass techniques like univariate linear regression (ULR), MLR, ordinary least squares (OLS), PLS, PCA, and principal component regression (PCR), among others. Furthermore, the OECD guideline proposes incorporating a priori feature selection using mechanistic principles or evolutionary methods such as the genetic algorithm (GA) and techniques like PCA or factor analysis (FA). Additionally, the guideline acknowledges the applicability of non-linear approaches, such as ANN, fuzzy clustering, and regression techniques, contingent upon the specific modeling objectives.

iii) OECD principle 3: A defined domain of applicability

The necessity to establish an applicability domain underscores the reality that QSARs are reductionist models inherently bound by constraints of the specific chemical structures, physicochemical characteristics, and action mechanisms for which dependable predictions can be generated. When a QSAR model is constructed using a particular set of chemical compounds, it occupies a distinct theoretical space and is acknowledged as a reliable predictor within that domain. Establishing the applicability domain of a model using the molecules in the training set becomes imperative to ascertain the credibility of predictions for molecules within the test set [138]. The applicability domain of a model relies on three primary features: a) structural information, b) physicochemical characteristics, and c) response patterns. Due to the potential involvement of multiple mechanistic foundations in various regulatory endpoints, QSAR models could be personalized for specific chemical classes that operate through the same mechanism of action. Sometimes, a single overarching QSAR model might struggle to differentiate between distinct chemical classes, thus failing to provide accurate estimates for a particular chemical class.

The OECD guidelines suggest either a) the development of multiple predictive models for the same endpoint on various domains of applicability together to give a global estimation or b) the use of statistical methods giving global modeling features across multiple mechanisms of actions concerning the same endpoint. By comparing the domain of chemicals concerning each designated regulatory endpoint, one can detect data gaps by appropriately using the application domain of existing QSAR models. The following points as identified by the OECD should be followed with respect to applicability domain of a QSAR model. Concerning the applicability domain of a QSAR model, it is recommended by the OECD to follow the subsequent guidelines:

a) Establishing confidence limits that describe the extent of an applicability domain.
b) Identifying the applicability domain for methods like structural alerts and fragment-based QSAR techniques.
c) Evaluating the effectiveness, constraints, and applicability of methods employed to determine an applicability domain.
d) Deploy integrated tools for applicability domain determination and other statistical operations.

iv) OECD principle 4: Appropriate measures of goodness-of-fit, robustness and predictivity

This principle emphasizes the necessity to provide two kinds of information:

a) The internal performance of the model through goodness-of-fit and robustness, using training set.
b) The predictability of that model, using a suitable test set.

The aim is to balance overfitting and underfitting the model's predictions under varying conditions. Statistical validation also serves to distinguish models that are based on chance correlations. A crucial strategy for assessing the internal stability and external predictability of a developed QSAR model involves dividing the dataset into training and test sets. In simpler words, these validation procedures are employed to confirm how closely a model's prediction matches experimental observations.

v) OECD principle 5: A mechanistic interpretation, if possible

The fifth principle from OECD focuses on "mechanistic interpretation" when possible. It can be challenging to explain the mechanics of a QSAR, but the principle suggests mentioning any available information to help future research in that area. Knowing how chemicals work in a process can help create similar compounds. QSAR modeling is a step-by-step process involving making guesses, testing them, and analyzing data. This helps improve the training set by looking at the chemicals and their traits. Sometimes, the sequence of guesses may not reveal all the mechanics for a specific goal. But this doesn't mean it failed; it might be in the early stages without all the mechanics. Sometimes, more information can be added from previous research. Overall, the fifth principle encourages sharing any mechanistic explanations found to enhance understanding of a particular goal.

1.10.5. Applications of QSAR

QSAR models have developed as essential tools in understanding and predicting the biological activity of chemical compounds. These models are extensively employed in various scientific disciplines, including drug discovery, materials science, and predictive toxicology. In drug discovery, QSAR facilitates the rapid screening of many chemical entities, aiding the identification of potential lead compounds against specific target enzymes. Moreover, QSAR plays a vital role in lead optimization and the prediction of pharmacokinetic profiles of pharmaceutical agents [139]. In addition, the QSAR utility extends to the agricultural sciences, which proves advantageous for evaluating the toxicity potential of chemicals, such as fungicides [140]. Owing to its broad applicability and validated reliability, QSAR has valid acceptance as an alternative approach to traditional animal testing and serves regulatory purposes.

1.11. Histone lysine methyl transferase

Histone methyltransferase (HMT) acts as an epigenetic 'writer', affecting gene transcription, DNA replication, and repair. It is vital for chromatin stability and gene expression by methylating lysine side chains on histones H3 and H4. HMT fall into two groups: histone lysine methyltransferases (HKMT) and protein arginine methyltransferases, based on their targeted residues. The majority of biological methylation reactions are facilitated by methyltransferases using S-adenosyl-L-methionine (SAM) as the methyl contributor. The human genome encodes around 100 potential HKMT candidates [141]. HKMTs form a group of enzymes responsible for transferring methyl groups from the cofactor SAM to lysine residues found on both histone and nonhistone substrates [142, 143]. The level of methylation occurring at the ε-nitrogen of a lysine side chain determines the creation of mono-, di-, and tri-methylated groups within the lysine residue (referred to as HKme1, HKme2, and HKme3). Among these, trimethylated variants, particularly trimers with trimethylation, are widely recognized as the most functionally impactful indicators [144].

In the year 2000, the initial lysine methyltransferase to be biochemically characterized was documented [145]. HKMTs have been linked closely to the progression of numerous cancers, holding immense potential as therapeutic targets [143]. In the past decade, substantial developments have been taken towards designing drugs aimed at HKMTs that participate in histone methylation and epigenetic control. However, these drugs are in an early developmental stage when compared to other therapeutic agents. A significant milestone was reached in 2020 when FDA granted approval for the first HKMT inhibitor, tazemetostat, for treating epithelioid sarcoma (ES) and subsequently for follicular lymphoma [146].

1.11.1. Biological function of HKMT in cancers

In the past two decades, different research groups have looked into how HKMT influences tumor characteristics. They found a complex connection between HKMT, which plays a role in epigenetic regulation, and the development of tumors and their spread. This has led to interest in targeting HKMT, like enhancer of zeste homolog 2 (EZH2), as a potential treatment [147, 148]. Understanding EZH2 better could help create a strong approach for managing cancer [149].

1.11.1.1.EZH2 in cancer

EZH2 operates as a catalytic constituent within the polycomb repressive complex 2 (PRC2), where its HKMT activity catalyzes the trimethylation of lysine at position 27 on histone H3 [150]. The primary role of EZH2 as an HKMT is facilitated through its su(var) 3-9, enhancer-of-zeste, trithorax (SET) domain. EZH2 can exert transcriptional repression or coactivation in a manner that depends on PRC2 or functions independently of it [151, 152]. For instance, EZH2 with PRC2-independent functions are implicated in interactions with non-PRC2 partners, like the androgen receptor and nuclear factor-kB, in the contexts of prostate and breast cancer [153, 154].

To date, multiple experiments have shown that EZH2 plays a vital role in controlling various gene expressions through epigenetic mechanisms across a range of human cancers. Increased levels of EZH2 have been observed in conditions such as lymphoma, lung cancer, prostate cancer, breast cancer, melanoma, retinoblastoma, and glioblastoma [149]. Notably, EZH2 has the ability to temporarily turn off genes related to B-cell differentiation, leading them away from the germinal centers within lymph nodes. In instances where the EZH2 gene is mutated, B cells remain trapped in the germinal centers, eventually transforming into tumor-forming cells, as seen in follicular lymphoma [155]. Jin et al. discovered that EZH2 is highly expressed in uveal melanoma (UM), where it enhances the proportion and self-renewal of cancer stem cells (CSC) using the microRNA (miR)-29c-disheveled segment polarity protein 2-b-catenin signaling pathway [156]. Additionally, it promotes migration and invasion through the Rho-GDP dissociation inhibitor-gamma-Rac1 axis. Their findings suggest that EZH2 could be a promising target for therapeutic intervention in UM [157]. Khan et al. identified high EZH2 expression in retinoblastoma (RB) tissues and cell lines [158]. Smith et al. found that c-Src/mechanistic target of rapamycin complex 1/PRC2 axis, plays an essential role in the development of mammary tumors driven by human epidermal growth factor receptor 2 (ErbB2) [159]. Additionally, Yang et al. revealed that E2F transcription factor 7 promotes cell proliferation, metastasis, and tumorigenesis in glioblastoma by regulating the EZH2-mediated phosphatase and tensin homolog (PTEN)/protein kinase B (AKT)/mechanistic target of rapamycin pathway. However, it is important to note that prolonged inhibition of EZH2 in glioblastoma might paradoxically encourage tumorigenesis, thus highlighting the dual nature of EZH2 role in tumorigenesis [160].

EZH2 is also implicated in targeted drug resistance. In non-small cell lung cancer (NSCLC), EZH2 triggers acquired resistance to gefitinib by activating the phosphoinositide 3-kinase (PI3K)/AKT [161]. Furthermore, Quan et al. found a negative association between EZH2 and resistance to epidermal growth factor receptor-tyrosine kinase inhibitors (EGFR-TKI) [162]. Sun et al. confirmed that EZH2 controls cisplatin resistance in ovarian cancer through the avian myelocytomatosis virus oncogene cellular homolog (c-Myc)-miR-137-EZH2 axis [163].

1.11.2. Drugs targeting HKMT

Until now, several inhibitors have displayed promising results against a variety of cancers, with a prominent application on targeting EZH2. These epigenetic drugs that counteract EZH2 are presented in **Figure 1.6**.

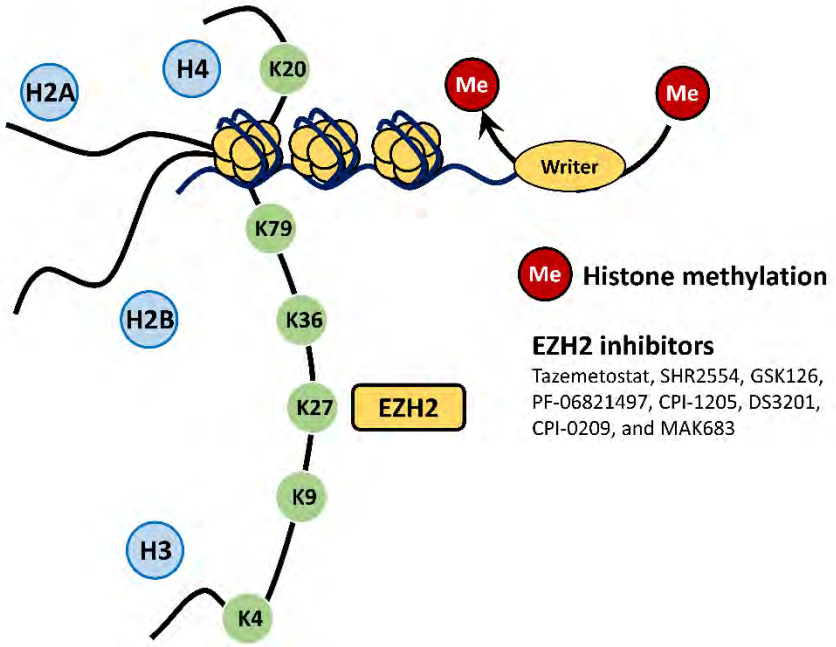

Figure 1.6. The current situation of clinical trials of EZH2 inhibitors.

1.11.3. Targeting EZH2 in cancer

Inhibitors that target EZH2 through small molecules can be categorized into three main groups based on their mechanisms: (1) The most extensive category comprises small molecules that inhibit EZH2 enzymatic activity. Notably, on January 23, 2020, tazemetostat (EPZ-6438), an orally available inhibitor of EZH2, received approval from the US FDA for treating locally advanced or metastatic Ewing sarcoma [164, 165]. Moreover, on June 18, 2020, the US FDA granted faster approval for tazemetostat use in treating follicular lymphoma [166]. Additionally, from July 29, 2016, to June 2, 2017, a multicenter, open-label, phase II study (NCT02860286) enrolled 74 patients with malignant pleural mesothelioma, and the findings have been documented [167]. Tazemetostat is also being examined in other hematological malignancies driven by specific molecular factors, lymphomas, prostate cancer, and various solid tumors (NCT02601950, NCT04179864, and NCT04204941) [168]. There are several other inhibitors sensitive to SAM with therapeutic potential, such as GSK126, CPI-1205, PF-06821497, SHR2554, and valemetostat [149]. A clinical trial involving GSK126 was discontinued due to reaching the maximum dose, and the treatment schedule showed insufficient evidence of clinical effectiveness to warrant further investigation (NCT02082977). At present, CPI-0209 is in the phase I/II clinical trials involving patients with advanced solid tumors and lymphomas (NCT04104776). Furthermore, other SAM-competitive inhibitors of EZH2, including PF-06821497, SHR2554, and CPI-1205, are currently undergoing clinical trials (NCT03460977, NCT03741712, and NCT02395601). (2) MAK683 is a small-molecule inhibitor, designed to block embryonic ectoderm development (EED) binding to H3K27. This action consequently prevents EZH2 activation. Currently, MAK683 is undergoing phase I/II trials for treating large B-cell lymphoma (DLBCL) (NCT02900651) [169]. (3) The last category presents a promising novel approach for inducing the degradation of EZH2. As noted earlier, EZH2 non-PRC2 activities involve coactivation of vital transcription factors like androgen receptor (AR), different from its role

as a methyltransferase. Therefore, the majority of existing EZH2 inhibitors, designed specially to target the catalytic function of EZH2, might not effectively counteract its noncanonical activities [149].

In 2018, a ground-breaking discovery introduced the EZH2-selective degrader (MS1943) to address EZH2 noncatalytic roles. Additionally, in 2021, the innovative EZH2-targeting proteolysis-targeting chimera (PROTAC) degrader MS177 was developed. Especially, MS177 revealed stronger anti-cancer properties against leukemia compared to typical EZH2 inhibitors [170].

1.12. Ghrelin O-acyltransferase

Ghrelin is a special hormone made up of 28 amino acids. It is the only hormone responsible for making humans feel hungry. It was found in 1999 while the researchers were looking for the natural ligands that turns on a receptor called growth hormone (GH) secretagogue receptor (GHS-R1a). When this receptor is turned on by ghrelin, it tells the body to release GH [171]. When compared to naturally occurring ghrelin from rat stomach extracts, synthetic ghrelin showed a shift in retention time, indicating a hydrophobic modification. This modification, identified as octanoylation of serine-3, is essential for ghrelin to activate GH secretion. Ghrelin, originally found in the stomach, has also been found in other tissues, contributing to its various physiological effects [172]. Ghrelin signaling affects growth hormone release, appetite stimulation, and energy balance [173]. Activation of GHS-R1a leads to the release of appetite-stimulating peptides like neuropeptide Y (NPY) and Agouti-related peptide (AgRP), increasing food intake [174]. Mice lacking crucial ghrelin enzymes experience reduced GH release and can develop severe hypoglycemia during starvation [175]. Ghrelin signaling also impacts energy balance in rodents lacking NPY and GH [176].

In 2008, two independent research teams identified ghrelin O-acyltransferase (GOAT) as the enzyme responsible for attaching a lipid to ghrelin [177, 178]. GOAT belongs to the membrane-bound O-acyltransferase (MBOAT) family and is primarily located in the stomach, with lesser amounts in the pancreas and gastrointestinal tract. GOAT adds an octanoyl group to ghrelin, a modification which is distinct and not found in other known proteins (**Figure 1.7**). It is also found in the hypothalamus and pituitary, regions where ghrelin is also present [179-181]. Ghrelin undergoes multiple processing steps before becoming a mature peptide hormone. Initially, it begins as a 117-amino acid preproghrelin, which is then transformed into proghrelin by signal peptidase cleavage [182]. Subsequently, GOAT adds an octanoyl group to proghrelin [177, 178], and finally, the C-terminal part of proghrelin is cleaved to yield the mature 28-amino acid hormone, which is secreted [182].

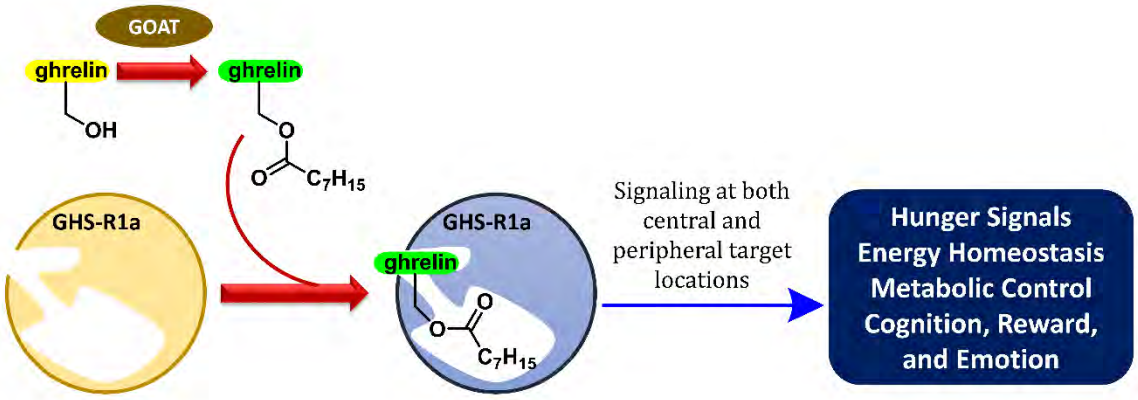

Figure 1.7. The role of GOAT in ghrelin signaling.

Understanding the structure of GOAT has been difficult because it is a type of protein which is part of cell membranes. The membrane topologies of GOAT within the cell membrane has been identified through specific experiments known as selective permeabilization experiments [183]. These studies show that MBOAT family enzymes, including GOAT with its eleven transmembrane domains, have complex structures within cell membranes [184]. The computational model of GOAT structure has revealed insights into how it acylates transmembrane proteins and lipids, guiding the design of inhibitors targeting GOAT [172].

1.12.1. Small molecule GOAT inhibitors

Inhibiting the acylation process of ghrelin by GOAT offers a promising approach to control ghrelin-related functions, as only acylated ghrelin activates signaling through GHS-R1a [185]. Ghrelin is exclusively known substrate for GOAT in the human proteome, making this system unique. Connecting this uniqueness, GOAT inhibition could potentially provide effective treatments for conditions influenced by ghrelin signaling, like diabetes and obesity, with a lower risk of side effects [186].

Initially, GOAT inhibitors mainly focused on peptide-based mimetics of substrates or products. However, these inhibitors often lacked potency in cell-based assays or not tested. Thus, the focus has shifted towards exploring small-molecule GOAT inhibitors, which hold better promise for therapeutic development. Recent improvements in this area have encouraged researchers to use GOAT as a useful target for ghrelin related health problems [187].

Pharmaceutical companies have taken an active role in searching for small molecule inhibitors of GOAT. Researchers at Eli Lilly published their findings on a series of aminopyrimidines as GOAT inhibitors (**Figure 1.8**). They identified the furanopyrimidine derivatives (**1a**) through a screening effort which utilized an ELISA assay to assess GOAT acylation activity [172, 188]. Another study eventually led to the identification of **LY3073084** (also known as GLWL01), which displayed excellent pharmacokinetic/pharmacodynamic properties and revealed good renal clearance against GOAT [189]. Currently, this compound is undergoing clinical evaluation for various disorders [172].

Takeda Pharmaceutical researchers have revealed different GOAT inhibitors characterized by distinct chemical structures. These inhibitors were constructed around an aromatic or heteroaromatic group, connected to a carboxylic acid via benzothiophene and benzofuran linkers. The development process involved screening approximately 500,000 small molecules using a homogeneous time-resolved fluorescence (HTRF) assay, resulting in the identification of 301 compounds. Subsequent evaluations, including ELISA assays to eliminate false positives, dose-response assessments, and substrate competition tests, identified 2-thiobenzoxazole (compound **1b**) as a lead compound, which was then optimized into benzothiophene (compound **1c**). Similar to the inhibitors reported by the Lilly group, both of these compounds competitively targeted octanoyl-CoA and displayed noncompetitive behavior against des-acyl ghrelin. *In vivo* studies with these compounds were also conducted [172, 190].

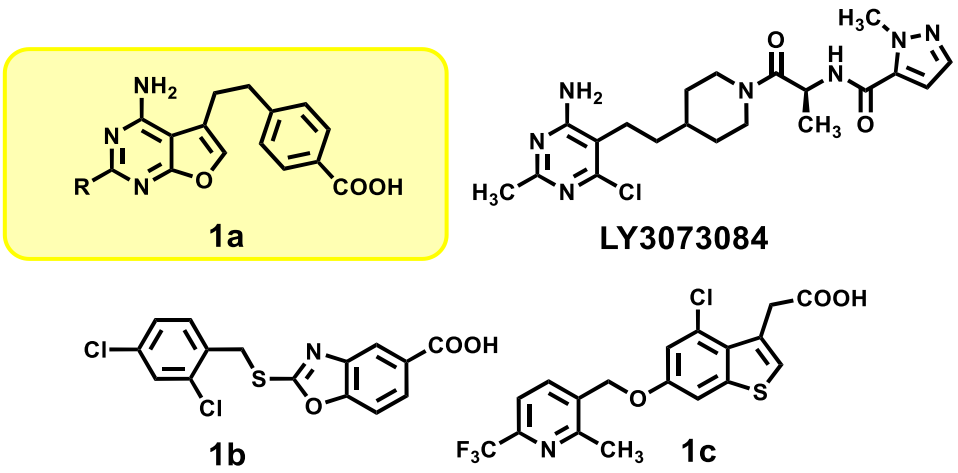

Figure 1.8. Heterocyclic small molecule from Eli Lilly and Takeda.

Other research groups have also reported their work on screening and improving GOAT inhibitors (**Figure 1.9**). GlaxoSmithKline group disclosed inhibitors similar to Takeda, highlighting the same benzothiophene acetic acid core (**1d**), but with the addition of a functionalized bicyclic pyridocyclopentyl ether, as shown in compounds **1e** and **1f**. This series exhibits higher potency against GOAT than Takeda compounds. Additionally, another series of GOAT inhibitors based on benzothiophene, represented by scaffold **1g** (**Figure 1.9**), was revealed. These inhibitors maintain the heterocyclic linkage to the 6-position of the benzothiophene but replace the ether with an ester linkage. They also substitute the acetic acid at the 3-position of the benzothiophene with an ether linkage at the 2-position connected to a functionalized pyridine. The most active compounds in this series include fluorine and methoxy substituents on the pyridine ring, as seen in compounds **1h** and **1i**.

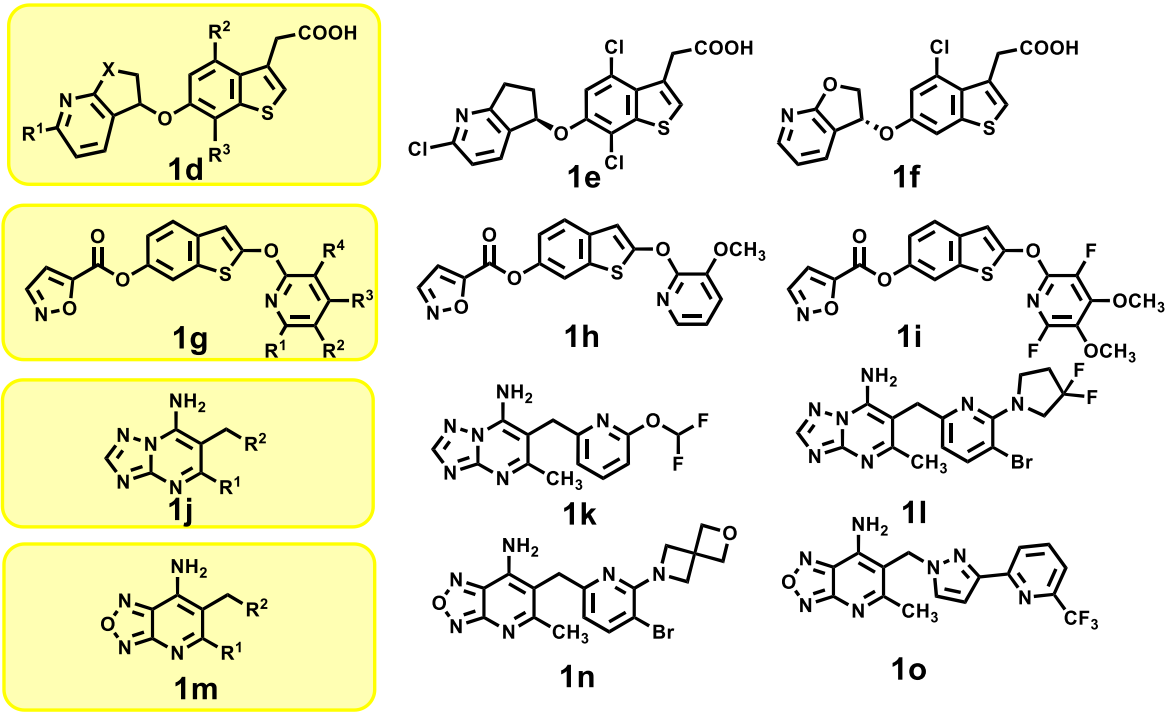

Figure 1.9. Heterocyclic small molecule GOAT inhibitors from other patents.

Recently, Boehringer Ingelheim researchers also revealed the most potent GOAT inhibitors. They introduced a series of 7-amino-[1,2,4]triazolo[1,5,-a]pyrimidines, with scaffold **1j**. The most effective compounds in this series included a methyl group at the 5-position of the pyrimidine and a benzylic pyridyl group at the 6 position, as seen in compounds **1k** and **1l** (**Figure 1.9**). These inhibitors typically had a fluorinated ether or amine at the 2-position of the pyridine and exhibited picomolar inhibitory activity against GOAT in ELISA assays. Furthermore, 7-amino-[1,2,5]oxadiazolo[3,4,-b]pyridines derivatives (scaffold **1m**), were reported to be even more potent GOAT inhibitors. The most active compounds, like **1n** and **1o**, showed exceptional potency. Future research will explore whether these significantly more potent GOAT inhibitors can effectively impact animal models of obesity and diabetes [172].