

# Understanding Toll-like Receptor Modulation Through Machine Learning

Inaugural-Dissertation

to obtain the academic degree

Doctor rerum naturalium (Dr. rer. nat.)

submitted to the Department of Biology, Chemistry, Pharmacy  
of Freie Universität Berlin

by

**Lihua Deng**

from Ganzhou, China

2021, Berlin



I hereby declare that I have completed the present work myself and have used no tools other than those listed here.

The presented thesis was prepared from October 2016 till October 2020 under the supervision of Prof. Dr. Gerhard Wolber at the Institute of Pharmacy of the Freie Universität Berlin.

Supervisor: Prof. Dr. Gerhard Wolber

Second examiner: Ass.-Prof. Dr. Johannes Kirchmair

Date of defense: 21.05.2021



# Contents

<b>Acknowledgements</b> . . . . .	<b>i</b>
<b>List of Abbreviations</b> . . . . .	<b>iii</b>
<b>Abstract</b> . . . . .	<b>vii</b>
<b>Zusammenfassung</b> . . . . .	<b>ix</b>
<b>1 Introduction</b> . . . . .	<b>1</b>
1.1 Therapeutic relevance of TLRs . . . . .	4
1.1.1 Therapeutic fields for TLR agonists . . . . .	6
1.1.2 Therapeutic fields for TLR antagonists . . . . .	9
1.2 Pharmacological/biological concepts used in <i>TollDB</i> . . . . .	10
1.3 Machine learning in drug discovery . . . . .	11
1.3.1 Machine learning model evaluation methods . . . . .	13
<b>2 Aim and Objectives</b> . . . . .	<b>17</b>
<b>3 Methods and Materials</b> . . . . .	<b>19</b>
3.1 Data collection and curation . . . . .	19
3.2 Web deployment for <i>TollDB</i> . . . . .	21
3.3 Data analysis for <i>TollDB</i> . . . . .	23
3.3.1 Basic data analysis and visualization . . . . .	23
3.3.2 Matched molecular pairs and activity cliffs . . . . .	23
3.4 Machine learning model development . . . . .	25
3.4.1 Dataset preparation for machine learning models . . . . .	25
3.4.2 Model selection and hyperparameter tuning . . . . .	28

3.4.3	Model evaluation and validation . . . . .	29
<b>4</b>	<b>Results . . . . .</b>	<b>31</b>
4.1	Database information . . . . .	31
4.1.1	Database schema . . . . .	31
4.1.2	Web application . . . . .	33
4.2	Basic statistical analysis for <i>TollDB</i> . . . . .	36
4.3	Chemical space analysis for <i>TollDB</i> . . . . .	48
4.4	Matched molecular pairs and activity cliffs study . . . . .	49
4.4.1	Case study of activity cliffs . . . . .	51
4.5	Machine learning study . . . . .	59
<b>5</b>	<b>Discussion . . . . .</b>	<b>69</b>
5.1	Data compilation . . . . .	69
5.2	Data analysis and docking studies . . . . .	70
5.2.1	Data analysis . . . . .	70
5.2.2	Docking studies . . . . .	71
5.3	Machine learning studies . . . . .	71
<b>6</b>	<b>Conclusion and Outlook . . . . .</b>	<b>77</b>
	<b>Appendix . . . . .</b>	<b>79</b>
A	MMPs for <i>TollDB</i> . . . . .	79
B	Machine learning results for random search and grid search . . .	88
	<b>Bibliography . . . . .</b>	<b>107</b>
	<b>List of Figures . . . . .</b>	<b>129</b>
	<b>List of Tables . . . . .</b>	<b>132</b>
	<b>Publications . . . . .</b>	<b>135</b>

# Acknowledgements

The present work was carried out at the Institute of Pharmacy of Freie Universität Berlin from 2016 to 2020.

Firstly, I would like to express my sincere gratitude to Prof. Dr. Gerhard Wolber for the nice and friendly supervision during the last years. I really enjoyed the harmonious atmosphere in his working group. Secondly, I would like to thank all my colleagues in the Computer-aided Drug Design lab, namely Marcel Bermúdez, Manuela Murgueitio, Dora Šriбар, Dmitri Stepanov, David Schallar, Trung Ngoc Nguyen, David Machalz, Christian Omieczynski, Szymon Pach, Jérémie Mortier, Alexandra Naß, Robert Schulz, Tessa Noonan etc. for their help and support in my life and study in Berlin. A special thank goes to Dongsheng Deng for his help with the design of the *TollDB* logo.

Gratefully acknowledged is the computing cluster Curta (former name Soroban) of the FU Berlin and the help of their support team is gratefully acknowledged. I also would like to thank Prof. Dr. Jörg Rademann for providing lab resources for initial trials for experiments at the first year of my Ph.D. and help from his group members. I would also express my gratitude to Johannes Kirchmair for reviewing the thesis and to all the committee members.

Furthermore, I would like to thank the Chinese Scholarship Council (CSC) for the financial support during my study in Germany. I also want to thank the Frauenförderung of the Fachbereich Biologie, Chemie, Pharmazie at Freie Universität Berlin for supporting my visit to the Drug Design Summer School in Vienna during the first year of my study. This conference provided a great opportunity for me to learn new techniques in drug design.

Finally, I want to thank my parents, my brother, and my friends, I would not have come so far without the continuing support from them.





# List of Abbreviations

ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
AI	Artificial intelligence
AP1	Activator protein 1
ASA	Water accessible surface area
AUC	Area under the ROC curve
COPD	Chronic obstructive pulmonary disease
CREB	Cyclic AMP-responsive element-binding protein
CTL	Cytotoxic T cells
DAMPs	Danger-associated molecular patterns
DNA	Deoxyribonucleic acid
dsRNA	Double-stranded RNA
DT	Decision tree
EC <sub>50</sub>	Half maximal effective concentration
ECD	Extracellular domain
ECFPs	Extended-connectivity fingerprints
FN	False negative
FP	False positive
FPR	False positive rate
GPUs	Graphical processing units
H-bond	Hydrogen bond
HSP	Heat shock proteins
IC <sub>50</sub>	Half maximal inhibitory concentration
IFN	Interferon
IKK	Inhibitor of NF- $\kappa$ B kinase

IL-1R	Interleukin-1 receptor
ILC2s	Group 2 innate lymphoid cells
IRAKs	IL-1R-associated kinases
IRFs	Interferon-regulatory factors
ITC	Isothermal titration calorimetry
JNK	JUN N-terminal kinase
$K_d$	Dissociation constant
kNNs	K-nearest neighbors
$\log P$ ; $\log P(o/w)$	<i>n</i> -octanol/water partition coefficient (logarithmic value)
LPS	Lipopolysaccharide
LR	Logistic regression
LRR	Leucine-rich repeat
LTA	Lipoteichoic acid
MAL	MyD88-adaptor-like protein
MALP-2	Macrophage-activating lipopeptide-2
MAPKs	Mitogen-activated protein kinases
MCC	Matthews correlation coefficient
MCS	Maximum common substructure
MKK	MAP kinase kinase
MMPs	Matched molecular pairs
MOE	Molecular Operating Environment
MPLA	Monophosphoryl lipid A
MQNs	Molecular quantum numbers
MST	Microscale thermophoresis assay
MW	Molecular weight
MyD88	Myeloid differentiation primary response protein 88
NF- $\kappa$ B	Nuclear factor- $\kappa$ B
ODN	Oligodeoxyribonucleotide
PAMPs	Pathogen-associated molecular patterns
PC	Principal component
PCA	Principal component analysis
PCR	polymerase chain reaction

QSAR	Quantitative structure-activity relationship
RDBMS	Relational database management system
RF	Random forest
RIP1	Receptor-interacting protein 1
RNA	Ribonucleic acid
ROC curve	Receiver operating characteristic curve
rRNA	Ribosomal RNA
SMILES	Simplified molecular input line entry system
SPR	Surface plasmon resonance
SQL	Structured query language
ssRNA	Single-stranded RNA
SVM; SVC	Support vector machine
TAB	TAK1-binding protein
TAK1	Transforming-growth-factor- $\beta$ -activated kinase
TBK1	TANK-binding kinase 1
T <sub>H</sub> 2	T helper 2 cells
TMD	Transmembrane domain
TN	True negative
TNF	Tumor necrosis factor
TNR	True negative rate
TLR	Toll-like receptor
<i>TollDB</i>	Toll-like receptor database
TP	True positive
TPR	True positive rate
TPSA	Topological polar surface area
TRAFs	TNF receptor-associated factors
TRAM	TRIF-related adaptor molecule
TRIF	TIR-domain-containing adaptor protein inducing interferon- $\beta$
VSA	Van der Waals surface area



# Abstract

Toll-like receptors (TLRs) represent one of the most fascinating and currently most widely studied immunologic targets, due to their crucial role in forming the first barrier in immune response. The structurally conserved TLRs consist of ten human subtypes (TLR1-TLR10), with a structurally broad range of natural ligands, including lipids, peptides, and ribonucleic acid (RNA), which challenges the rational design of drug-like TLR ligands. Therefore, despite their enormous therapeutic potential as powerful regulators of inflammatory pathways, only few TLR modulators (e.g., Imiquimod) are currently in clinical use.

Since no complete and up-to-date repository for known TLR modulators is currently available, we carefully collected and manually curated data to create a Toll-like receptor database (*TollDB*), the first database which includes all reported small organic drug-like molecules targeting TLRs and detailed pharmacological assay conditions used for their characterization. *TollDB* is freely accessible via <https://tolldb.drug-design.de> and provides three different search possibilities including a ligand-centered simple search, an advanced search that can retrieve information on biological assays and a structure search.

Currently, *TollDB* contains 4925 datapoints describing 2155 compounds tested in 36 assay types using 553 different assay conditions. Among all the 2155 compounds, 1278 are not reported as TLR ligands by *ChEMBL* database. Users can retrieve information about the measured inactives and multi-target TLR ligands from *TollDB*. After statistical analysis for *TollDB*, we compared the chemical space covered by compounds in *TollDB* to that covered by the compounds in *DrugBank*. Next, we explored the matched molecular pairs (MMPs) and activity cliffs, then used docking to explain the activity cliffs between MMPs. After a thorough analysis of the entire database, we used a selected dataset from *TollDB* to train machine learning models to

distinguish active ligands for different subtypes. These validated models can be used for prioritizing hits from virtual screening for chemical synthesis or for biological testing.

The curated database can be directly used in many ways, for example, as a validation dataset for pharmacophore model evaluation, as a virtual screening library for drug-repurposing or as reference for pharmacological assay design. *TollDB* represents a unique and useful resource for various research fields such as medicinal chemistry, immunology, computational biology and promotes the use of artificial intelligence in modern drug design campaigns.

# Zusammenfassung

Toll-like Rezeptoren (TLRs) sind aufgrund ihrer entscheidenden Rolle bei der Bildung der ersten Barriere der Immunantwort eines der faszinierendsten und derzeit am häufigsten untersuchten immunologischen Ziele. Die strukturell konservierten TLRs weisen zehn menschliche Subtypen (TLR1-TLR10) auf. Sie umfassen ein breites strukturelles Spektrum natürlicher Liganden, einschließlich Lipiden, Peptiden und RNA, was das rationale Design von arzneimittelähnlichen TLR-Liganden herausfordernd macht. Daher werden derzeit trotz ihres enormen therapeutischen Potenzials als starke Regulatoren von Entzündungswegen nur wenige TLR-Modulatoren (z. B. Imiquimod) klinisch eingesetzt.

Da derzeit keine vollständiges und aktuelles Respository für bekannte TLR Modulatoren verfügbar ist, haben wir sorgfältig Daten gesammelt und manuell überprüft, um eine Toll-like-Rezeptor-Datenbank *TollDB* zu erstellen. Diese Datenbank enthält alle uns bekannten kleinen organischen arzneimittelähnlichen Moleküle mit detaillierten pharmakologischen Testbedingungen, die für ihre Charakterisierung verwendet wurden. *TollDB* ist unter <https://tolldb.drug-design.de> frei zugänglich und bietet drei verschiedene Suchmöglichkeiten, darunter eine Liganden zentrierte einfache Suche, eine erweiterte Suche, mit der Informationen zu biologischen Assays abgerufen werden können, und eine strukturelle Suche.

Derzeit enthält *TollDB* 4925 Datenpunkte, die 2155 Verbindungen beschreiben, die in 36 in vitro Testtypen unter Verwendung von 553 verschiedenen Testbedingungen getestet wurden. Von allen 2155 Verbindungen sind 1278 nicht in der *ChEMBL* Datenbank enthalten. Benutzer können bei der *TollDB* auch Informationen zu den gemessenen inaktiven und Multi-Target-TLR-Liganden erhalten. Nach der statistischen Analyse für *TollDB* haben wir den von den Verbindungen in *TollDB* abgedeckten chemischen Raum mit dem von den Verbindungen in *DrugBank* abgedeckten verglichen.

Wir haben die *matched molecular pairs* und *activity cliffs* untersucht. Nachdem wir ein umfassendes Verständnis der Daten in der *TollDB* erlangt haben, haben wir die Daten verwendet, um Modelle für maschinelles Lernen zu trainieren, um aktive Liganden für verschiedene Subtypen zu identifizieren. Diese validierten Modelle können zur Priorisierung von Treffern aus dem virtuellen Screening zur Synthese oder zum Testen verwendet werden.

Zusammenfassend kann die Datenbank in vielen Aspekten direkt verwendet werden, beispielsweise als Validierungsdatensatz für die Bewertung eines Pharmakophormodells, als virtuelle Screening-Bibliothek für die Umfunktionierung von Arzneimitteln oder als Referenzsubstanz, für das Design des pharmakologischen Assays. *TollDB* stellt eine einzigartige und nützliche Ressource für verschiedene Forschungsbereiche wie medizinische Chemie, Immunologie und Computerbiologie dar und fördert den Einsatz künstlicher Intelligenz in modernen Wirkstoffdesign.



# 1. Introduction

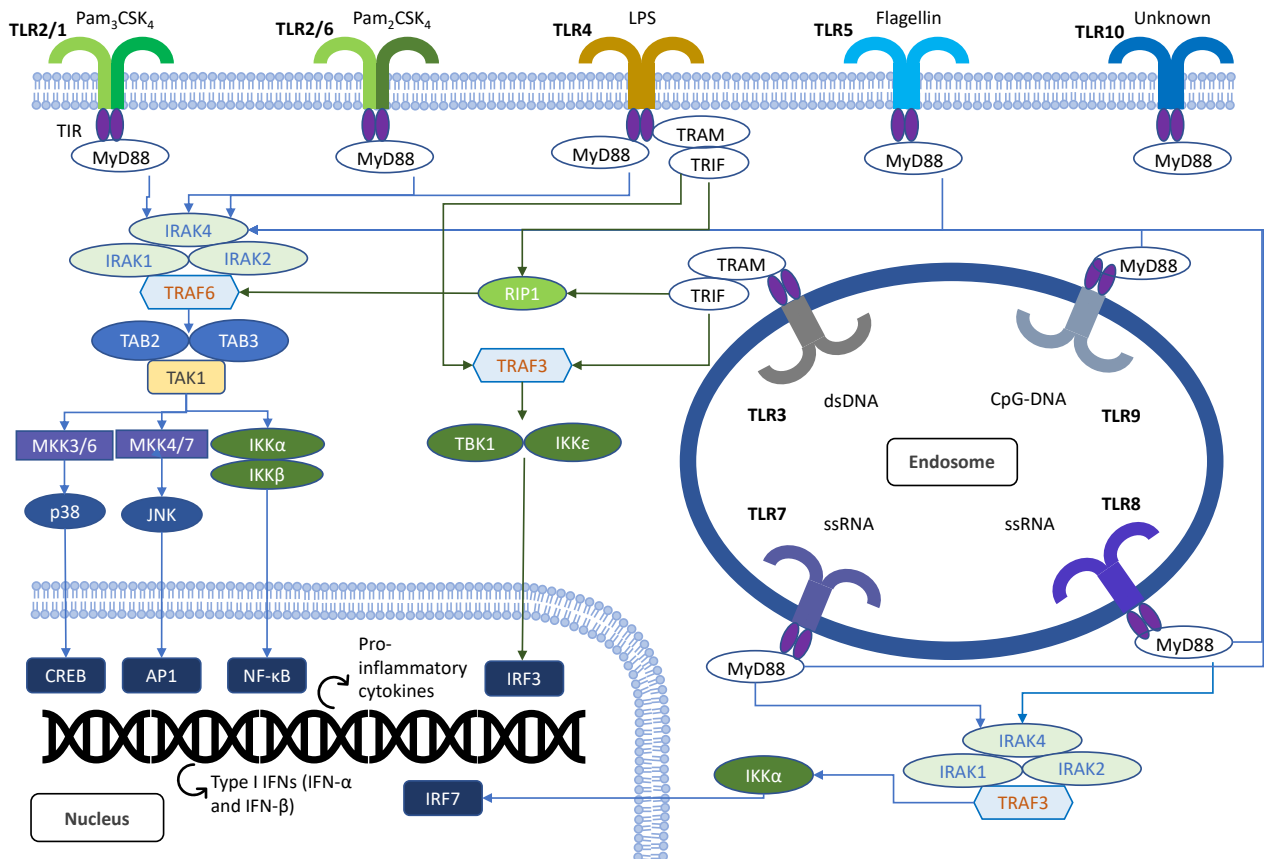
TLRs are a family of structurally conserved proteins that received their name from their similarity to the protein coded by the toll gene identified in *Drosophila* by Christiane Nüsslein-Volhard and Eric Wieschaus in 1985 [1]. A total of 13 TLRs have been discovered, 10 TLRs (TLR1-10) are found in human and 12 TLRs (TLR1-9 and TLR11-13) are expressed in rodents [2, 3]. The TLR family can be divided into two subgroups, extracellular and intracellular TLRs, depending on their cellular localization. Among these TLRs, TLR1, TLR2, TLR5, TLR6 and TLR10 are localized on the cell surface, whereas the TLR3 and TLR7, TLR8 and TLR9 are found in the endoplasmic reticulum, endosome, and lysosome. The subcellular localization of TLR4 is unique because it is localized in both the plasma membrane and endosomal vesicles [4]. However, TLR10-13 remain poorly characterized and their function is unclear [5, 6]. The native ligands for TLRs include lipoteichoic acid (LTA) from gram-positive bacteria or macrophage-activating lipopeptide-2 (MALP-2) originally isolated from *Mycoplasma fermentans* for TLR1, 2, and 6; lipopolysaccharide (LPS) from gram-negative bacteria for TLR4 and flagellin for TLR5; double-stranded RNA (dsRNA) for TLR3 and single-stranded RNA (ssRNA) for TLR7 and TLR8; TLR9 recognizes unmethylated CpG motifs of bacterial deoxyribonucleic acid (DNA). TLRs are highly expressed on immune cells and their presence and distribution vary by immune cell type [7]. TLRs are central to the innate and adaptive immune response and were considered as the first line in the immune defense [8, 9]. Thus, targeting small molecule TLR modulators has a great potential in developing prophylactic/therapeutic agents.

TLRs have major roles in the activation of the innate immune response against invading microbial pathogens [10] through recognizing specific sets of pathogen-associated molecular patterns (PAMPs) or danger-associated molecular patterns (DAMPs). Ligand types for TLRs range very broadly, including components of bacterial cell walls like LPS,

along with RNA or DNA immune complexes or DAMPs like heat shock proteins (HSP60, HSP70) [11]. All (patho)physiological TLR modulators are structurally complex, which challenges the rational design of drug-like TLR modulators. However, several studies led to the identification of small organic molecules with TLR modulating properties [12, 13]. These small molecules have the potential to be developed as therapeutic options against various diseases [8] such as sepsis [14], autoimmune diabetes [15] and metabolic syndrome [16], or used as vaccine adjuvants [17]. There are several TLR-targeting molecules that have been developed as marketed drugs, including **Imiquimod**, a TLR7 agonist that has been approved by the FDA for treatment of keratosis [18–20], condylomata acuminata [21] and basal cell carcinoma [22], **Resiquimod (R848)**, a dual TLR7 and TLR8 agonist granted as an orphan drug for the treatment of cutaneous T cell lymphoma in the European Union [23, 24] and **Rintatolimod**, a TLR3 agonist used in severe cases of myalgic encephalomyelitis/chronic fatigue syndrome [25].

TLRs belong to the class of integral membrane type I glycoproteins, which have three major domains: the extracellular domain (ECD) with 16-28 of leucine-rich repeat (LRR) motifs, the transmembrane domain (TMD), and the cytoplasmic domain (similar to that of interleukin-1 receptor; IL-1R), which is known as the Toll/IL-1R (TIR) domain [26, 27]. In 2005, the first crystal structure of the ECD of TLR reported was TLR3 [28, 29], and crystal structure analysis of other later-reported TLRs provided more information for elucidating the ligand/receptor binding mechanism [30]. Despite the differences in ligand interactions, the overall shape of the TLR-ligand complexes is strikingly similar: upon the ligand binding, two ECDs form an M-shaped homo- or heterodimer sandwiching the ligand molecule, thus bringing the TMD and TIR domain of the dimer in close proximity to trigger a downstream signaling cascade [31].

After ligand engagement, TLRs trigger multiple signaling pathways, including one depending on the adaptor molecule, termed MyD88, and another that is MyD88 independent, but depends on TIR-domain-containing adaptor protein inducing interferon- $\beta$  (TRIF) [33, 34]. MyD88 is a universal adaptor molecule used by almost all TLRs except TLR3 and it plays a crucial role in TLR signal transduction as a component for a “shared” signaling pathway. An overview of the TLR signaling pathways is shown in Figure 1.1. TLR signaling is initiated by ligand-induced dimerization of receptors. Following this, the Toll-IL-1-resistance (TIR) domain of TLRs engages TIR domain-



**Figure 1.1:** MyD88-dependent and MyD88-independent signaling pathways for TLRs. Figure modified from [32]. Note that TLR4 localizes at both the plasma membrane and the endosomes. dsRNA, double-stranded RNA; IKK, inhibitor of NF- $\kappa$ B kinase; LPS, lipopolysaccharide; MKK, MAP kinase kinase; RIP1, receptor-interacting protein 1; rRNA, ribosomal RNA; ssRNA, single-stranded RNA; TAB, TAK1-binding protein; TAK, TGF- $\beta$ -activated kinase; TBK1, TANK-binding kinase 1.

containing adaptor proteins (either myeloid differentiation primary-response protein 88 (MyD88) and MyD88-adaptor-like protein (MAL), or TIR domain-containing adaptor protein inducing interferon- $\beta$  (IFN  $\beta$ ), i.e. TRIF and TRIF-related adaptor molecule (TRAM). Engagement of the signaling adaptor molecules stimulates downstream signaling pathways that involve interactions between IL-1R-associated kinases (IRAKs) and the adaptor molecules TNF receptor-associated factors (TRAFs). This leads to the activation of mitogen-activated protein kinases (MAPKs), JUN N-terminal kinase (JNK) and p38, and finally to the activation of transcription factors. The two important families of transcription factors that are activated downstream are nuclear factor- $\kappa$ B (NF- $\kappa$ B) and interferon-regulatory factors (IRFs). Other transcription factors, such as cyclic AMP-responsive element-binding protein (CREB) and activator protein 1 (AP1), are also important. A major consequence of TLR signaling is the induction of various pro-inflammatory cytokines, and in the case of the endosomal TLRs, the induction of type I interferon (IFN) [32].

## 1.1 Therapeutic relevance of TLRs

TLRs are widely accepted to be present on immune cells and vast numbers of studies reported the presence of TLR message in leukocytes, with different TLR isoforms being present on specific subsets [35]. According to a study in 2002, which used quantitative real-time polymerase chain reaction (PCR) to systematically examine the expression of mRNAs encoding all known human TLRs, most tissues tested expressed at least one TLR, and several (spleen, peripheral blood leukocytes) expressed all subtypes. Monocyte-like THP-1 cells regulate TLR mRNAs in response to a variety of stimuli including phorbol esters, LPS, bacterial lipoproteins, live bacteria, and cytokines.

The immune response can discriminate between self and non-self, recognize specific pathogens, and uses an immunological memory to learn about the threat and enhance the immune response accordingly, thus protecting humans from potentially harmful disease-causing organisms or pathogens, like bacteria or viruses. When the human immune system is hyperresponsive, conditions including organ rejection, cardiovascular diseases, and autoimmune diseases (e.g., multiple sclerosis, allergy) may occur, on the contrary, when the immune system is hyporesponsive, conditions like cancer and

infectious disease or sepsis may occur [36]. Therefore, regulating the immune response can be a therapeutic avenue to treat such diseases. Since TLRs are engaged in not only innate immunity but also adaptive immunity, they are well-suited as targets for modulating the immune response to cure related diseases.

Studies have shown that TLRs are involved in a wide spectrum of diseases [8, 37–41]. As shown in Table 1.1, TLR agonists are potential agents to be used as vaccine adjuvants, and for the treatment of allergy, cancer, and infectious diseases. TLR antagonists are designed to reduce inflammation caused by infection or autoimmune disease. In the past two decades, many endeavors have been dedicated to delineating this relationship and compiling data regarding the TLR involvement in various diseases [42].

**Table 1.1:** TLR family as potential drug targets

<b>Disease</b>	<b>TLR involved</b>	<b>Therapeutic approach</b>	<b>references</b>
<b>Cancer</b> (including colon cancer, gastric cancer, breast cancer, melanoma, hepatocellular carcinoma, lung cancer, glioma, prostate cancer, ovarian cancer, cervical squamous cell carcinomas, chronic lymphocytic leukemia or used as vaccine adjuvants)	TLR1-9	Agonist	[19, 43–46]
<b>Allergic diseases</b> (including asthma, allergic rhinitis)	TLR4, 7, 8, 9	Agonist	[19, 43, 47]
<b>Infectious diseases</b> (anti-bacterial/anti-viral activity)	TLR2, 3, 4, 7, 8, 9	Agonist	[19, 43, 48–51]
<b>Acute/chronic inflammatory diseases</b> (including diabetes, chronic obstructive pulmonary disease)	TLR2, 4	Antagonist	[15, 52–54]
<b>Neuropathic pain, chronic pain</b>	TLR4	Antagonist	[55, 56]
<b>Autoimmune diseases</b> (including systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), systemic sclerosis [57], Sjögren’s syndrome, multiple sclerosis (MS))	TLR1-9	Antagonist	[19, 43, 58, 59]
<b>Sepsis</b>	TLR2, 4, 9	Antagonist	[19]

Upon binding of ligands with TLRs, the ligand-receptor complexes lead to the activation/inhibition of transcription factors and cytokine production. A signaling

pathway modulator does not directly interact with the TLRs but modulates the downstream signaling pathway, which could potentially either enhance or decrease the effects of a TLR ligand. The pathway modulator (that does not bind directly to TLRs) could also be targeted for potential therapeutic effect.

Despite the great potential for TLR ligands in therapy, it remains challenging for their development and application. TLR agonists have the potential to cause chronic inflammation and the non-specific activation of immune cells, and as an anti-tumor or anti-infectious agent they have the possibility of activating self-reproduction that might result in autoimmune diseases. Many TLR agonists under clinical studies were withdrawn from further development due to either lack of efficacy or serious side effects [18]. The major drawback to the use of TLR antagonists is an increase in susceptibility to infectious agents and tumors. These adverse effects highlight the number of issues that must be taken into consideration when designing TLR ligands.

### 1.1.1 Therapeutic fields for TLR agonists

Immunotherapy based on TLR agonists represents a promising way for the prevention and/or treatment of several disorders including cancer, allergy, and infectious diseases.

#### **Cancer**

Functional TLRs are not only expressed on immune cells but also on cancer cells, thus implicating TLRs in tumor biology [60]. Increasing bodies of evidence have suggested that TLRs act as a double-edged sword in cancer cells as uncontrolled TLR signaling provides a microenvironment that is necessary for tumor cells to proliferate and evade the immune response. Alternatively, TLRs can induce an antitumor immune response in order to inhibit tumor progression [61]. Activation of TLRs leads to the induction of cytokines such as IL-12, IL-10, IL-6, TNF- $\alpha$ , and types I and II IFN. In addition, it also enhances the activation of CD4<sup>+</sup> helper T cells and CD8<sup>+</sup> cytotoxic T cells (CTL), which leads to induction of Th1, Th2, and Th17 responses [62]. TLR agonists can therefore serve as immunotherapeutics or vaccine adjuvants for the treatment or prevention of cancer.

Several TLR agonists are being explored for cancer immunotherapy [63]. Monophosphoryl lipid A (**MPLA**) has been approved for prophylaxis of HPV-associated cervical

cancer (used as a human vaccine adjuvant) [64]. Also, the TLR7 agonist, **Imiquimod**, an imidazoquinoline derivative, has been approved by Food and Drug Administration (FDA) as a therapeutic agent for basal cell carcinoma and genital warts [65].

A phase 1 study of TLR8 agonist **VTX-2337** (alternatively: **Motolimod**) in treating patients with recurrent or persistent ovarian epithelial, fallopian tube, or peritoneal cavity cancer was successfully completed [66]. A phase 1b multicenter pre-surgical study to evaluate immune biomarker modulation in response to **VTX-2337** in combination with Nivolumab in subjects with resectable squamous cell carcinoma of the head and neck (SCCHN) is now under phase 1 recruiting [67]. The phase 1 study of a TLR4 agonist **GLA-SE** (Glucopyranosyl Lipid A in Stable Emulsion) in patients with Merkel cell carcinoma was completed in March 2018 [68], **GLA-SE** was also used as an adjuvant in a pilot study of peptide vaccine (melanoma antigen recognized by T-cells 1 (MART-1) antigen) for patients with resected melanoma [69]. The phase 1/2 dose-escalation study in patients with relapsed or refractory Waldenström’s macroglobulinemia using TLR7/8/9 antagonist **IMO-8400** (an oligonucleotide) was terminated due to lack of efficacy [70]. Phase 2 trial with intradermal **IMO-2125** (alternatively: **Tilsotolimod**) in pathological tumor stage pT3-4 cN0M0 melanoma (INTRIM) is under phase 2 study [71]. TLR9 agonist **SD-101**, in combination with Ibrutinib and local radiation, is now under phase 2 study for relapsed or refractory grade 1-3A follicular lymphoma [72]. Other clinical studies relating to TLR ligand on cancer immunotherapy before 2019 were summarized in references [73, 74].

### **Allergic disease: asthma and allergic rhinitis**

Allergies, also known as allergic diseases, are a number of conditions caused by hypersensitivity of the immune system to allergens. Common allergens include pollen and food. According to the site of contact with the allergen, different clinical manifestations may develop in the airways, skin, or gastrointestinal tract. The frequency of allergic diseases has increased over the last century [75]. Allergic rhinitis (AR) and asthma are two common allergic diseases of the respiratory system. Both the innate and adaptive immune systems are relevant for the development of asthma.

Asthma is mainly driven by type 2 immune responses, which comprise increased airway eosinophils, CD4<sup>+</sup> T helper 2 (T<sub>H</sub>2) cells. Recently, group 2 innate lymphoid

cells (ILC2s) have been identified that may play an important role in non-allergic asthma [76]. The recruitment of T<sub>H</sub>2 and ILC2s and later activation of transcription factors lead to the secretion of IL-4, IL-5, and IL-13 [77, 78].

Allergic rhinitis represents another allergic disease of the respiratory system. It is clinically defined as a symptomatic disorder of the nose induced after allergen exposure by IgE-mediated inflammation [79]. It affects people of all ages, peaking in the teenage years. Although allergic rhinitis is not a serious illness, it is clinically relevant because it underlies many complications. It is a major risk factor for poor asthma control and affects quality of life as well as productivity at work or school [80]. In 2011, a TLR8 agonist was used in phase 1b/2a trials for allergic rhinitis [81]. Compounds that target TLRs have been found to suppress airway inflammation, eosinophilia and airway hyperresponsiveness in various animal models of allergic inflammation [82].

A TLR7 agonist, **GSK2245035**, was studied for its effect on the allergen-induced asthmatic response but was withdrawn after revision of new data [83, 84]. It was also used in a clinical study for treating allergic rhinitis [85–87]. **CYT003**, a TLR9 agonist, was also withdrawn [88, 89]. This indicates that a great effort is still needed for developing TLR agonists for the treatment of asthma and allergic rhinitis.

## **Infectious diseases**

Infectious diseases are caused by pathogenic microorganisms, such as bacteria, viruses, parasites, or fungi; the diseases can be spread, directly or indirectly, from one person to another. TLRs can induce a multitude of inflammatory cytokines and mediators and play a major role in viral clearance. This has led to the discovery that TLR agonists can be utilized to control viral infections [49]. The MyD88-dependent pathway is important for protection against bacterial infection [90]. TLR agonists have been harnessed as anti-microbial agents or as adjuvants [17, 91]. Other development of TLR agonists as anti-bacterial/anti-viral agents are summarized in review [92].

Starting from the end of 2019, the COVID-19 epidemic [93] has spread all over the world and posed a great threat to public health. This pandemic had a profound impact on people's lives [94] and the global economy [95]. The TLR signaling pathways, specifically the TRIF-dependent pathway, which activate the type I IFNs and inflammatory factors, is important for enhancing the protection for uninfected cells. Therefore, TLR3



and TLR4 have been highlighted as crucial and can be targeted for generating host defense to COVID-19 [96].

### 1.1.2 Therapeutic fields for TLR antagonists

TLRs exhibit homeostatic roles in immunity. Therefore, modulating the immune response by using TLR agonists or antagonists might be of therapeutic value. Their use in clinical trials to treat septic shock and autoimmune disease shows great potential [97].

#### **Acute/chronic inflammatory diseases**

Inflammation is an essential process in response to injury and infection. Manifested in the form of heat, pain, redness, and swelling, inflammation represents a critical process as the body adapts to restore homeostasis after infection or injury. However, excessive inflammation contributes to the development of inflammatory and autoimmune diseases [5].

The pathogenesis of chronic obstructive pulmonary disease (COPD) is driven by chronic inhalation of noxious particles, often cigarette smoke, that persistently stimulates innate and inflammatory responses [98]. TLR2 and TLR4 play an important role in the immunoregulation of the inflammatory process in COPD [99], providing a new target for COPD treatment. TLRs and their downstream protein kinases may be potential targets for the treatment of traumatic brain injury [100]. Recent studies indicate that TLR activation could be the molecular basis for the development of metabolic syndrome-induced inflammation, thus providing a new strategy for developing TLR antagonists to suppress unwanted metabolic syndrome-associated inflammatory response [101].

#### **Autoimmune diseases**

Under certain circumstances, dis-regulation of the inflammation process can lead to pathological conditions, such as autoimmune diseases, including rheumatoid arthritis, multiple sclerosis, and atherosclerosis. The pivotal function of TLR2/4 in the pathogenesis of autoimmune diseases is summarized in a review by Liu et al. [58]. Activation of TLR signaling pathways leads to the pro-inflammatory cytokine production and direct/indirect T cell activation, which are now considered to be major factors in the

development of autoimmunity [59, 102]. This is another promising application of TLR antagonists in treatment.

Although it failed in the clinical trial for cancer treatment, **IMO-8400** was used in phase 2 study for patients with moderate to severe plaque psoriasis [103] and demonstrated clinical improvement.

## 1.2 Pharmacological/biological concepts used in *TollDB*

With considerable information on structure, function and signaling known for the TLRs, and their relationship with numerous diseases, researchers have targeted TLRs to discover potential therapeutic agents. During the experimental testing procedure, many different materials and experimental methods were used.

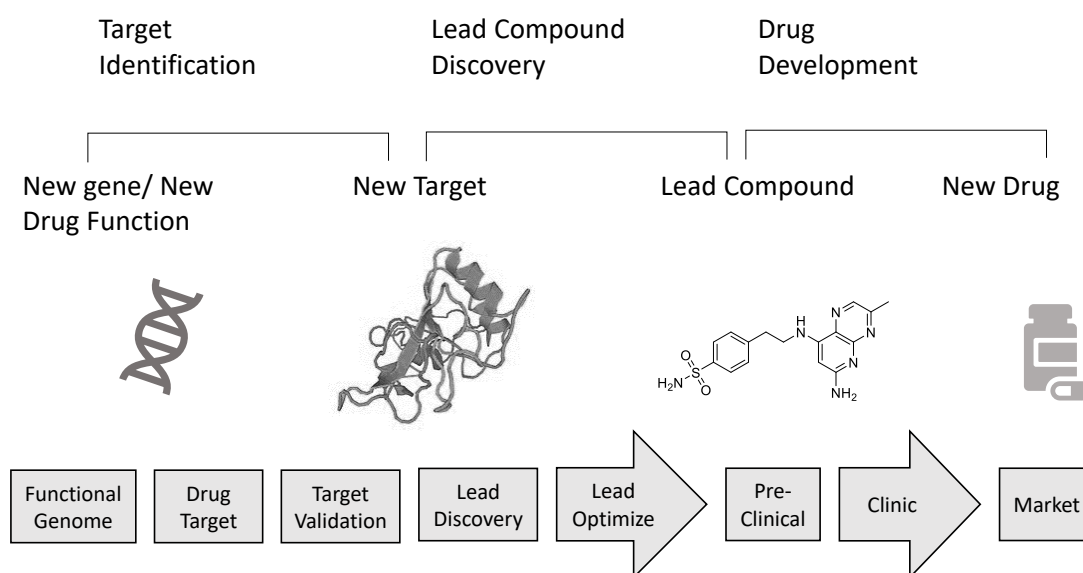
The TLR modulators database we built is referred to as *TollDB* in all following sections. All assays collected in *TollDB* have been classified as two main types, namely functional assay and binding assay. Functional assays were used to investigate whether the molecule tested is relating to a particular cellular pathway or biological process. Binding assays were used to test the binding affinity between the target protein and the testing molecule. Functional assays in *TollDB* mainly include gene reporter assays [104], cytokines related assays [105, 106], mRNA related assays [107, 108], transcription factor related assays [109, 110] and cell proliferation related assays [111], depending on the content that was being measured. Binding assays in *TollDB* include surface plasmon resonance (SPR) [112], isothermal titration calorimetry (ITC) [113], fluorescence polarization [114] and microscale thermophoresis assay (MST) [115]. These assays also represent the main methods for measuring ligand-acceptor binding affinity.

Tested effects in *TollDB* were categorized as antagonism and agonism, depending on whether the testing compound has a positive, activating effect, or a negative, inhibiting effect on the signaling pathway. We also curated the cell line category for *TollDB* according to the cell line type used in the functional assays, where the two categories comprise primary cell line and immortalized cell line. Primary cell lines are cells that are directly derived from normal embryonic or adult tissue, which are propagated in culture. These cells are considered to be genetically identical to cells in the tissue of origin [116]. Immortalized cell lines are comprised of a single cell type that can be

serially propagated in culture either for a limited number of cell divisions (approximately thirty) or otherwise indefinitely [117].

### 1.3 Machine learning in drug discovery

The drug discovery process (Figure 1.2) in general can be divided into three parts: (1) target identification; (2) lead discovery; and (3) clinical trials [118]. *TollDB* focuses on lead discovery and lead optimization. Lead design is the most decisive step in the process of drug discovery and rational drug design makes it more likely to find new structures possessing the required properties and biological activity. With the fast development of new computer hardware such as graphical processing units (GPUs) that make parallel processing faster, drug discovery has taken advantage of abundant, high-quality data to make use of machine learning or deep learning for accurate predictions, thus speeding up the process and reducing failure rates. Opportunities to apply machine learning may occur in all stages of drug discovery [119].

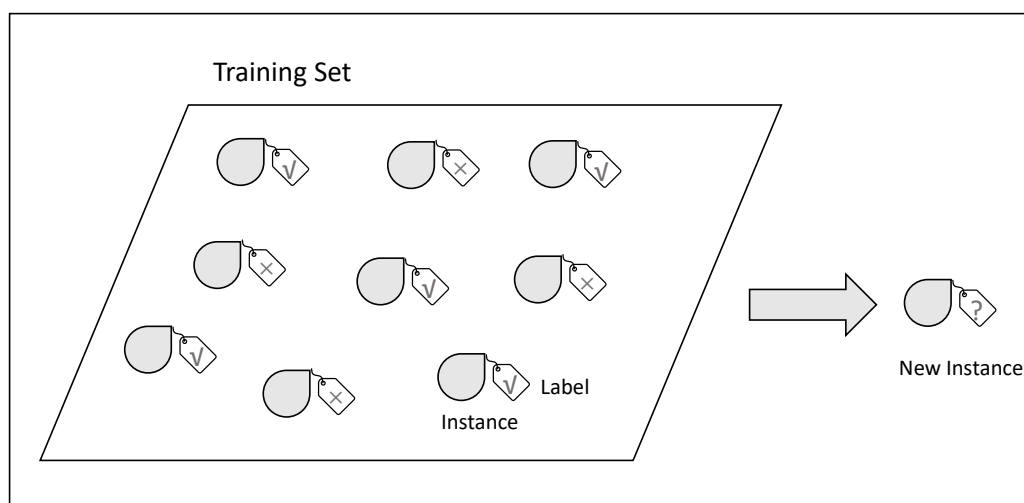


**Figure 1.2:** Drug discovery process. Figure modified from [118].

Artificial intelligence (AI), in particular machine learning and deep learning, have been successfully employed for drug discovery and design, including virtual screening, physicochemical and ADMET properties predictions, drug repurposing and so on, all

summarized in a review by Yang et al. [120]. For application to TLRs, the first reported use of random forest-based approaches [121, 122] to predict novel mouse TLR9 agonists based on an in-house experimentally validated single-stranded DNA oligonucleotides dataset was reported by Khanna and colleagues [123]. This will be discussed in the Discussion section.

The term “machine learning” was coined in 1959 by Arthur Samuel, with the general definition that machine learning is the field of study that gives computers the ability to learn without being explicitly programmed [124]. Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: “A computer program is said to learn from experience  $E$  with respect to some class of tasks  $T$  and performance measure  $P$  if its performance at tasks in  $T$ , as measured by  $P$ , improves with experience  $E$ .” [125]. Depending on the amount and type of supervision a model gets during training, machine learning systems can be classified into four major categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning. In supervised learning, the training data fed to the algorithm includes the desired solutions, called labels. A typical supervised learning task is classification (example shown in Figure 1.3), another is to predict a target numeric value, given a set of features called predictors. This sort of task is called regression [126]. In our current work, we concentrate on building models for prediction of potential active molecules. Thus, classification models are used.



**Figure 1.3:** A labeled training set for supervised learning (e.g., spam classification). Figure modified from [126]

In practical machine learning projects, data is as important as algorithms. Sometimes the data source cannot be used directly, this can be due to several reasons: missing values, imbalanced dataset, noise, too many variables, specific domain restriction of the algorithms, etc. Therefore, most of the times the dataset needs to be preprocessed. Furthermore, data preprocessing has a huge impact on model performance [127–129].

Common data types include numerical (such as height) and categorical (such as a label), and each of them could be further subdivided, numerical variables can be subdivided as integer variables and floating-point variables, and categorical variables can be subdivided as boolean (dichotomous), ordinal, or nominal variables.

Preprocessing of the dataset includes extracting and transforming data, handling of missing values, encoding for categorical features, scaling and normalization for numerical features etc. It is very common to encounter categorical features in a dataset, and these categorical features can be ordinal or not. However, machine learning algorithms can only read numerical values. Thus, it is essential to encode categorical features into numerical values. The commonly used categorical encoding methods provided by scikit-learn include label encoding and one-hot encoding.

In machine learning, hyperparameters are used to control the learning process and are set before the learning process begins. In scikit-learn [130, 131] hyperparameters are passed in as arguments to the constructor of the model classes. The most commonly used methods for optimizing hyperparameters include grid search [132] and random search [133]. The benefit of grid search is that it is guaranteed to find the optimal combination of parameters supplied. The drawback is that it can be computationally expensive and time consuming. Random search differs from grid search mainly in that it searches the specified subset of hyperparameters randomly instead of exhaustively. The major benefit of random search is decreased processing time. Note that random search does not guarantee the optimal combination of hyperparameters.

### 1.3.1 Machine learning model evaluation methods

In the field of machine learning and specifically the problem of classification tasks, different performance measurement methods are used [134]. In the case of a supervised binary classification problem, a confusion matrix (see Table 1.2) is often used. Each row of the matrix represents the instances in a predicted class while each column represents

the instances in an actual class (or vice versa), i.e., reports the number of false positives (FP), false negatives (FN), true positives (TP) and true negatives (TN). Terminologies derived from a confusion matrix are shown in the following paragraphs [135].

**Table 1.2:** Confusion matrix

	Actual Positive (1)	Actual Negative (0)
Predicted Positive (1)	TP	FP
Predicted Negative (0)	FN	TN

Accuracy is used for evaluating the performance of machine learning models, and for binary classification, accuracy can be calculated in terms of positives and negatives as in Equation 1.1 [136]. It works well only if there are equal number of samples belonging to each class. Accuracy is not always perfect for model evaluation, especially for imbalanced datasets [137].

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1.1)$$

We use a balanced accuracy score to avoid inflated performance estimates on imbalanced datasets. Balanced accuracy score is the macro-average of recall scores per class or, equivalently, raw accuracy, where each sample is weighted according to the inverse prevalence of its true class. Thus, for balanced datasets, the balanced accuracy is equal to accuracy. In the case of binary classification, balanced accuracy is equal to the arithmetic mean of sensitivity (true positive rate, TPR) and specificity (true negative rate, TNR). See Equation 1.2.

$$balanced\ accuracy = \frac{1}{2} \left( \frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) \quad (1.2)$$

Precision focuses on false positive errors. See Equation 1.3.

$$precision = \frac{TP}{TP + FP} \quad (1.3)$$

Recall is also called the true positive rate (TPR) or sensitivity, and corresponds to the proportion of positive data points that are correctly considered as positive, with respect to all positive data points. For a binary classification problem, it can be calculated using Equation 1.4.

$$recall = TPR = \frac{TP}{TP + FN} \quad (1.4)$$

The false positive rate (FPR) (fall out) is equal to one minus specificity or true negative rate (TNR). Defined in Equation 1.5.

$$FPR = \frac{FP}{FP + TN} \quad (1.5)$$

The area under the ROC curve [138] (AUC [139]) is one of the most widely used metrics for evaluation [140]. It is used for binary classification problems. The AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example. AUC is the area under the curve of plotting FPR vs. TPR at different points in the range [0, 1]. Since AUC is scale-invariant, it measures how well predictions are ranked, rather than their absolute values. It is also classification-threshold-invariant, measuring the quality of the model's predictions irrespective of what classification threshold is chosen. Thus, we also used this to compare model performance.

The F1 Score is the harmonic mean between precision and recall, see Equation 1.6. The range for the F1 Score is [0, 1]. It tells you how precise your classifier is (how many instances it classifies correctly), as well as how robust it is (it does not miss a significant number of instances).

$$F1 = 2 \cdot \frac{1}{\frac{1}{precision} + \frac{1}{recall}} \quad (1.6)$$

The Matthews correlation coefficient (MCC) is a balanced measure of prediction quality which not only takes TP and FP into account, but also TN and FN, thus can be used if the classes are of very different sizes. It is calculated according to Equation (1.7).

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP) \cdot (TP + FN) \cdot (TN + FP) \cdot (TN + FN)}} \quad (1.7)$$

While there is no perfect way of describing the confusion matrix of true and false positives and negatives by a single number, the Matthews correlation coefficient is generally regarded as being one of the best matrices [141]. Thus we used MCC as the primary measure of model performance for our binary classification problems to

compare the used algorithms in order to choose the best for a specific problem and at the same time take accuracy score, balanced accuracy score, and the AUC into account.



## 2. Aim and Objectives

TLRs are a family of structurally conserved, single-pass membrane-spanning proteins that play important roles in the activation of the innate immune response against invading microbial pathogens. They recognize specific sets of PAMPs or DAMPs. Ligand types for TLRs range from components of bacterial cell walls like LPS, to RNA or DNA immune complexes, to DAMPs like HSP60, HSP70. All (patho)physiological TLR modulators are structurally very complex, which challenges the rational design of drug-like TLR modulators. Despite a large number of TLR modulators emerging in publications in the past two decades, there is no complete and curated database available that containing all published small molecule TLR modulators.

The aim of this thesis is to make use of all available information for TLRs to develop a strategy for assisting rational drug design through machine learning approaches. As a starting point, we collected all available information about published small molecule TLR modulators and transformed them into a MySQL database. This process is not only fundamental for the data-driven approaches including investigation of chemical space coverage, matched molecular pairs and activity cliffs, but also for the development of predictive machine learning models. To achieve our goals, we proceed according to the following steps:

1. Literature search, reading, data collection and curation, database compilation.
2. Web application development for easier information retrieving and publication.
3. Data analysis and visualization for *TollDB*, chemical space analysis to get a deeper understanding of the database.
4. Development of machine learning models including data preprocessing, algorithm selection, feature selection, hyperparameter tuning and model evaluation.

The constructed database will provide a more profound understanding of TLR ligand profiles. The developed web application will help scientists from various fields in targeting TLRs for drug development, and the optimized machine learning models will provide a useful tool for filtering hits from virtual screening campaigns, prioritizing compounds for synthesis and testing, or even predicting TLR related off-target effects when designing ligands for other targets.

## 3. Methods and Materials

### 3.1 Data collection and curation

For collecting all original papers related to TLR ligands, our reference sources include the following: SciFinder, PubMed, Web of Science, Google Scholar, Google. To obtain research articles containing information on TLR modulators, systematic searches were performed using various keywords such as “toll-like receptor”, “toll-like receptor ligand”, “TLR 2 (or other subtypes such as 1, 3, 4, 5, 6, 7, 8, 9, 10) ligand”, “TLR 2 (or other subtypes such as 1, 3, 4, 5, 6, 7, 8, 9, 10) inhibitor (activator, agonist, antagonist)”, “small molecule toll-like receptor ligand” and so on. Articles describing prediction methods without experimental data, book chapters and patents were excluded. The rest of the search articles were manually checked by cautious reading through the abstract to discriminate whether it was focusing on small molecule modulator for TLRs with reported biological assay/testing. Review articles were checked for their cited papers. For modulators that have been examined in more than one study or tested in different labs, multiple entries were included.

More detailed inclusion criteria for the data from a publication were as follows:

1. Publication is accessible in the public domain
2. Publication is the primary source for the data
3. Related activity testing assays are part of the publication

Compounds from publications that satisfy the aforementioned criteria were collected, curated and compiled manually in Instant JChem [142]. Other criteria for inclusion of compounds were as follows:

1. Compounds with detailed testing information in the original publication

2. Compounds with a molecular weight (MW) of less than 700 Da
3. Compounds that were not cytotoxic when tested at the effective concentrations

It is noteworthy to mention that we included information about stereochemistry if available, i.e., we considered stereo isomers as well as racemates if available as unique entries, since bioactivity is likely to differ between them. Additional information about related testings from the publications were collected and curated, which includes the following aspects (see Table 3.1):

**Table 3.1:** Data collection and curation

Column name <sup>a</sup>	Description
<i>TargetName</i>	the TLR subtype that an assay is targeting
<i>TargetOrganism</i>	organism or the origin of target TLR gene
<i>AssayType</i>	categorized as functional assay or binding assay
<i>CellLineName</i>	cell line used in the assay
<i>CellLineCategory</i>	categorized as primary cell line or immortalized cell line
Stimuli <sup>b</sup>	if used in some antagonistic-effect related assays
<i>ActivityType</i> <sup>b</sup>	IC <sub>50</sub> , EC <sub>50</sub> , K <sub>d</sub> or other
<i>EffectType</i>	categorized as agonism or antagonism
<i>Result</i>	testing result for the corresponding activity type
<i>Unit</i>	unit for the result
<i>ResultLabel</i>	categorized as active or inactive

<sup>a</sup>: in italic if used as column names in *TollDB*.

<sup>b</sup>: information included in *ConditionDescription* if exists.

The criteria for determining whether a compound is active/inactive include the following: (1) compounds are defined clearly as active or inactive in the original publications; (2) experimental data for all compounds that have an obvious threshold in IC<sub>50</sub>, EC<sub>50</sub>, K<sub>d</sub>, or activation/inhibition rate; (3) for those that the authors did not note clearly as active or inactive, and those that do not have an obvious threshold, the following criteria were applied: (i) if the negative control is provided, the compound that generates an effect that is significantly different compared to the negative control is considered as active, otherwise considered as inactive; (ii) if the positive control is provided, the compound that generates at least half of the effect of the positive control is considered as active, otherwise considered as inactive; (iii) when neither (i) nor (ii) applies, compounds used for further testing are considered as active and those only tested in the initial screening assays but without further testings are considered

as inactive. The overall workflow for publication selection, compound selection and compound labeling is shown in Figure 3.1.

We took advantage of the friendly graphical user interface (GUI) of Instant JChem for collecting data from publications, especially for the 2D structure collection. We then transferred the whole database into a relational MySQL database. Later modification or updating of the database was performed using the same approach. We checked the uniqueness of the compounds in *TollDB* using canonical isomeric SMILES (simplified molecular input line entry system) [143, 144] before transferring the database to MySQL.

Canonical SMILES are used as the unique string that encodes the connection table of a molecule, but with no chiral or isotopic information. Consequently, two stereoisomers always share the same canonical SMILES, since their stereo information was ignored during the canonicalization process. Canonical isomeric SMILES encodes isotopic and stereo information. Due to the unambiguity of canonical isomeric SMILES, they can be used as universal identifiers for a specific chemical structure (absolute SMILES [145]). Since we recorded the molecule structures as they were tested in publications, we conducted a preprocessing step using Molecular Operating Environment (MOE) version 20190102 [146] to remove the salts before transforming molecules to canonical isomeric SMILES by Open Babel [147, 148] and used them for duplicate checking in *TollDB*.

We used MOE to calculate all available 2D descriptors and some 3D descriptors (including ASA<sup>1</sup>, E<sub>sol</sub><sup>2</sup>, vol<sup>3</sup>, VSA<sup>4</sup>) for compounds in *TollDB*. Molecular quantum numbers (MQNs) [149] and molecular fingerprints were calculated using RDKit [150]. Descriptors are referred to as features in our machine learning studies described in Section 3.4.

## 3.2 Web deployment for *TollDB*

Before deployment, *TollDB* was transferred to MySQL Workbench [151]. The *TollDB* web interface was developed with Bootstrap [152], a popular responsive development framework including HTML, CSS and JavaScript, which enables integration of all data for user-friendly searching and visualization. Marvin JS [153], Smiles Drawer [154]

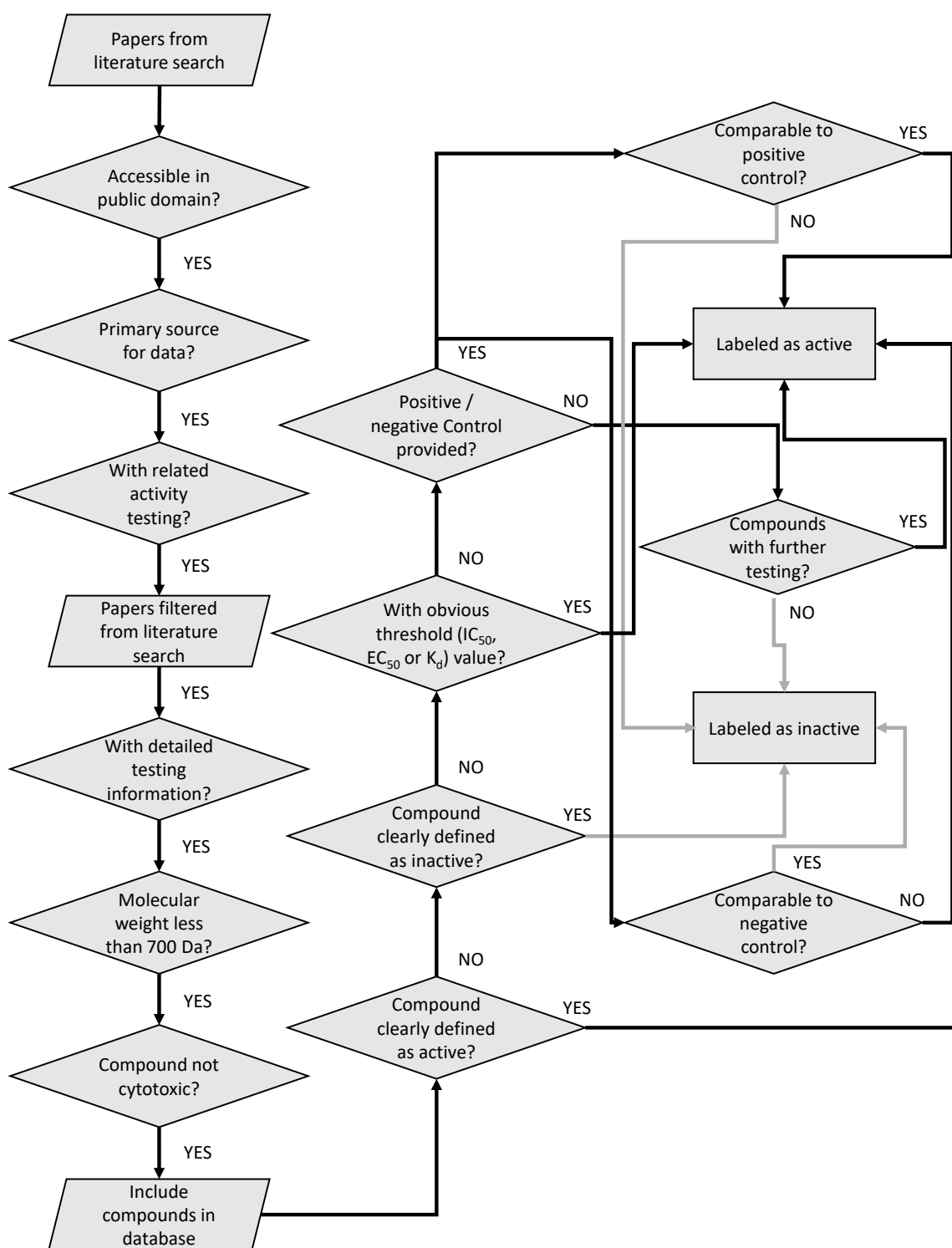
---

<sup>1</sup>ASA: Water accessible surface area

<sup>2</sup>E<sub>sol</sub>: solvation energy

<sup>3</sup>vol: Van der Waals volume

<sup>4</sup>VSA: Van der Waals surface area



**Figure 3.1:** Criteria for publication selection, compound selection and compound labeling in data collection process.

and RDKit were employed for structure search and 2D structure depiction on the web application. Python [155] with flask package was used for server-side scripting. The *TollDB* website is compatible with most major browsers.

### 3.3 Data analysis for *TollDB*

#### 3.3.1 Basic data analysis and visualization

Data analysis and visualization for *TollDB* were mainly conducted using Python with the numpy, pandas, matplotlib, and seaborn packages. The basic calculation included count number, quartile values, minimum value, maximum value, median value, skewness, kurtosis, variance and standard deviation of all descriptors for each ligand type.

For chemical space visualization we used principal component analysis (PCA) [156] and compared the chemical space between *TollDB* and *DrugBank*. Descriptors used for PCA analysis include: water accessible surface area (ASA)<sup>5</sup>, number of hydrogen-bond acceptors (a\_acc), number of hydrogen-bond donors (a\_don), weight (or molecular weight, MW), number of heavy atoms (a\_heavy), number of hydrophobic atoms (a\_hyd), fraction of rotatable bonds (b\_rotR), number of chiral centers (chiral), solvation energy (E\_sol), sum of formal charges (FCharge), *n*-octanol/water partition coefficient (log *P*(o/w)), number of rings (rings), Van der Waals volume (vol), Van der Waals surface area (VSA)<sup>6</sup>, topological polar surface area (TPSA). These were calculated with MOE.

#### 3.3.2 Matched molecular pairs and activity cliffs

Drug discovery projects frequently discover that a small structural change causes a major change to a property of interest. It is a central premise of medicinal chemistry that molecules that are structurally similar have similar biological activities [157]. In drug design, traditional medicinal chemists apply this in synthesis to modify hit compound expecting discovery of compounds with an improved desired property. Sometimes this similarity principle can fail, in which case digging into molecular similarity or diversity

---

<sup>5</sup>Water accessible surface area calculated using a radius of 1.4 Å for the water molecule. A polyhedral representation is used for each atom in calculating the surface area.

<sup>6</sup>van der Waals surface area. A polyhedral representation is used for each atom in calculating the surface area.

would be beneficial for the quantitative structure-activity relationship (QSAR) study of a particular target.

### Matched molecular pairs and activity cliffs search

Matched molecular pairs (MMPs) have generally been defined as “a pair of molecules that differ only by a particular, well-defined, structural transformation (represented by a substructure) such as a ring or an R-group” [158], and one of the key advantages of MMP analysis over other data analysis and modeling techniques is that it directly deals with the chemistry and measured data, ensuring clear interpretation of the results [159, 160]. Many researchers have developed software for identifying matched molecular pairs, such as MedChem Toolkit (from OpenEye) [161], mmpdb [162], Drug-Guru [163], WizePairZ [164], VAMMPIRE [165] and others [166]. The MMP concept has been further developed into Matched Pair Series [167, 168] or Matched Molecular Series (MMS) [169] to describe a set of compounds (not only a pair) differing by only a single chemical transformation.

In this work, we used the Automated Matched Pairs node from Erlwood knime open source cheminformatics in KNIME [170, 171] to detect and output all the possible matched molecular pairs in *TollDB*. Since the algorithm is not based on maximum common substructure (MCS) detection [172, 173], it is comparatively fast and efficient. [161, 174, 175] The workflow in the KNIME Analytics Platform is shown in Figure 3.2. Subsequently, we conducted docking studies trying to explain the activity difference between matched molecular pairs.

### Docking study for activity cliff pairs

Docking studies were carried out using GOLD (Version 5.8.1, Genetic Optimization for Ligand Docking, CCDC software, Cambridge, UK) [176]. The binding site was defined as all protein residues within 6 Å of the bound ligand, protein residues were kept rigid during docking and 15 diverse poses were generated for each ligand, using GoldScore as scoring function [176]. The crystal structure of TLR2/TLR1 with **Pam<sub>3</sub>CSK<sub>4</sub>** (PDB ID: 2Z7X) [177] was used for the docking study of TLR2 agonists, for TLR8 agonists the crystal structure of TLR8 with **CL097** (PDB ID: 3W3J) [178] was used, and the crystal structure of TLR8 with **CUCPT9b** (PDB ID: 5WYZ) [179] was



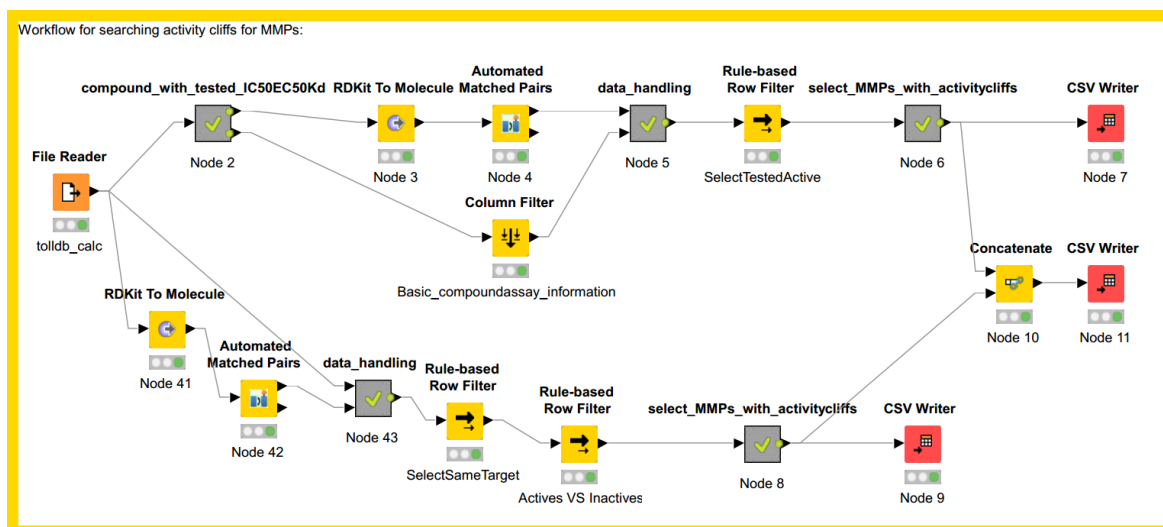


Figure 3.2: KNIME workflow applied for searching activity cliffs between MMPs.

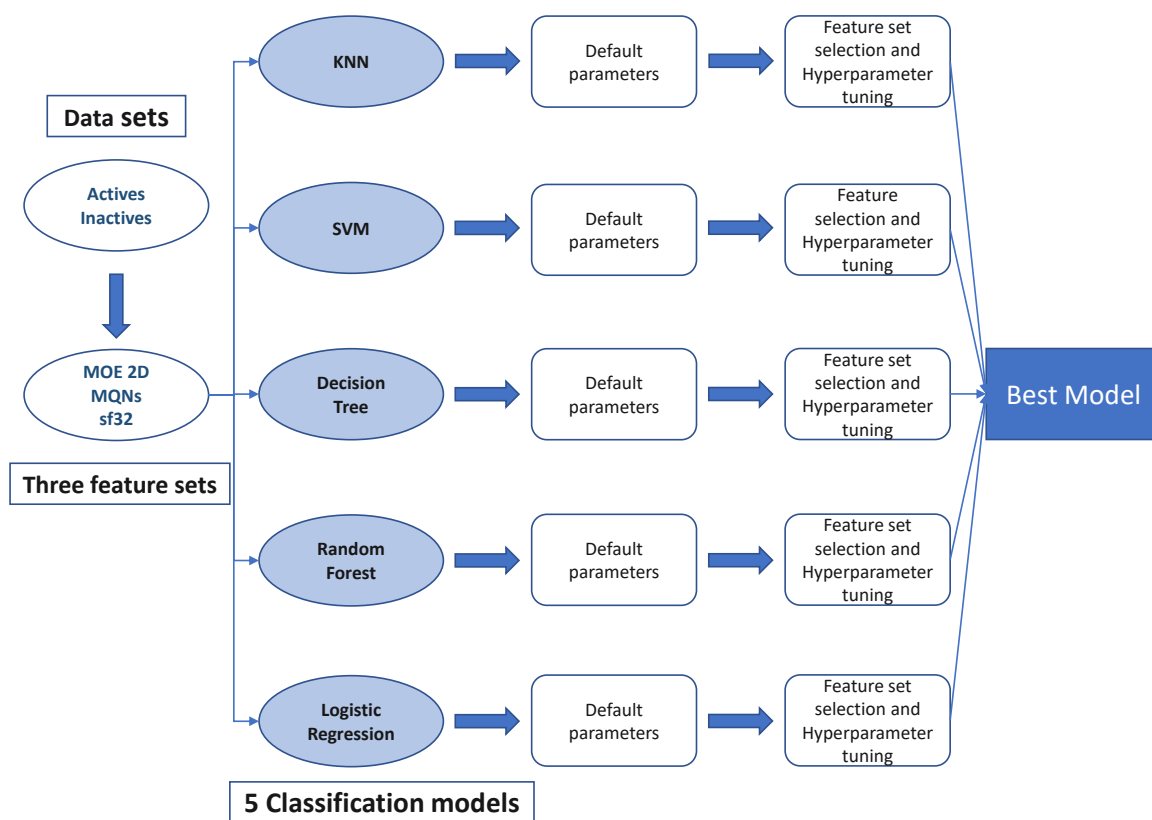
selected for the docking study of TLR8 antagonists. Docking poses were minimized using MMFF94 [180] force field and inspected in LigandScout (Version 4.2, Inte:ligand, Vienna, Austria) [181, 182].

### 3.4 Machine learning model development

The overall machine learning workflow is shown in Figure 3.3.

#### 3.4.1 Dataset preparation for machine learning models

We used MOE to calculate all available 2D descriptors and the open-source cheminformatics toolkit RDKit to calculate the MQNs for all compounds in *TolIDB*. Apart from these two feature sets, we also manually selected a feature set as a subset of the MOE 2D feature set (containing 206 features, later referred as *moe2d*). We selected the subset in the following manner: for the numerical features, we first removed all the constant features. Constant features are the type of feature that contains only one value for all the output of the instances in the dataset, which provides no information in discriminating different classes. We then removed all the constant and near-constant features (similar to constant feature, we consider that the features that have less than 5 unique values for all the output of all instances of the whole *TolIDB* database as near-constant features), and also removed those features that are mathematically closely related. For example, the MOE 2D feature set includes a series of descriptors like “Kier1”, “Kier2”,



**Figure 3.3:** Machine Learning workflow applied for a specific prediction task.

“Kier3”, “KierA1”, “KierA2”, “KierA3”, and “KierFlex”<sup>7</sup> [183, 184] and we only kept one of them (i.e., the “KierFlex”), removing all the others. In cases where two features had a correlation coefficient above 0.9, we removed one of the features of the pair. After that, the categorical features were carefully inspected and selected. These processes finally resulted in a feature set that contains 32 features (referred to as the sf32 feature set later).

In order to initiate the exploration of different machine learning algorithms, three different molecular descriptor sets; (1) all 2D descriptors available in MOE (referred to as moe2d feature set later), (2) 42 MQNs (referred to as MQNs feature set later), and (3) a manually selected subset from moe2d (i.e., sf32 feature set) were prepared as described previously. These three descriptor sets are referred to as feature sets later, and different descriptors are referred to as features. The purpose of feature selection was to choose the best feature set in terms of accuracy, speed, and computing space. Different datasets were selected for different models, for example, in predicting TLR2 agonism, all the compounds that have been tested towards TLR2 agonism were selected as the dataset for the TLR2 agonism prediction model. Data preprocessing was conducted for the dataset of each model before using it in training. The preprocessing steps include: (1) removing the features that have the same value in the whole data set, i.e., the constant features; (2) removing one of the numerical features from a pair that have a correlation coefficient with each other of above 0.9; (3) checking if there are distinct compounds that have the same values for all features but are labeled as a different class,

---

<sup>7</sup>KierFlex: Kier molecular flexibility index:  $(\text{KierA1})(\text{KierA2}) / n$ . The Kier and Hall chi connectivity indices are calculated from the heavy atom degree  $d_i$  (number of heavy neighbors) and  $v_i$ . The  $v_i$  are calculated using a connection table approximation. For a heavy atom  $i$  let  $v_i = (p_i - h_i) / (Z_i - p_i - 1)$  where  $p_i$  is the number of s and p valence electrons of atom  $i$ ,  $Z$  denotes the atomic number of an atom, lone pair pseudo-atoms (LP) are given an atomic number of 0. Heavy atoms are atoms that have an atomic number strictly greater than 1 (not H nor LP);  $h$  denotes the hydrogen count; the number of hydrogens an atom is (or should be) attached, this count includes all hydrogen atoms that are necessary to fill valence, a trivial atom is an LP pseudo-atom or a hydrogen with exactly one heavy neighbor;  $d$  denotes the heavy degree, which is the number of heavy atoms to which an atom is bonded. That is,  $d$  is the number of bonded neighbors of the atom in the hydrogen suppressed graph;  $n$  denotes the number of atoms in the hydrogen suppressed graph,  $m$  is the number of bonds in the hydrogen suppressed graph and  $a$  is the sum of  $(r_i/r_c - 1)$  where  $r_i$  is the covalent radius of atom  $i$ , and  $r_c$  is the covalent radius of a carbon atom. Also, let  $p_2$  denote the number of paths of length 2 and  $p_3$  the number of paths of length 3. Kier1 is first kappa shape index:  $(n - 1)^2/m^2$ ; Kier2 is second kappa shape index:  $(n - 1)^2/m^2$ ; Kier3 is third kappa shape index:  $(n - 1)(n - 3)^2/p_3^2$  for odd  $n$ , and  $(n - 3)(n - 2)^2/p_3^2$  for even  $n$ ; KierA1 is first alpha modified shape index:  $s(s - 1)^2/m^2$  where  $s = n + a$ ; KierA2 is second alpha modified shape index:  $s(s - 1)^2/m^2$  where  $s = n + a$ ; KierA3 is third alpha modified shape index:  $(n - 1)(n - 3)^2/p_3^2$  for odd  $n$ , and  $(n - 3)(n - 2)^2/p_3^2$  for even  $n$  where  $s = n + a$ .

and, if so, removing them, as this is due to the existence of isomers in the database; (4) converting all categorical features (if there are any) using “OneHotEncoder” in scikit-learn.

The quality of the training dataset determines the quality of the resulting machine learning models [185]. Missing or insufficient data, mislabeling of the target variable, and irrelevant features may complicate the learning task and hinder the performance of the trained models. The datasets selected from *TollDB* all contain validated data from reliable sources. Additionally, we applied methods to deal with missing values and checked the inconsistency of labeling (for compounds that have the same values for all feature sets, such as isomers). For filling up missing values with common strategies, scikit-learn provides a “SimpleImputer” [186, 187]. The four main strategies are the mean, the most frequent (mode), the median and the constant, which replace each attribute’s missing value with the corresponding mean (or mode or median or a constant) of that attribute. Before transforming the data, we determined whether there were ordinal or nominal features among the categorical features and transformed them using different transforming methods from scikit-learn, i.e., “OrdinalEncoder” for ordinal data, and “OneHotEncoder” for nominal data. “StandardScaler” was used when measured features were in different scales and did not contribute equally to the analysis, thus potentially creating a bias.

All compounds are labeled based on their testing results and stored in *TollDB*. For a specific model, for example, to build a model to predict if a compound shows TLR7 agonist activity, we selected all the compounds that were tested towards TLR7 agonistic activity. In this way, compounds that have been tested on other TLR subtypes at the same time are also selected.

### 3.4.2 Model selection and hyperparameter tuning

Scikit-learn is a Python module integrating a wide range of state-of-art machine learning algorithms for medium-scale supervised and unsupervised problems, and we used this module for all machine learning studies.

Initially, we explored the performance of five algorithms, K-Nearest Neighbors (kNNs), Logistic Regression (LR), Support Vector Machine (SVM, or SVC), Decision Tree (DT) and Random Forest (RF) classifiers on three feature sets (i.e., moe2d,

MQNs and sf32), using the default parameters for each algorithm. We then applied 10-fold cross-validation with grid search or random search for a carefully prepared hyperparameter searching space for each algorithm to try to get the best performance for each algorithm and each feature set. We then compared the accuracy, balanced accuracy, MCC score and AUC in order to determine the best feature set and the best algorithm for our models.

Since we aimed to find the optimal combination of hyperparameters for the given parameter searching space, both grid search with cross-validation and random search with cross-validation were used for the selected feature set and algorithm. With the number of combinations in hyperparameter searching space getting larger, we can see that random search performs better than grid search considering the comparable results and the relatively fast calculation. The random search method outweighs the grid search when there was a large number of hyperparameter combinations and when using more complex algorithms such as RF. Thus, random search is the preferred method in later studies due to its computational efficiency.

### **3.4.3 Model evaluation and validation**

The whole dataset selected for a specific prediction model was divided into two parts beforehand in a ratio of 4:1 using stratified sampling methods. We used stratified methods because in some models, the number of active and inactive compounds differs a lot, and stratified methods ensure that the relative active/inactive compound ratio is approximately preserved in training and testing set. 80% of the data from the selected dataset was used for training and the remaining 20% of the data was used as a testing set to calculate the final performance of the model.

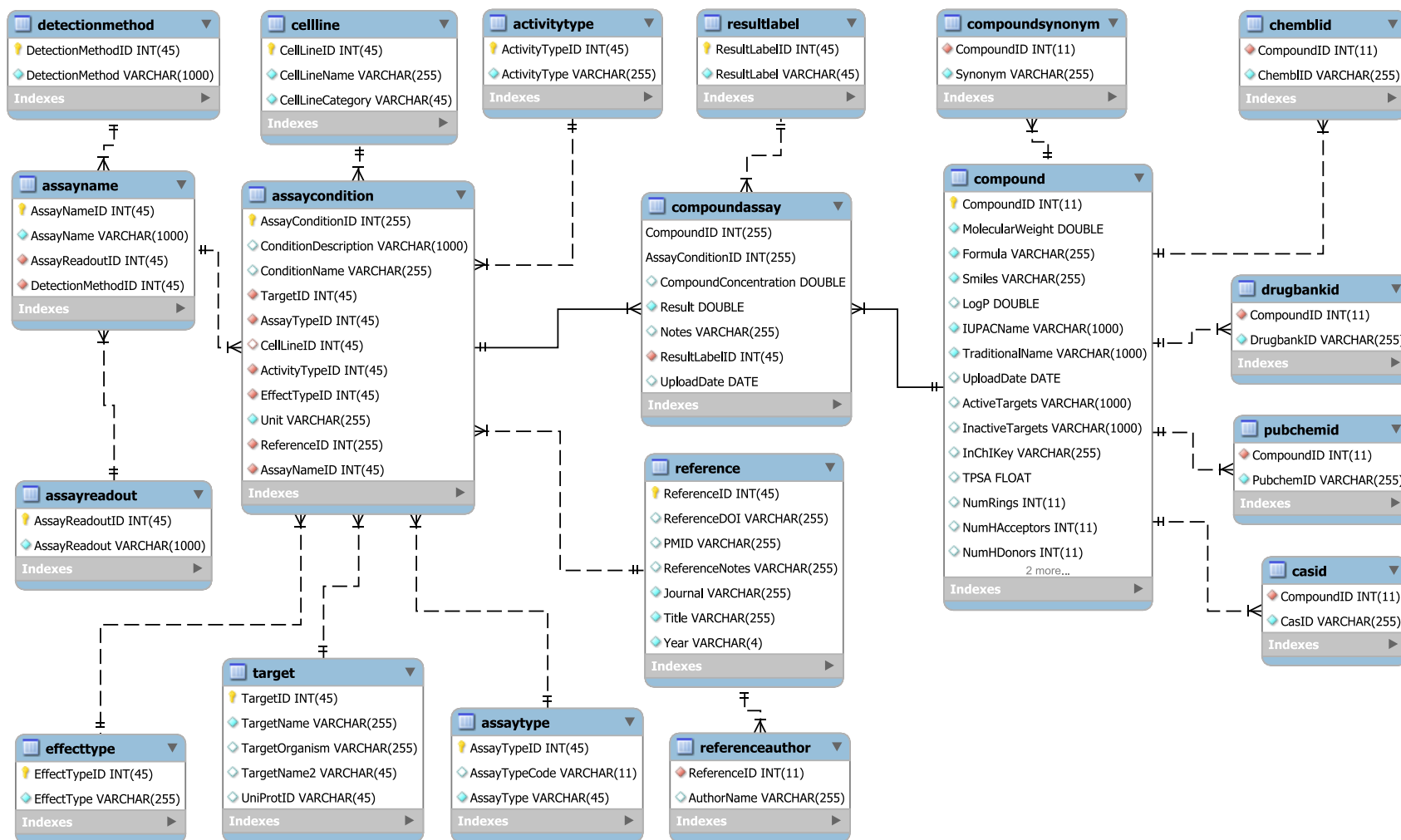


## 4. Results

### 4.1 Database information

#### 4.1.1 Database schema

*TollDB* was developed as a local MySQL database; the database schema is shown in Figure 4.1. MySQL is an open source relational database management system (RDBMS) based on structured query language (SQL). The main tables in *TollDB* comprise *compound*, *assaycondition*, *compoundassay* and *reference*. The *compound* table shows general information about the molecules, with distinct compounds as entries in it; the *assaycondition* table includes basic biological testing conditions; the *compoundassay* table includes testing results for compounds, using a combination of *CompoundID* and *AssayConditionID* as the primary key to ensure uniqueness; the *reference* table provides the possibility for tracing back to the original articles.

Figure 4.1: Database schema for *TollDB*.



### 4.1.2 Web application

A web application was developed for *TollDB*, which can be accessed at <https://tolldb.drug-design.de/> and provides a user-friendly interface for convenient data searching and browsing. There are mainly three searching possibilities: “Simple Search”, “Advanced Search” and “Structure Search”.

#### Simple Search

All the distinct TLR ligands are shown by clicking at the “Show all data” button on the right side below the query builder. Users can use the query builder at the top to retrieve specific TLR ligands from the database.

The query builder takes several distinct query parameters. Users can specify:

- specific TLR target
- host organism of target TLR
- tested effect type (agonism or antagonism)
- molecular weight
- $\log P$


By clicking “Show/hide example queries”, the dropdown list of default query examples will be shown. A click on an example query will automatically fill the query field accordingly. “Run query” on the right side below the query builder will perform a search. Users can query with a single query parameter or with a conjunction of the query parameters. Query parameters can be grouped and connected by Boolean operators “AND” and “OR”. To add further parameters or create a group, “Add rule” and “Add group” can be used. “AND” will return the intersection between compounds queried by specified parameters (or group of parameters), while “OR” will return the union. Note that TLR2 refers to cases where the authors worked with the TLR2 monomer specifically, or when the authors did not specify whether TLR2/1 or TLR2/6 heterodimer was used.






The “Run query” button can be clicked after the query mask was updated. In response to the query parameter provided, a table of all matching molecules will be

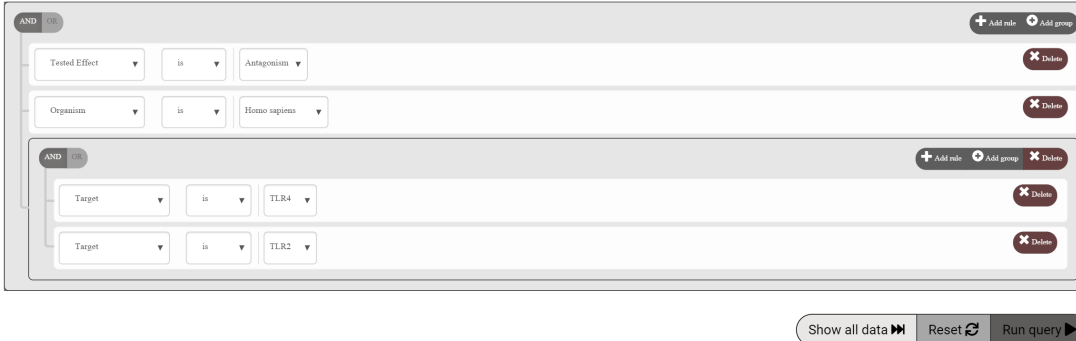
shown. When multiple fields are used for creating a query, the results will show molecules that satisfy all conditions. A screenshot of this search page is shown in Figure 4.2.

**Simple Search**  
Enables the exploration of ligand properties.  
(Use [Advanced Search](#) for retrieving bioactivity data and assay properties)

Show/hide example queries

The following example queries were pre-created to explain the power of the search mask below. Just click a link to automatically fill the query mask accordingly. Clicking on the  next to the query shows the rationale for performing the respective example search. Please click the "Run Query" button after the query mask was updated.

-  Show all human TLR7 agonists. 
-  Show all human dual TLR2/TLR4 antagonists. 
-  Show TLR3/TLR7/TLR8/TLR9 ligands with defined log P. 



[Show all data](#) [Reset](#) [Run query](#)

**Figure 4.2:** Screenshot of example query for Simple Search.

## Advanced Search

The query builder at the top of the “Advanced Search” page allows the user to retrieve the bioactivity data for all reported molecules tested against different TLRs from the database. It must be kept in mind that “Simple Search” only contains ligands tested as actives against different TLRs, while “Advanced Search” additionally contains inactive molecules. The query builder takes several distinct query parameters. Apart from those queries that users can specify in “Simple Search”, users can specify additional queries as follows:

- flag for activity (active or inactive)
- tested effect type (agonist or antagonist)
- assay type (functional or binding)
- activity type (how the activity is expressed, e.g.,  $IC_{50}$ ,  $EC_{50}$ ,  $K_d$ , relative activation or inhibition rate)


- cell line used for testing
- cell line category used for testing (primary cells or immortalized cells)
- activity value (when activity type set to IC<sub>50</sub>, EC<sub>50</sub>, K<sub>d</sub>, users can set the activity value range or threshold)
- result unit (unit for the activity value)






The basic rules for the query builder on the “Advanced Search” page are the same as on the “Simple Search” page. For more examples, please refer to the query examples below “Advanced Search”. The results table displays information related to bioactivity data of retrieved molecules from experiments that satisfy the input query.


Additionally, users can access information about a specific compound or assay by clicking on Structure or Assay Name, respectively. This will navigate to a new page with a basic overview of the compound or overview of compounds tested under the chosen Assay Name. A screenshot for the “Advanced Search” page is shown in Figure 4.3.

**Advanced Search**  
Explore bioactivity data and assays.

Show/hide example queries

The following example queries were pre-created to explain the power of the search mask below. Just click a link to automatically fill the query mask accordingly. Clicking on the  next to the query shows the rationale for performing the respective example search. Please click the “Run Query” button after the query mask was updated.

- [Show tested inactives toward TLR2.](#) 
- [Show compounds with tested binding affinity towards human TLR2.](#) 
- [Show antagonists tested towards TLR2 in human primary cell lines.](#) 
- [Show agonists with reported EC50 value in HEK293-hTLR7.](#) 
- [Show human TLR ligand with a defined EC50, IC50 or Kd value that is less than 10000 nM.](#) 



Activity Label is Active

Tested Effect is Agonism

Activity Type is EC50

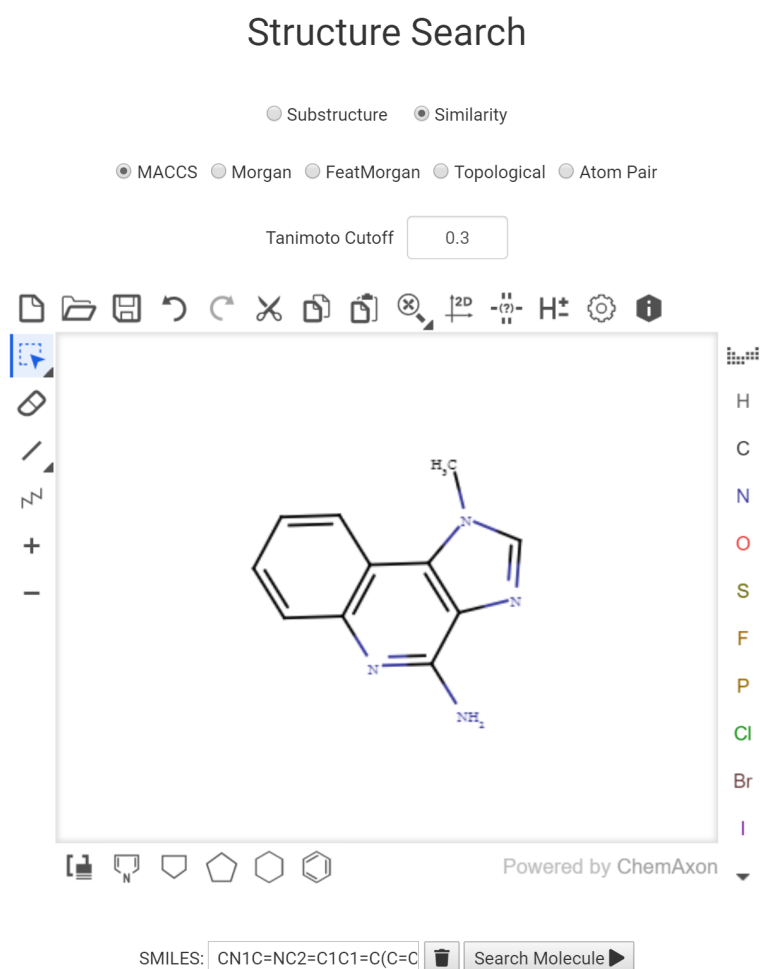
Cell Line is HEK293-hTLR7

**Figure 4.3:** Screenshot of example query for Advanced Search.

## Structure Search

“Structure Search” includes “Substructure Search” and “Similarity Search”. “Substructure Search” allows searching for a specified substructure. Enabling “Similarity Search”

shows the fingerprints used for calculating the Tanimoto similarity [188] between molecules. Users can choose which fingerprints to use for similarity calculation and even specify a threshold for the Tanimoto similarity for filtering the result table. Users can either input SMILES, upload the molecular structure file, or draw the molecule. Similarity is expressed as the Tanimoto similarity between specified fingerprints generated from respective SMILES. A screenshot of the “Structure Search” page is shown in Figure 4.4.



**Figure 4.4:** Screenshot of example query for Structure Search.

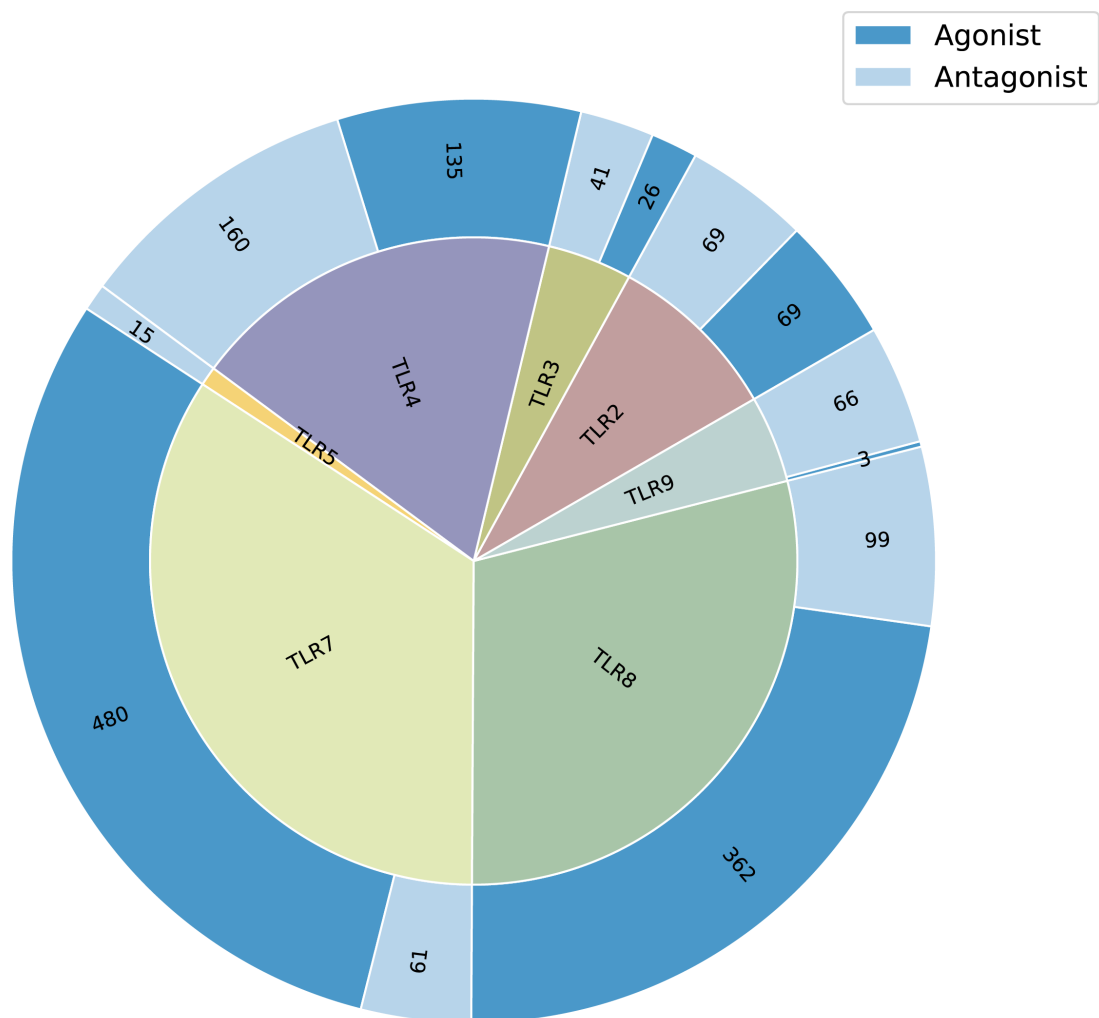
## 4.2 Basic statistical analysis for *TollDB*

The current version of *TollDB* contains a total of 2155 compounds, of which 1329 are tested active towards TLR targets and the remaining are tested inactive. Of all

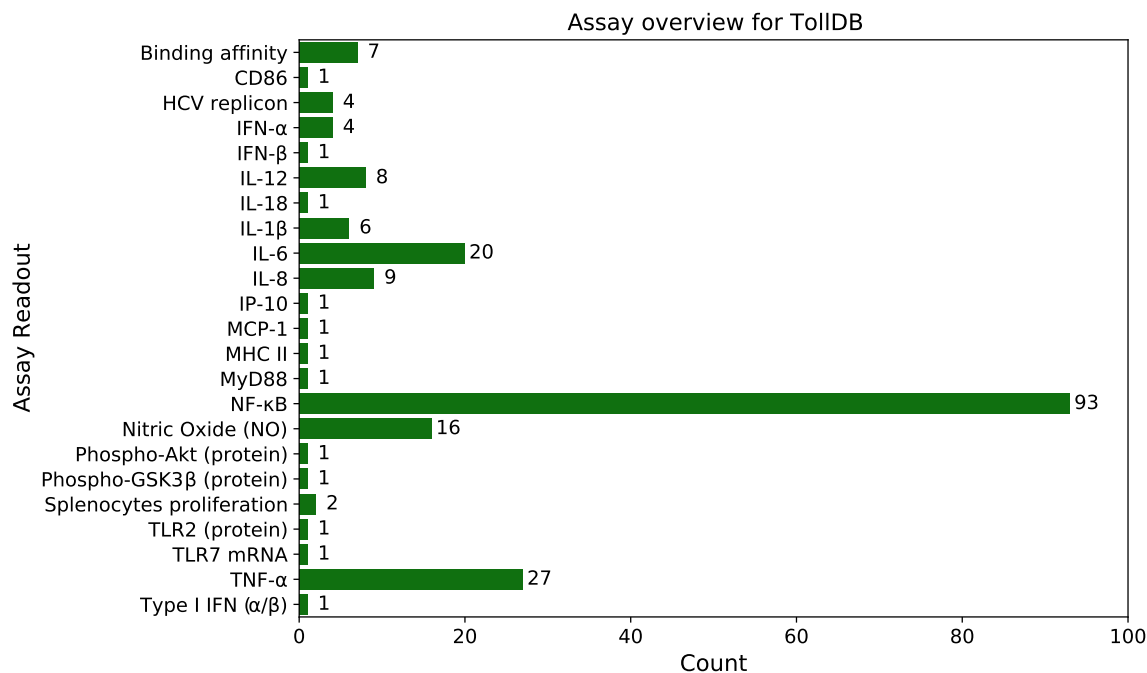
the active compounds, 861 are agonists and 468 are antagonists. There are also 248 compounds that have multi-TLR effects, including five dual TLR2/TLR4 antagonists, two TLR2/7/8 antagonists, two TLR3/TLR4 antagonists, two TLR3/TLR5 antagonists, one TLR3/8/9 agonist, two TLR4/7/8/9 antagonists, three TLR4/TLR9 antagonists, 212 dual TLR7/TLR8 agonists, 19 dual TLR7/TLR8 antagonists, ten TLR7/TLR9 antagonists, two TLR4/7/8/9 antagonists and one TLR2/3/4/5/7/8/9 antagonist. The number of active molecules targeting each TLR subtype is shown in Figure 4.5. As shown on Figure 4.5, the number of discovered agonists of TLR7 and TLR8 differs a lot compared to their corresponding antagonists, and so far there is neither a reported TLR5 agonist nor a TLR10 ligand, thus, TLR5 and TLR10 were not considered for later analysis. Judging by the discovered ligand number for each TLR subtype, research interest mainly focuses on TLR2, TLR4, TLR7 and TLR8.

Compounds were tested in 36 assay types using 553 conditions with a total number of 4921 datapoints. Of the 36 assay types, 32 are functional assays and 4 are binding assays. For functional assays, the main categories comprise reporter gene related assays, cytokine related assays, mRNA related assays, transcription factor related assays and cell proliferation related assays. The methods used for binding assays are mainly surface plasmon resonance (SPR), isothermal titration calorimetry (ITC) and fluorescence polarization. The overall information about the assays shows in Figure 4.6. Binding assays were rarely used in testing and most of the biological testing involves functional testing for activation or inhibition of cytokines or cytokine-related proteins, or mRNA expression. A gene reporter assay for NF- $\kappa$ B was the most frequently used testing method for determining TLR-related activity.

Although TLR modulators can be found in current databases such as *ChEMBL* [189, 190], *BindingDB* [191], *PRRDB* [192, 193], *ImmtorLig.DB* [194] or the IUPHAR/BPS Guide to Pharmacology [195], there is only a small overlap with *TollDB*. No other database that contains all known drug-like TLR ligands or focuses primarily on small TLR modulating compounds is available. We retrieved all molecules tested for activity against TLRs from *ChEMBL* version 25 (accessed on 10/15/2019) [196] that have an MW below 700 Da and were published in peer-reviewed journals. The total number of distinct compounds is 881 and all of these compounds were included in *TollDB*. Moreover, *TollDB* contains more recently published small organic compounds and,



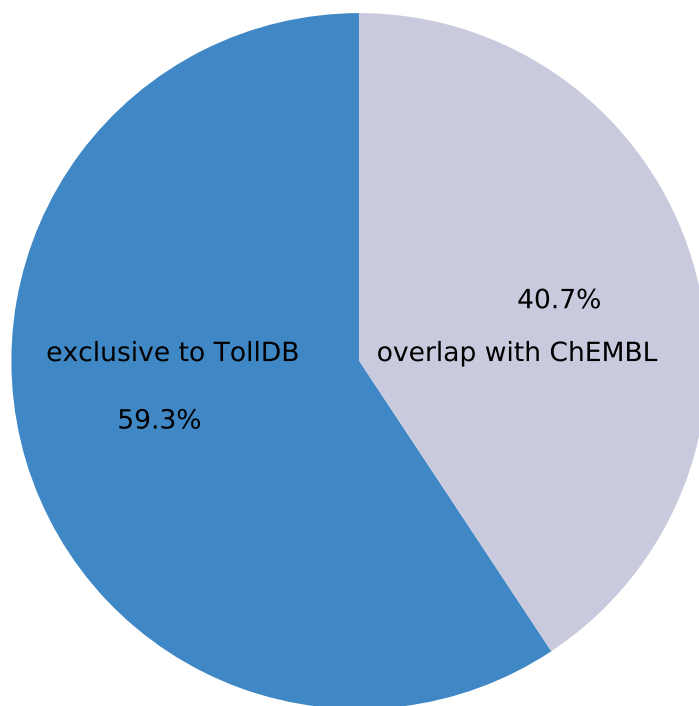
**Figure 4.5:** Number of active molecules targeting each TLR subtype. Multi-target molecules are included in all the subtypes that they belong to. The outer shell shows the ligand distribution according to the function.



**Figure 4.6:** Assay condition number count for the same assay readout.

compared with *ChEMBL*, 1278 compounds (about 59.3%) were exclusively reported in *TolIDB*, as shown in Figure 4.7.

To get an overview of the compounds in *TolIDB* for each activity type, we first compared the distribution of two key physicochemical properties; molecular weight (MW) and *n*-octanol/water partition coefficient ( $\log P$ ) between agonists and antagonists at different TLR subtypes (shown in Figure 4.8). Figure 4.8A shows that the mean MW of TLR antagonists is a bit smaller than that of TLR agonists ( $p \leq 0.001$ ), and this trend is recapitulated at nearly all subtypes except for TLR8 and TLR9. The *p*-values for the difference in MW between agonists and antagonists for TLR2, TLR3, TLR4 are all less than 0.001, and for TLR7 the *p*-value is less than 0.05. If we distinguish the dual TLR7/TLR8 ligand, these differences are shown in Figure 4.9A, which distinguishes dual TLR7/TLR8 ligands from pure TLR7 and 8 ligands. The mean  $\log P$  values of TLR agonists and TLR antagonists are almost the same ( $p > 0.05$ ), see Figure 4.8B. However, for agonists and antagonists of different TLR subtypes, as exemplified by the TLR2, 3, and 4 ligands, the agonists usually show a higher  $\log P$  value than their corresponding antagonists, with *p*-values all less than 0.001. For TLR7, if we exclude the large number of dual functioning compounds (shown in Figure 4.9B), this trend is also apparent (i.e., the  $\log P$  value for pure TLR7 agonists is greater than pure TLR7



TollDB compound number: 2155

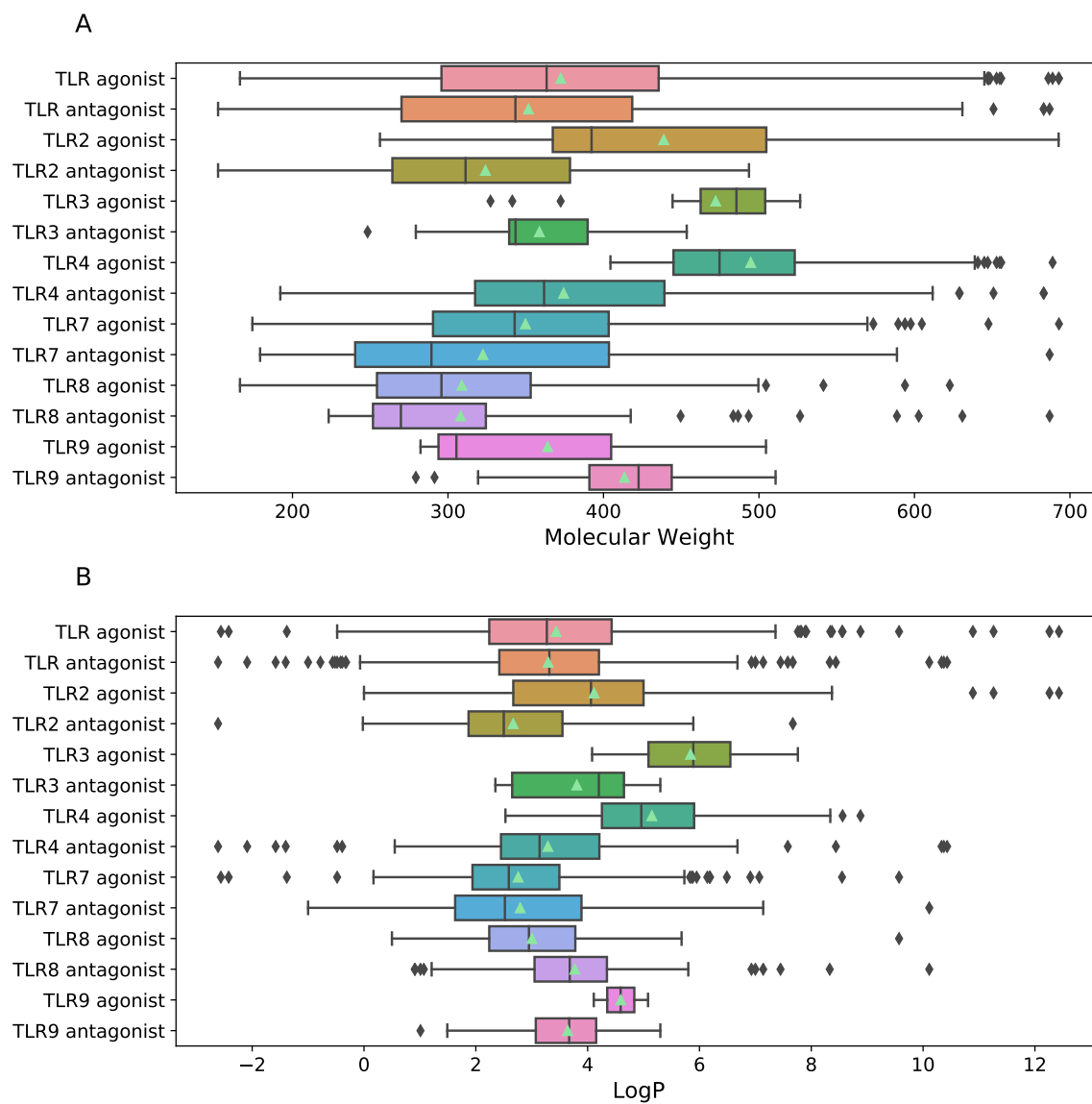
**Figure 4.7:** Composition of *TollDB* with comparison to the portion overlapping with *ChEMBL*.

antagonists).

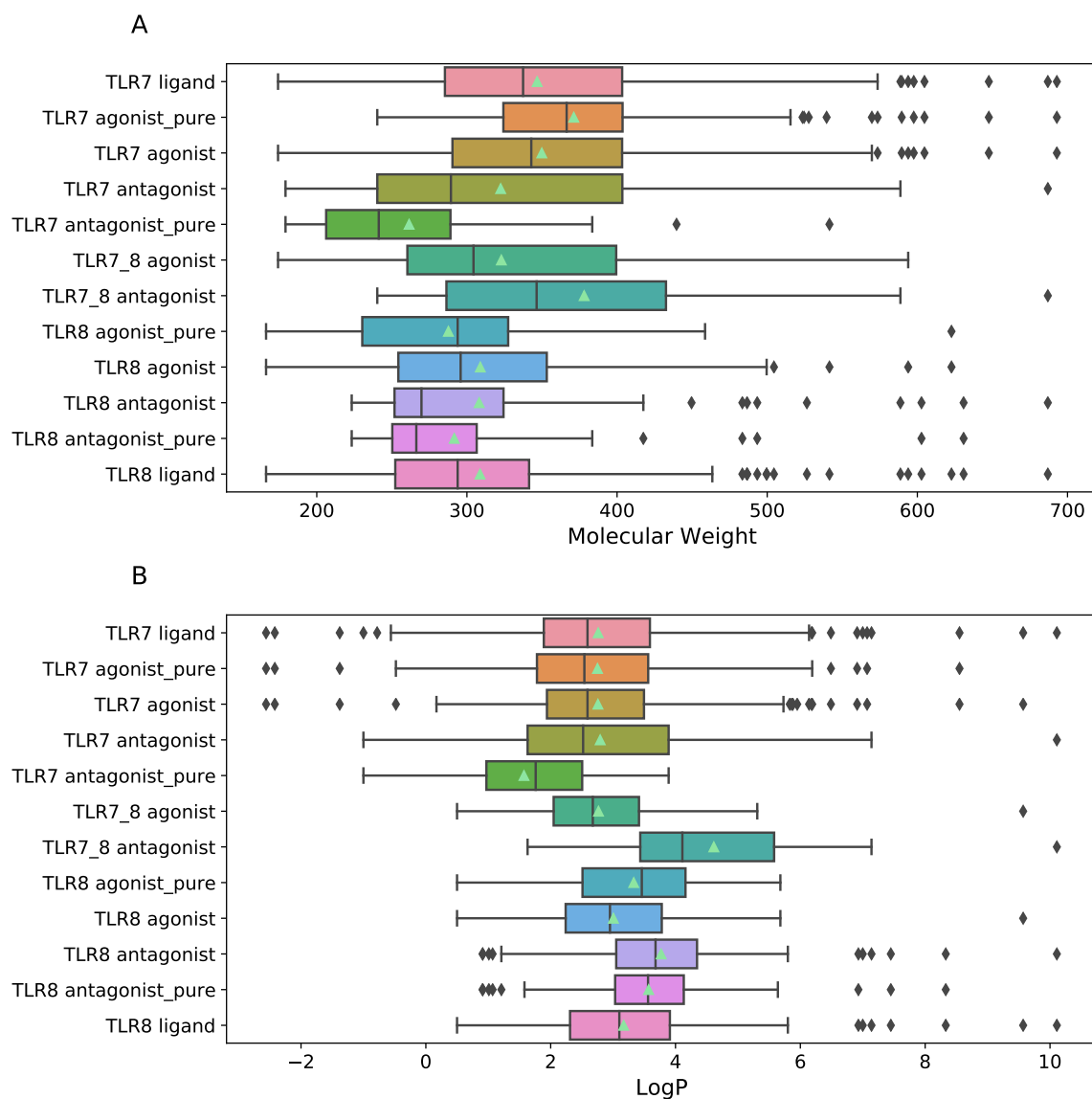
The number of compounds targeting each TLR subtype featuring a certain number of chiral atoms is shown in Figure 4.10. It is evident that most TLR ligands contain fewer than three chiral atoms. TLR4 has the biggest variety in ligand chiral atom count. After a detailed examination of TLR4 ligand structures, it is apparent that this is mainly due to series of compounds featuring a core structure containing multiple chiral centers, such as pyranose derivative-containing compounds, or other examples of multiple chiral centers, such as monocyclic or fused-ring compounds.

The polar surface area (PSA) is defined as the sum of the surface areas of polar atoms in a molecule, which correlates to membrane transportability of drugs [197]. In 2000, Ertl et al. [198] used the sum of PSA values of polar fragments of a molecule to define the topological polar surface area (TPSA) index. TPSA is now a widely used physicochemical property for the prediction of molecular transport properties [199]. The distribution of TPSA for TLR subtypes are shown in Figure 4.11. As shown in Figure 4.11A, the TPSA of most TLR ligands is in the range of 60-120 Å<sup>2</sup>. We also can

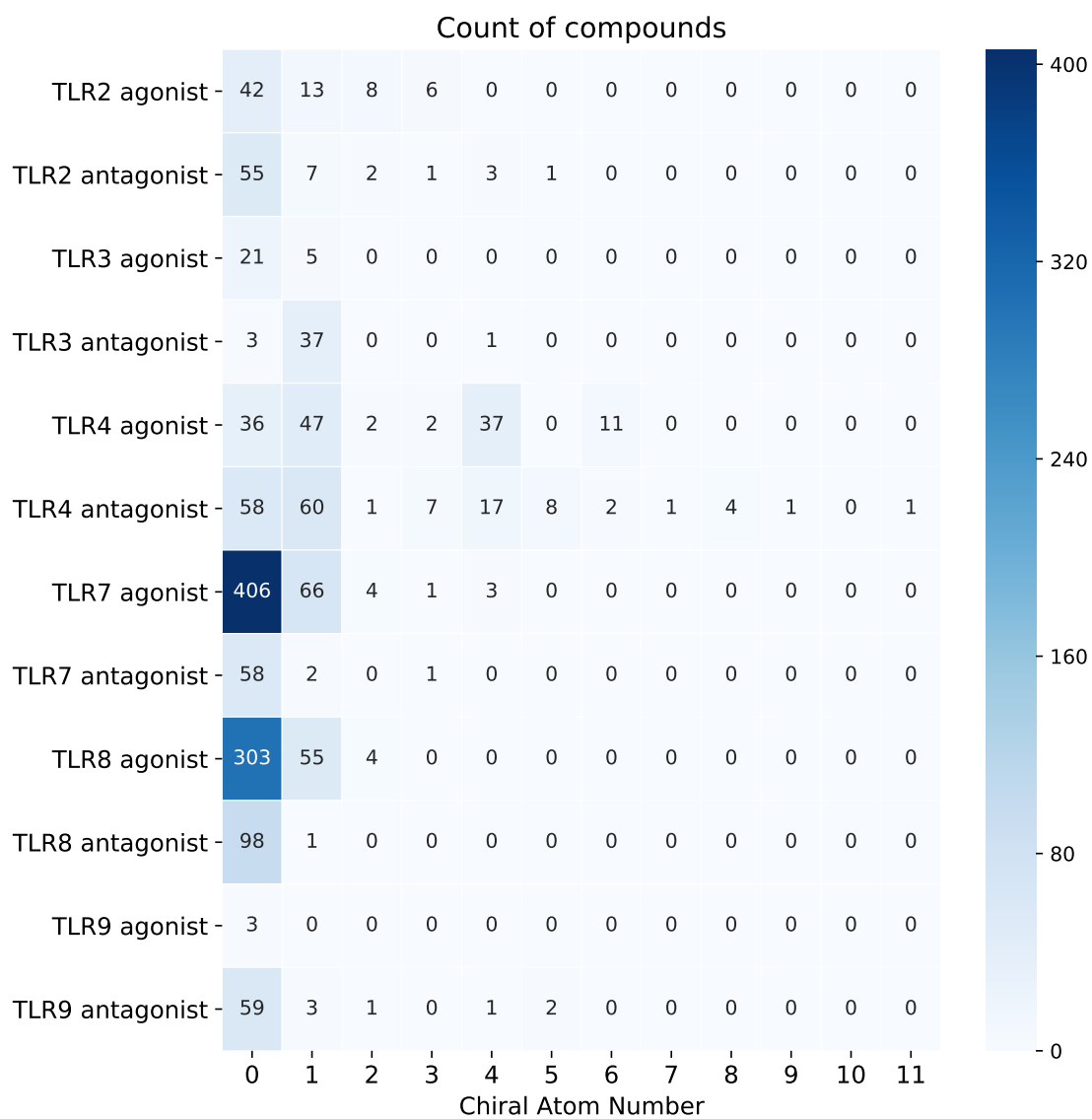




**Figure 4.8:** Box plot illustrating the distribution of (A) MW and (B) log  $P$  for different types of TLR ligands. Green triangles indicate the mean value.

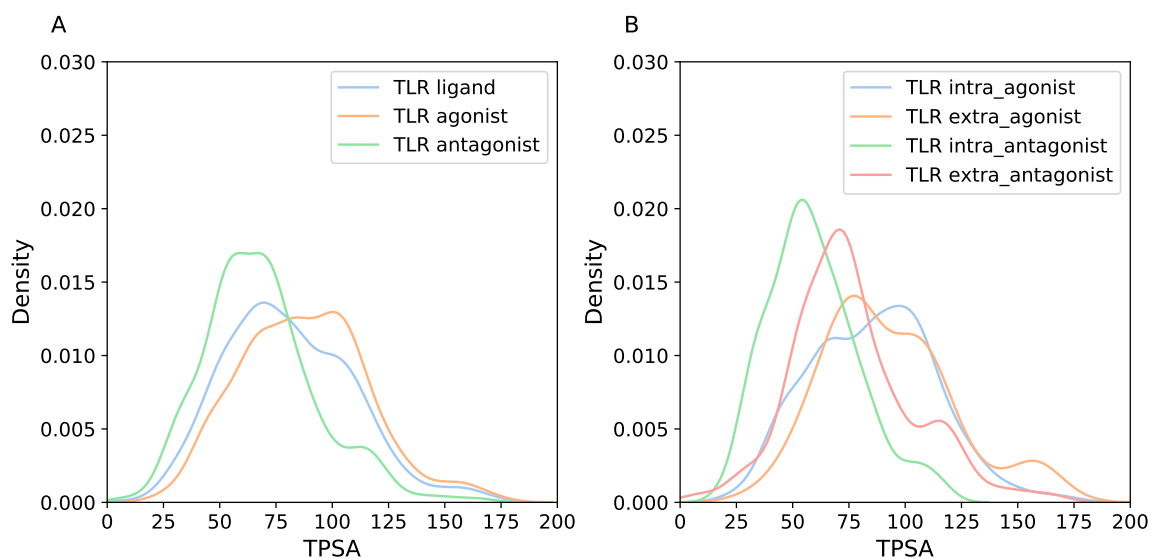


**Figure 4.9:** Box plot illustrating the distribution of (A) MW and (B) log  $P$  for TLR7 ligands (compounds that activate or inhibit TLR7), TLR7 agonist\_pure (compounds tested on TLR7 and TLR8 but turn out to be specifically activate on TLR7), TLR7 agonist (compounds that can activate TLR7), TLR7 antagonist (compounds that can inhibit TLR7), TLR7 antagonist\_pure (compounds tested towards TLR7 and TLR8 that specifically inhibit TLR7), TLR7\_8 agonist (compounds that can activate both TLR7 and TLR8), TLR7\_8 antagonist (compounds that can inhibit both TLR7 and TLR8), TLR8 agonist\_pure (compounds tested towards TLR7 and TLR8 but turn out to be specifically activate on TLR8), TLR8 agonist (compounds that can activate TLR8), TLR8 antagonist (compounds that can inhibit TLR8), TLR8 antagonist\_pure (compounds tested towards TLR7 and TLR8 that specifically inhibit TLR8) and TLR8 ligand (compounds that activate or inhibit TLR8). Green triangles indicate the mean value.



**Figure 4.10:** Compound count for TLR subtypes with different numbers of chiral atoms.

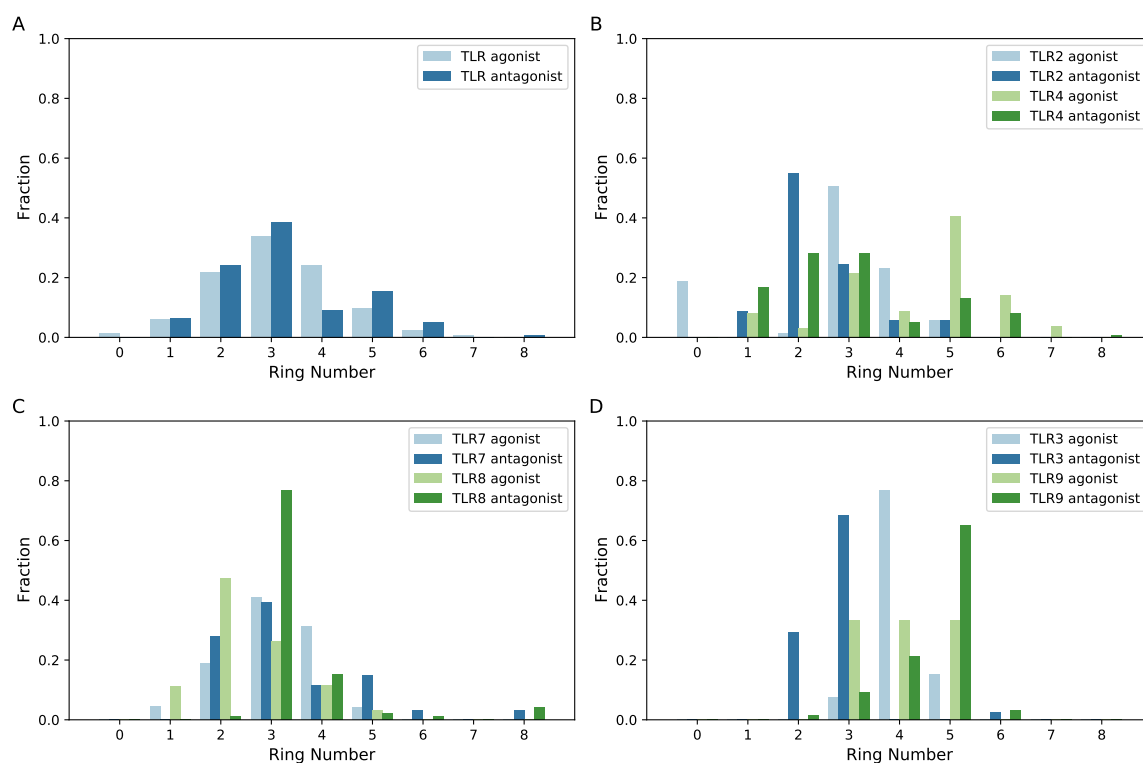
see from Figure 4.11B, that for TLR intracellular antagonists, the TPSA displays a clear left shift compared to TLR extracellular antagonists ( $p \leq 0.001$ ). This is reasonable since molecules with a smaller TPSA are more lipophilic, and thus more likely to permeate cell membranes. For agonists, the TPSA distribution between intracellular and extracellular ligands does not differ much ( $p = 0.002 > 0.001$  if considering the same significance level).



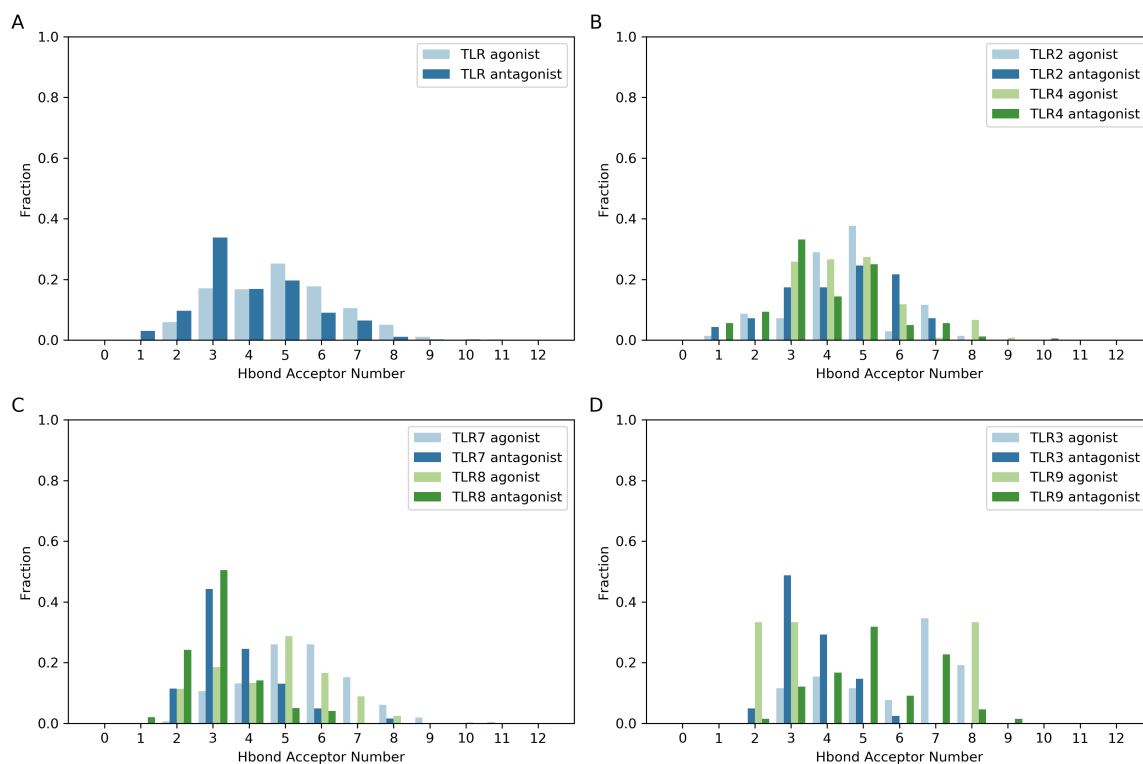
**Figure 4.11:** TPSA distribution of (A) TLR ligands, TLR agonists and TLR antagonists and (B) TPSA distribution of intracellular agonists (including all TLR3, TLR7, TLR8 and TLR9 agonists), extracellular agonists (including TLR2 and TLR4 agonists), intracellular antagonists (including all TLR3, TLR7, TLR8 and TLR9 antagonists) and extracellular antagonists (including TLR2 and TLR4 antagonists).

The distribution of number of rings for ligands of different TLR subtypes is shown in Figure 4.12. As shown in Figure 4.12A, most TLR ligands contain 2-5 rings. For different TLR subtypes, the ring number distribution does not differ much (as shown in Figure 4.12B-D).

H-bond donors and H-bond acceptors are commonly used molecular descriptors. The distribution of H-bond acceptors for ligands of different TLR subtypes is shown in Figure 4.13. As we can see in Figure 4.13A, most TLR ligands have between three to six H-bond acceptors. The number of H-bond acceptors do not differ much for different TLR subtypes (TLR2, TLR4, TLR3, TLR9) or between agonists and antagonists (see Figure 4.13(B,D)). For TLR7 and TLR8, agonists have a slightly higher number of H-bond acceptors than antagonists (see Figure 4.13C).

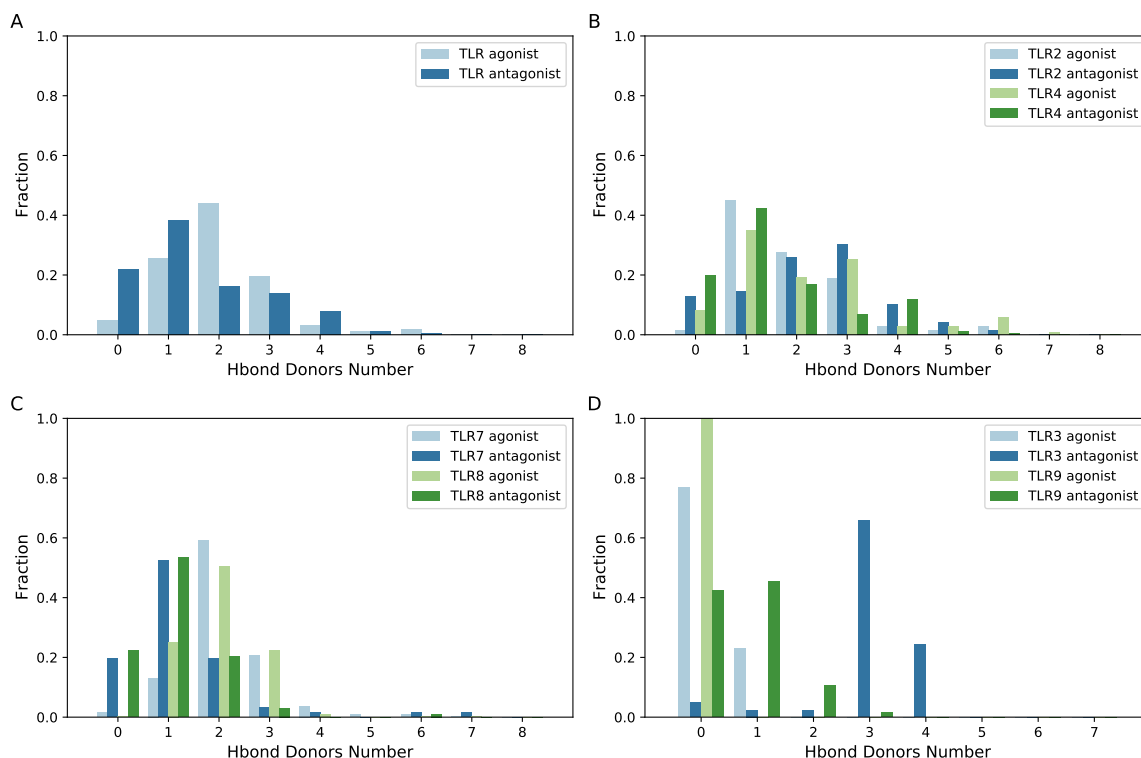


**Figure 4.12:** Ring number distribution for ligands of different TLR subtypes.



**Figure 4.13:** Distribution of H-bond acceptors for TLR subtypes.

The distribution of H-bond donors for ligands of different TLR subtypes is shown in Figure 4.14. As we can see in Figure 4.14A, most TLR ligands have between one to three H-bond donors. The number of H-bond donors do not differ much for different TLR subtypes (TLR2, TLR4, TLR3, TLR9) or between agonists and antagonists (see Figure 4.14(B,D)). For TLR7 and TLR8, agonists have a slightly higher number of H-bond donors than antagonists (see Figure 4.14C).



**Figure 4.14:** Distribution of H-bond donors for TLR subtypes.

Studies have shown that reduced molecular flexibility, as measured by the number of rotatable bonds, and low polar surface area or total hydrogen bond count (sum of donors and acceptors) are important predictors of good oral bioavailability [200]. The distribution of rotatable bonds for TLR subtypes is shown in Figure 4.15. As shown in Figure 4.15A, most TLR ligands have fewer than 10 rotatable bonds. For TLR2, TLR4, TLR3 and TLR9, the number of the rotatable bonds of agonists and antagonists shows almost the same distribution (see Figure 4.15(B,D)). Interestingly, TLR7 and TLR8 antagonists show fewer rotatable bonds than their corresponding agonists (see Figure 4.15C).

The distribution of ASA for TLR subtypes is shown in Figure 4.16. Figures 4.16A-C show that the ASA of TLR agonists and antagonists follows the same trend. For

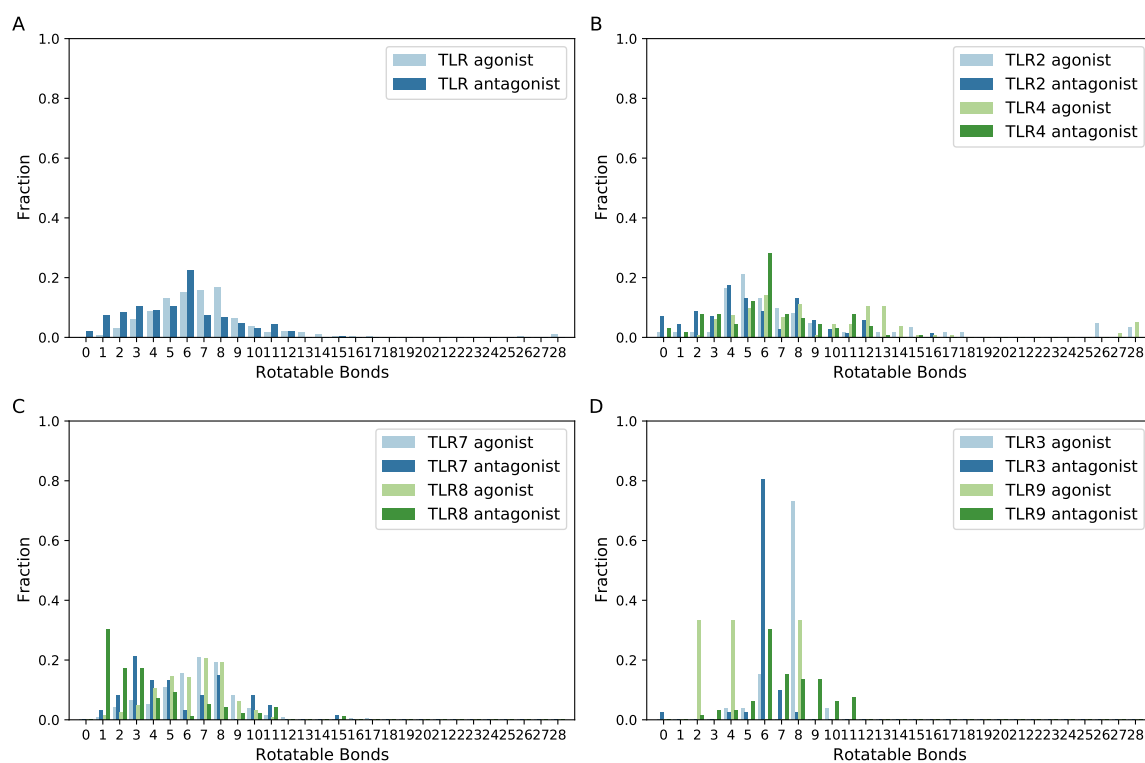


Figure 4.15: Distribution of rotatable bonds for TLR subtypes.

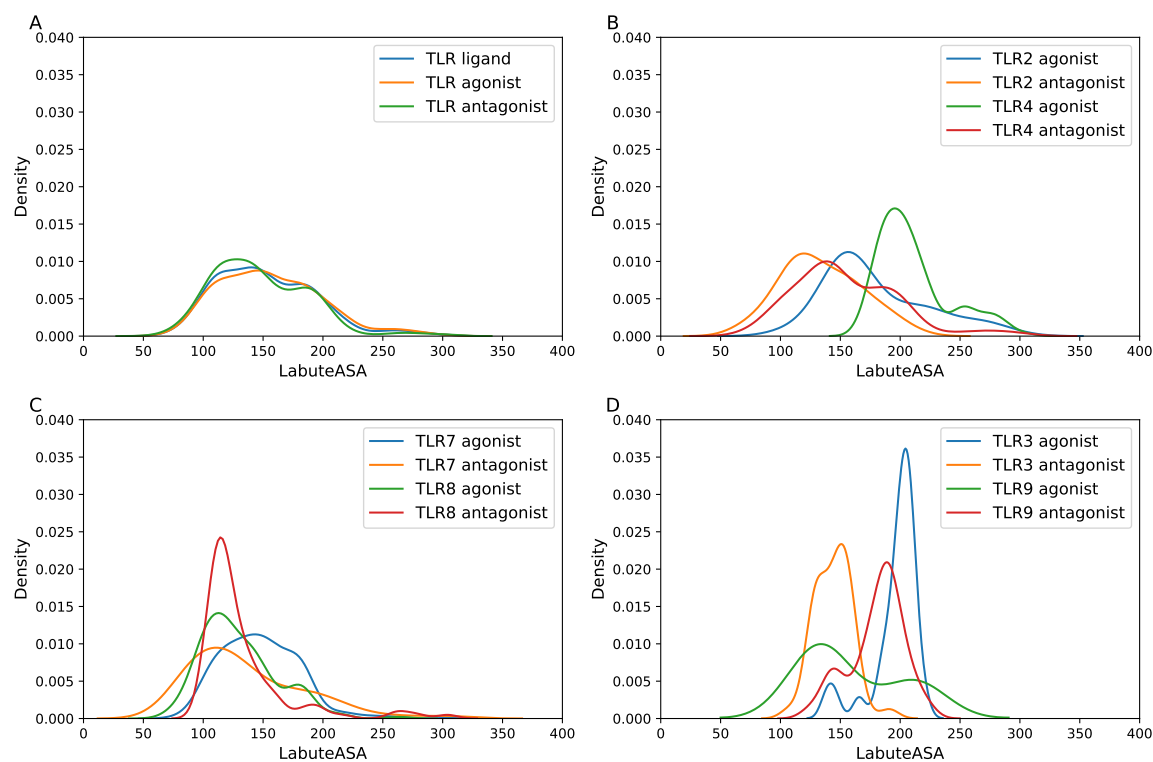
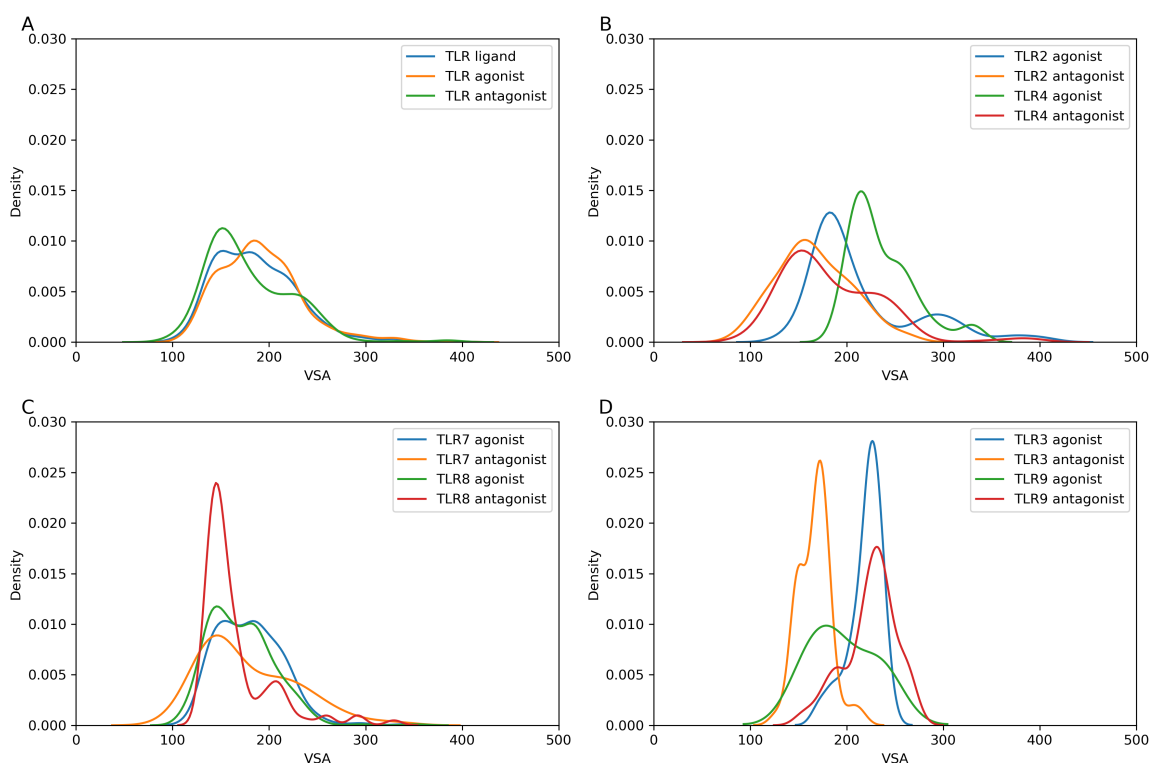


Figure 4.16: Distribution of ASA for TLR subtypes.

TLR2 and TLR4, the ASA of their agonists is shifted to the right compared to their corresponding antagonist. TLR7 ligands show the same tendency as TLR2 and TLR4 ligands, while the ASA of TLR8 ligands (agonists and antagonists) do not show the same trend; TLR8 agonists and antagonists have nearly the same distribution.

The distribution of VSA [201] for TLR subtypes is shown in Figure 4.17. For TLR2, TLR3, TLR4 and TLR9, agonists have a larger VSA value than antagonists but for TLR7 and TLR8, the distribution between agonists and antagonists is nearly the same.



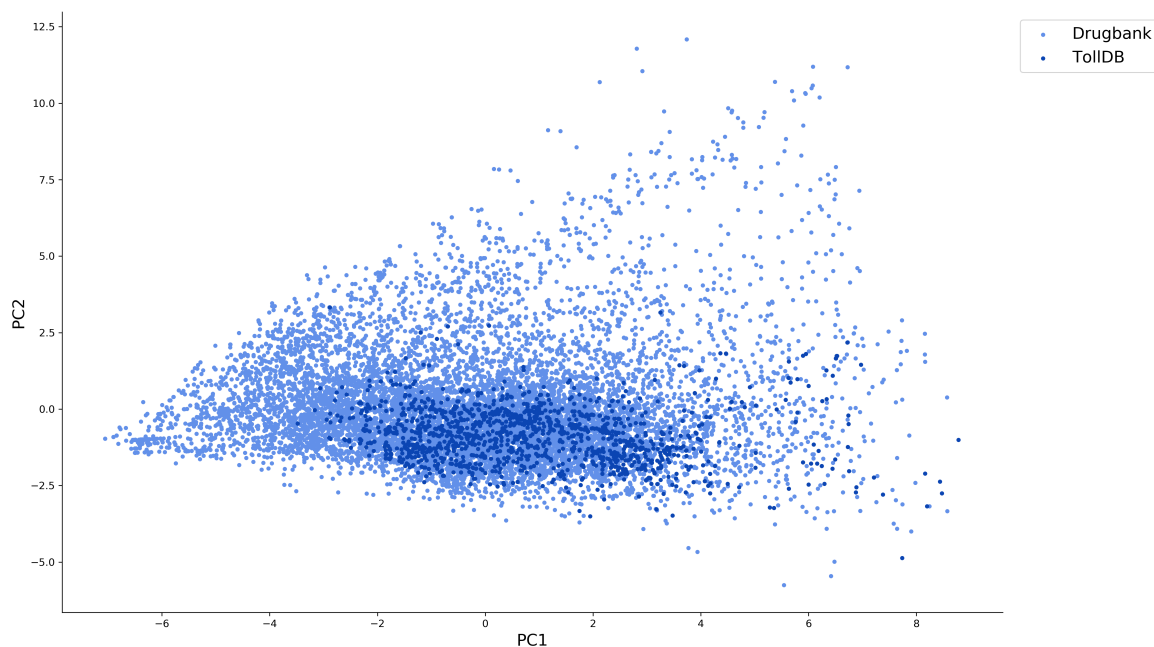
**Figure 4.17:** Distribution of VSA for TLR subtypes.

### 4.3 Chemical space analysis for *TollDB*

In order to determine the chemical space covered by *TollDB*, we compared it with that of molecules from *DrugBank* [202], a database of drugs and drug candidates. PCA based on 15 relevant physicochemical properties was applied, and Table 4.1 provides detailed information on used descriptors. A total of 9998 molecules from *DrugBank* with an MW of less than 700 Da were taken into consideration, including about 94% of the molecules from *DrugBank*. Figure 4.18 shows a scatter plot of the first two principal components (PCs). The loadings of the first and second PC are listed in Table 4.1.



PC1 and PC2 explain 45% and 22% of the total variance, respectively. Ligand entries of *TollDB* are within the chemical space of *DrugBank* molecules. This underlines the drug-like properties for most known TLR small-molecule ligands.



**Figure 4.18:** Scatter plot of the second PC against the first PC for the two data sets based on 15 relevant physicochemical properties. Dark blue dots represent compounds from *TollDB* and light blue dots represent compounds from *DrugBank*.

As we can see from Table 4.1, there is no descriptor that dominates PC1 or PC2. Instead, features correlated with the size of a molecule, such as MW, number of heavy atoms, water accessible surface area, Van der Waals volume and Van der Waals surface area are the major contributors to PC1, and the major contributors to PC2 are number of hydrogen bond donors, number of hydrogen bond acceptors, topological polar surface area and number of chiral atoms.

## 4.4 Matched molecular pairs and activity cliffs study

A matched molecular pair (MMP) is defined as a pair of compounds that only differ by a particular, well-defined, structure transformation (represented by a substructure) such as a ring or an R-group [158]. We used the “Automated Matched Pairs” node from Erlwood open source cheminformatics in KNIME [170, 171] to detect and output all the possible matched molecular pairs in *TollDB*. For the formation of MMP-cliffs, the following structural and potency criteria were applied. For a qualifying MMP, the heavy

**Table 4.1:** Loadings of the first two components resulting from PCA analysis.

Principal component	PC1	PC2
<b>Variance(%)</b>	<b>0.45</b>	<b>0.22</b>
<b>Principal component</b>	<b>0.45</b>	<b>0.67</b>
ASA	0.37	-0.04
a_acc	0.19	0.44
a_don	0.08	0.48
Weight	0.37	0.03
a_heavy	0.38	-0.00
a_hyd	0.34	-0.22
b_rotR	0.05	0.10
chiral	0.12	0.28
E_sol	-0.01	-0.04
FCharge	-0.00	-0.04
log $P(o/w)$	0.17	-0.44
rings	0.27	-0.15
vol	0.38	-0.07
VSA	0.37	-0.05
TPSA	0.16	0.46

ASA: water accessible surface area; a\_acc: number of H-bond acceptor atoms; a\_don: number of H-bond donor atoms; Weight: molecular weight; a\_heavy: number of heavy atoms; a\_hyd: number of hydrophobic atoms; b\_rotR: fraction of rotatable bonds; chiral: number of chiral centers; E\_sol: solvation energy; FCharge: sum of formal charges; log  $P(o/w)$ : log *n*-octanol/water partition coefficient; rings: number of rings; vol: Van der Waals volume; VSA: Van der Waals surface area; TPSA: topological polar surface area.

atom number of the transformed part from either molecule of the MMPs is no more than the heavy atom number of the common core part of the MMPs. Furthermore, the potency difference between compounds in an MMP that meets the structural criteria has to be at least two orders of magnitude (100-fold) in comparable assays. The workflow used is shown in Figure 3.2.

1373 molecules in *TollDB* with defined  $IC_{50}$ ,  $EC_{50}$ ,  $K_d$  values yielded a total of 9023 unique MMPs, and 275 of these pairs met our cliff criteria. After evaluating whether the activity is comparable (i.e., the testing condition is similar), the number of MMPs was reduced to 65 (including one enantiomer pair). Thus, ~0.72% of all MMPs represented activity cliffs (results shown in appendix A Table 6.1). Of all these 65 pairs, both compounds in the pair stem are from the same paper. The TLR targets for those pairs that feature activity cliffs are mainly TLR4, TLR7 and TLR8.

The activity differences between the MMPs range from 100 to 7200 fold. Several representative pairs that possess activity cliffs are shown in Table 4.2. Other MMPs with activity cliffs are shown in the appendix.

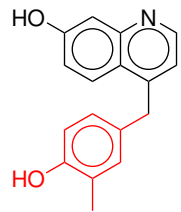
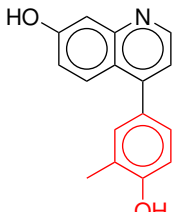
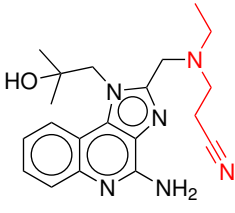
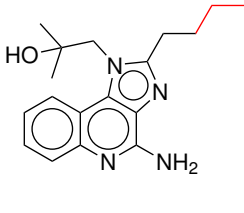
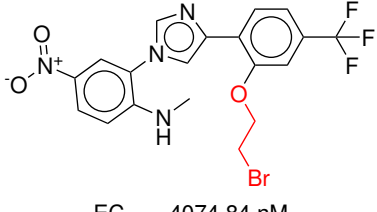
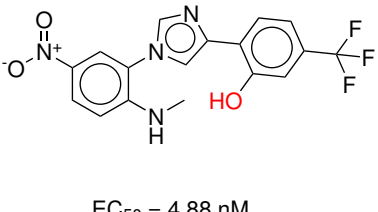
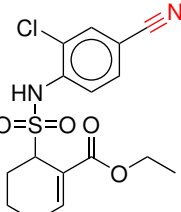
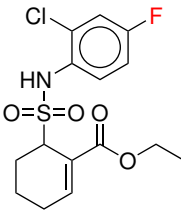
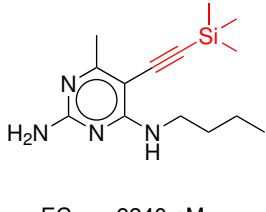
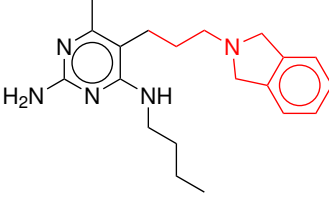
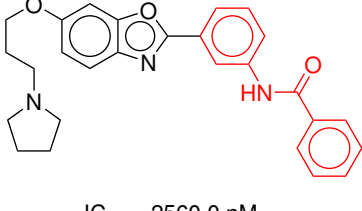
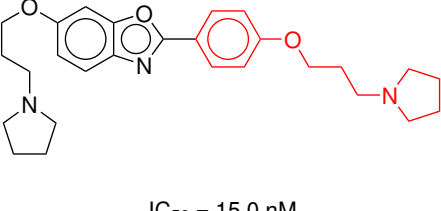
#### 4.4.1 Case study of activity cliffs

All the activity cliffs for MMPs in *TollDB* can be found in Table 6.1 in the appendix. The compound pairs used for docking studies were chosen from the pairs where any compound of the pair has an activity value of less than 10000 nmol (i.e., 10  $\mu$ mol). Selected pairs were then docked into the corresponding target protein in an attempt to rationalize the activity cliffs. The following section discusses example cases where molecular docking was used to explain activity cliffs between MMPs.

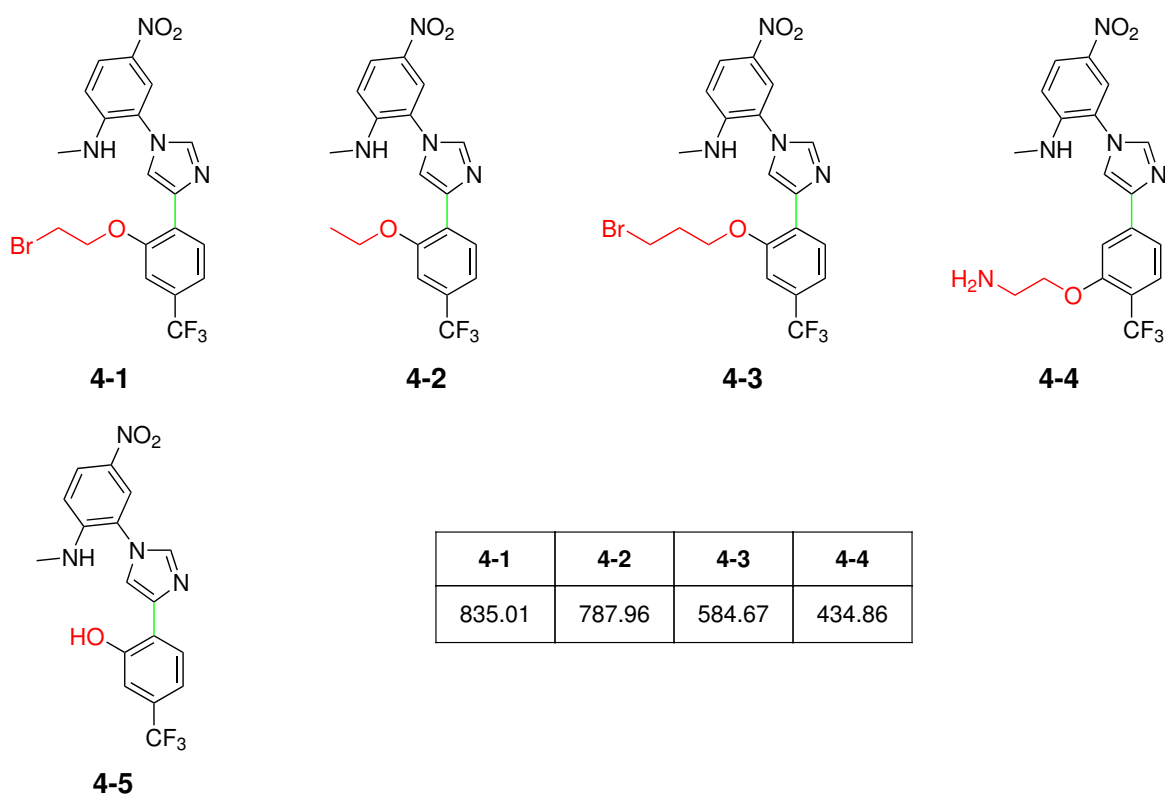
##### TLR2 agonists

For TLR2, the activity difference between compound pairs **4-1** and **4-5**, **4-2** and **4-5**, **4-3** and **4-5**, **4-4** and **4-5** are shown in Figure 4.19. The table in Figure 4.19 shows the activity comparison (the  $IC_{50}$ ,  $EC_{50}$  or  $K_d$  value for the corresponding compound divided by the  $IC_{50}$ ,  $EC_{50}$  or  $K_d$  value for compound **4-5**) between these pairs. As we can see directly from the structures, compounds with substitutions at the phenolic hydroxyl group (compounds **4-1**, **4-2**, **4-3** and **4-4**) show a greater decrease in activity compared to compounds without a substitution (compound **4-5**).

**Table 4.2:** Examples of activity cliffs between MMPs.

Target	Structure1	Structure2	AD <sup>a</sup>
TLR8 antagonism	 <p>IC<sub>50</sub> = 2370 nM</p>	 <p>IC<sub>50</sub> = 0.7 nM</p>	3385.71
TLR7 agonism	 <p>EC<sub>50</sub> = 8390 nM</p>	 <p>EC<sub>50</sub> = 8.58 nM</p>	977.86
TLR2 agonism	 <p>EC<sub>50</sub> = 4074.84 nM</p>	 <p>EC<sub>50</sub> = 4.88 nM</p>	835.01
TLR4 antagonism	 <p>IC<sub>50</sub> = 1600 nM</p>	 <p>IC<sub>50</sub> = 3.2 nM</p>	500
TLR8 agonism	 <p>EC<sub>50</sub> = 9240 nM</p>	 <p>EC<sub>50</sub> = 29.0 nM</p>	318.62
TLR9 antagonism	 <p>IC<sub>50</sub> = 2560.0 nM</p>	 <p>IC<sub>50</sub> = 15.0 nM</p>	170.67

AD<sup>a</sup>: Activity difference, equal to the activity value (IC<sub>50</sub>/EC<sub>50</sub>/Kd) for structure1 divided by the activity value (IC<sub>50</sub>/EC<sub>50</sub>/Kd) for structure2.



**Figure 4.19:** Activity cliff examples for TLR2 agonists. The activity comparison in the table refers to the listed compounds in the table compared to compound **4-5**, i.e., the activity value (IC<sub>50</sub>/EC<sub>50</sub>/K<sub>d</sub>) for the listed compounds divided by activity value (IC<sub>50</sub>/EC<sub>50</sub>/K<sub>d</sub>) for compound **4-5**.

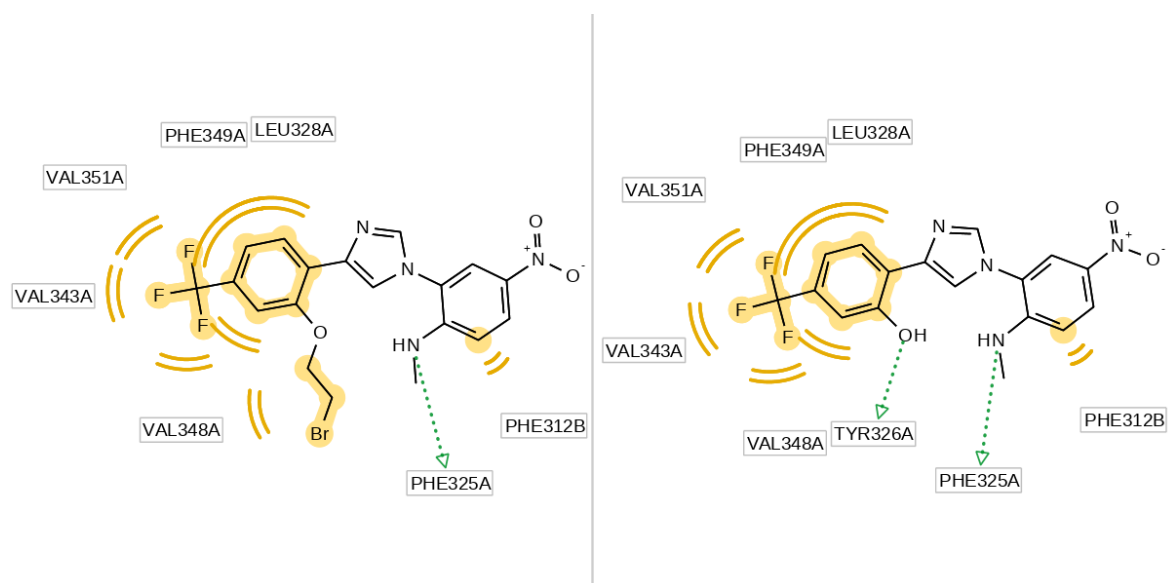
We docked those compound pairs into the TLR1-TLR2 heterodimer (PDB ID: 2Z7X) and compared their interactions. The results for compound pair **4-1** and **4-5** are shown in Figure 4.20 and Figure 4.21 as an example. The main difference in interactions between the two compounds is the H-bond donor interaction formed between the hydroxyl group in compound **4-5** and the tyrosine residue (TYR326A) on TLR2. This interaction is lost for compound **4-1** due to the substitution on the hydroxyl group, resulting in the dramatic activity decrease seen for compound **4-1** compared to compound **4-5**. This indicates the importance of this H-bond interaction between the ligand and the protein. Compounds **4-2**, **4-3** and **4-4** have the same interaction pattern as compound **4-1**. Compared to compound **4-5**, the other four compounds have lost the H-bond interaction due to their respective substitutions at the hydroxyl group. This substitution causes the substituted trifluoromethyl phenol portion to adopt a different orientation, i.e., the benzene ring plane is made to rotate along the single bond (shown in Figure 4.19 in the structure in green color) between the imidazole and the benzene moiety, to avoid the steric hindrance between the substitution chain and the leucine residue (LEU328A) in TLR2. This rotation causes the substituted moiety (shown in Figure 4.19 in red) to be oriented towards the pocket in TLR2, providing a further hydrophobic contact to the valine residue (VAL348A). However, this hydrophobic interaction cannot compensate for the loss of the H-bond donor interaction.

### **TLR8 agonists**

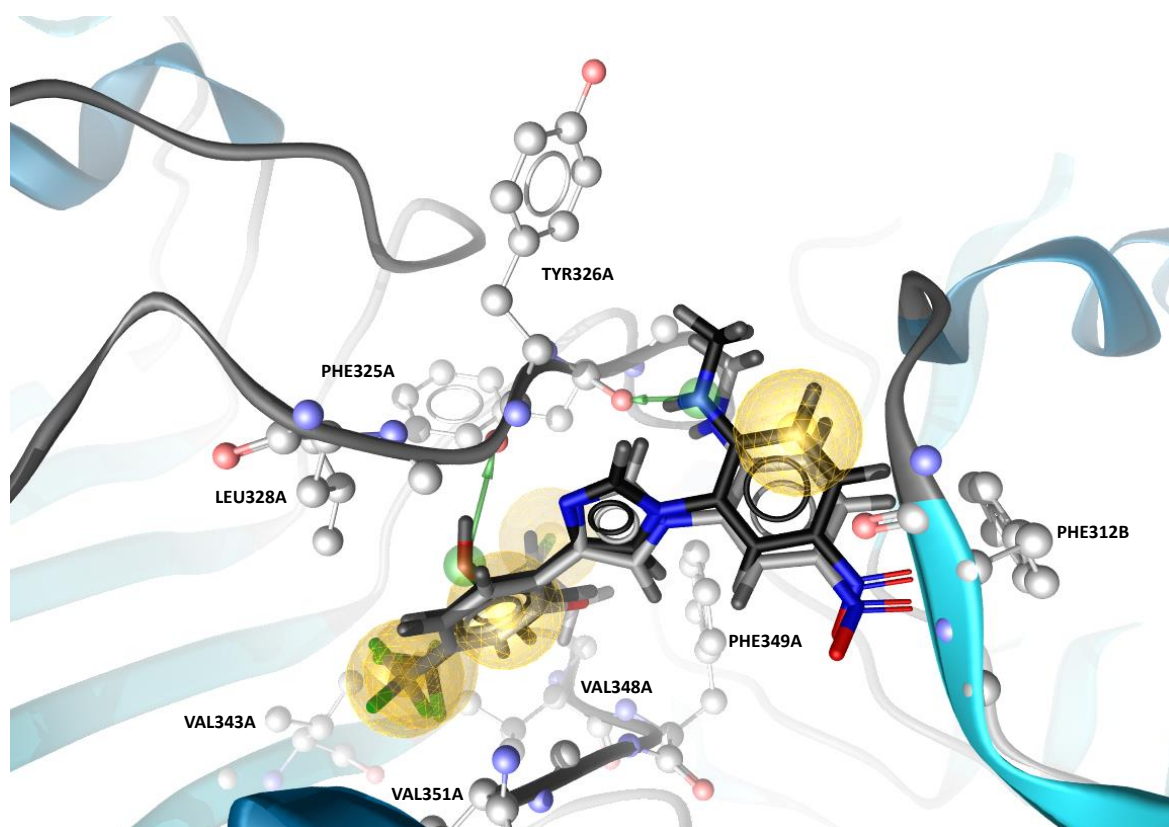
For TLR8, the activity difference between pairs **4-6** and **4-10**, **4-7** and **4-10**, **4-8** and **4-10**, **4-9** and **4-10** are shown in Figure 4.22. Table in Figure 4.22 shows the comparison in activity (the  $IC_{50}$ ,  $EC_{50}$  or  $K_d$  value for the corresponding compound divided by the  $IC_{50}$ ,  $EC_{50}$  or  $K_d$  value for compound **4-10**) between these pairs.

We docked those compound pairs in the TLR8 protein (PDB ID: 3W3J) and compared their interactions. As an example, **4-10** and **4-6** are shown in Figure 4.23 and Figure 4.24, respectively.

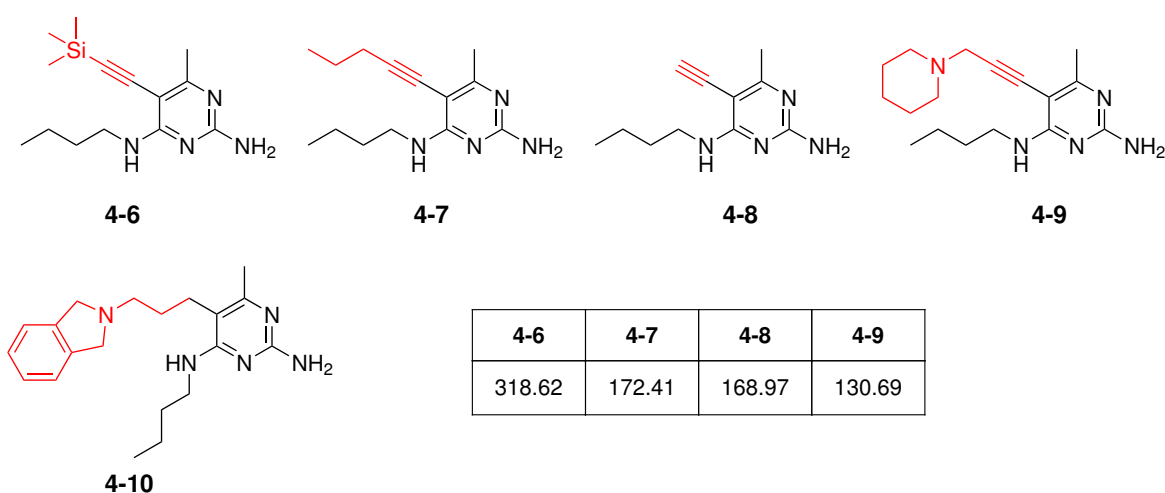
We can see from Figure 4.23 that the pyrimidine ring forms an aromatic interaction with PHE405B and the nitrogen in position 1 serves as an H-bond acceptor for THR574A. The protonated nitrogen in position 3 of the pyrimidine ring serves as an H-bond donor for residue ASP543A and is also a positive ionizable center. The primary amino



**Figure 4.20:** 2D representation of ligand-protein interactions for TLR2 agonists. Interaction pattern between compound **4-1** and the protein (left), interaction pattern between compound **4-5** and the protein (right). Interactions are color-coded: hydrophobic – yellow sphere; H-bond donor – green arrow.



**Figure 4.21:** Predicted binding poses for TLR2 agonists. Compound **4-1** with carbons colored in silver and compound **4-5** with carbons colored in black. Both compounds are shown in stick representation with nitrogens, oxygens, and fluorines colored in blue, red, and yellow, respectively. Important interacting protein residues are shown in ball-and-stick style. Interactions are color-coded: hydrophobic – yellow sphere; H-bond donor – green arrow.

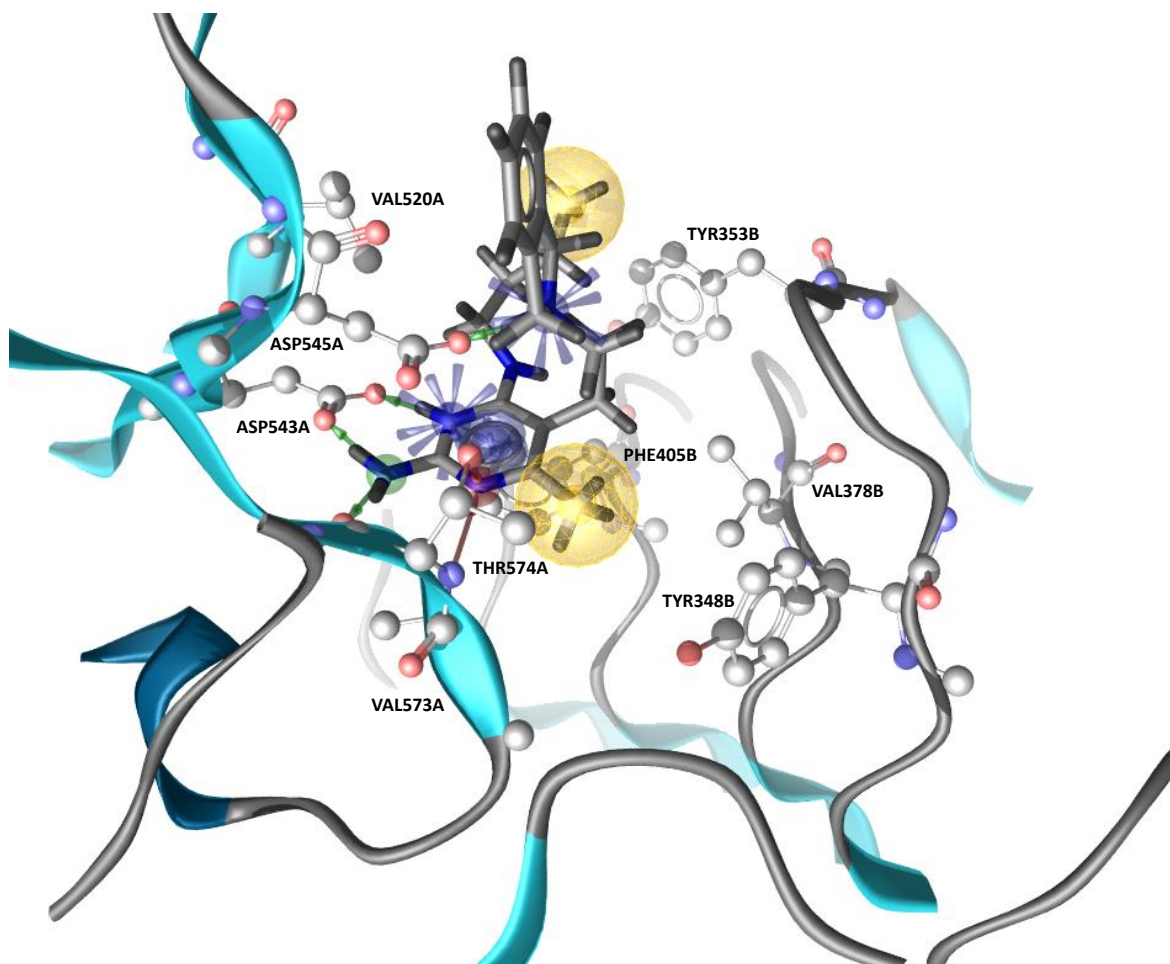


**Figure 4.22:** Activity cliff examples for TLR8 agonists. The activity comparison corresponds to the listed compounds in the table compared to compound **4-10**, i.e., the activity value ( $IC_{50}/EC_{50}/K_d$ ) for the listed compounds divided by the activity value ( $IC_{50}/EC_{50}/K_d$ ) for compound **4-10**.

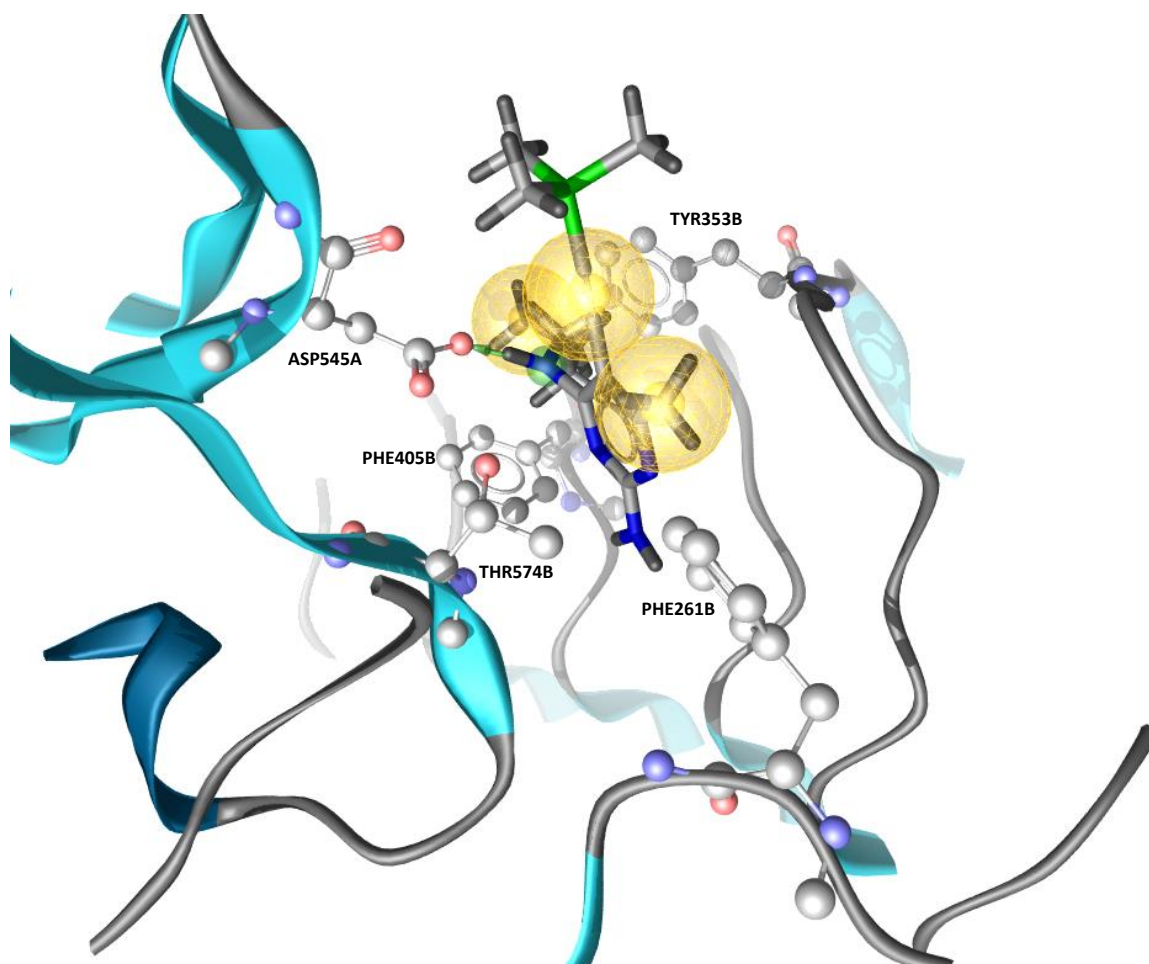
substitute group in position 2 serves as an H-bond donor to the THR754A backbone and ASP545A sidechain. There are two hydrophobic interaction areas. One is the methyl substitute at position 6 of the pyrimidine, surrounded by residues THR574A, PHE405B, VAL378B, VAL573A and TYR348B, and the other is the methyl group, the tail of the N-butylamino substitute in position 4 of the pyrimidine ring, surrounded by residues VAL520A and TYR353B. Apart from these, the protonated nitrogen on the isoindoline ring serves as an H-bond donor is also a positive ionizable center. The interaction between compound **4-10** and the protein residues cause the compound to adopt a configuration so that the isoindoline ring plane is perpendicular to the pyrimidine ring plane.

For compounds **4-6**, **4-7**, **4-8** and **4-9**, a triple bond exists in the substitution of position 5 in the pyrimidine ring. These molecules therefore become less flexible. The amino group, the pyrimidine ring, and the triple bond are positioned in the same plane. This forces the pyrimidine ring into a sub-optimal orientation when compared to compound **4-10**) and may adopt the binding pose shown in Figure 4.24, using compound **4-6** as an example. The two hydrophobic interaction areas and the H-bond donor to residue ASP545A remain, and a new hydrophobic interaction forms involving the triple bond. The overall decrease of interactions between the ligand and the protein residues may weaken the ligand-protein affinity, thus resulting in reduced activity for these compounds (compared to compound **4-10**).





**Figure 4.23:** Predicted binding pose for TLR8 agonists. Compound 4-10 is shown in stick representation with carbons and nitrogens colored in grey and blue, respectively. Important interacting protein residues are shown in ball-and-stick style. Interactions are color-coded: hydrophobic – yellow sphere; H-bond donor – green arrow; H-bond acceptor – red arrow; positive ionizable center – blue stars; aromatic interaction – blue sphere/circle.



**Figure 4.24:** Predicted binding pose for TLR8 agonists. Compound 4-6 is shown in stick representation with carbons, nitrogens, and silicones colored in grey, blue, and green, respectively. Important interacting protein residues are shown in ball-and-stick style. Interactions are color-coded: hydrophobic – yellow sphere; H-bond donor – green arrow.

## 4.5 Machine learning study

The main goal of the machine learning study was to build machine learning models that could distinguish small molecules (either as agonists or antagonists) that have high activity or low activity at different TLR targets. For a specific effect type for each TLR subtype, we used the tested positive examples and tested negative examples in *TollDB* as the data source for model construction.

In initial experiments, we used five different machine learning algorithms (starting with the default parameters) and three feature sets. The three feature sets comprise moe2d feature set, MQNs feature set and sf32 feature set as mentioned before. Taking the TLR2 agonism model as an example, the results are shown in Figure 4.25 and Figure 4.26. Results for other models are in appendix B (Table 6.2 and Table 6.3). For TLR2 agonist activity, different combinations of algorithms and feature sets with two different searching methods (random and grid search) result in different accuracy score, balanced accuracy score, MCC and AUC values. For TLR2 agonistic model with random search, the accuracy scores range from 0.827 to 0.889. The balanced accuracy scores range from 0.585 to 0.790, the MCC values range from 0.245 to 0.518, and AUC ranges from 0.706 to 0.924. While with grid search, the accuracy scores range from 0.827 to 0.877, the balanced accuracy scores range from 0.585 to 0.790, the MCC values range from 0.245 to 0.545, and AUC ranges from 0.720 to 0.919. We can see clearly that the results from random search and grid search are comparable. If we compare the results from using different feature sets, it is apparent that the sf32 feature set used the least number of features without losing too much accuracy. Furthermore, of the five different algorithms used, kNN and RF performed well using the sf32 feature set on the TLR2 agonist activity prediction model.

Comparing the results for other models, the accuracy scores of the initial trials do not differ between different algorithms or between different feature sets. The results obtained using random search or grid search also do not differ much. Taking into account the balanced accuracy score, MCC score, the AUC, and also the computational efficiency, we finally selected two algorithms and one feature set for the final models. The two algorithms comprise kNN and RF, and the sf32 feature set as the best feature set. When considering the same algorithm and the same feature set, the scores after

hyperparameter tuning using random search and grid search are similar. But if taking the computational cost into account, random search proves to be more efficient than grid search. This is obvious especially when dealing with a huge hyperparameter searching space, a big training set or a relatively more complex algorithm. Thus, random search was used for the later hyperparameter tuning process.

The detailed scores for the machine learning study are shown in Table 6.2 and Table 6.3 in appendix. We can choose the best algorithms for different models to make use of them in assisting drug design.

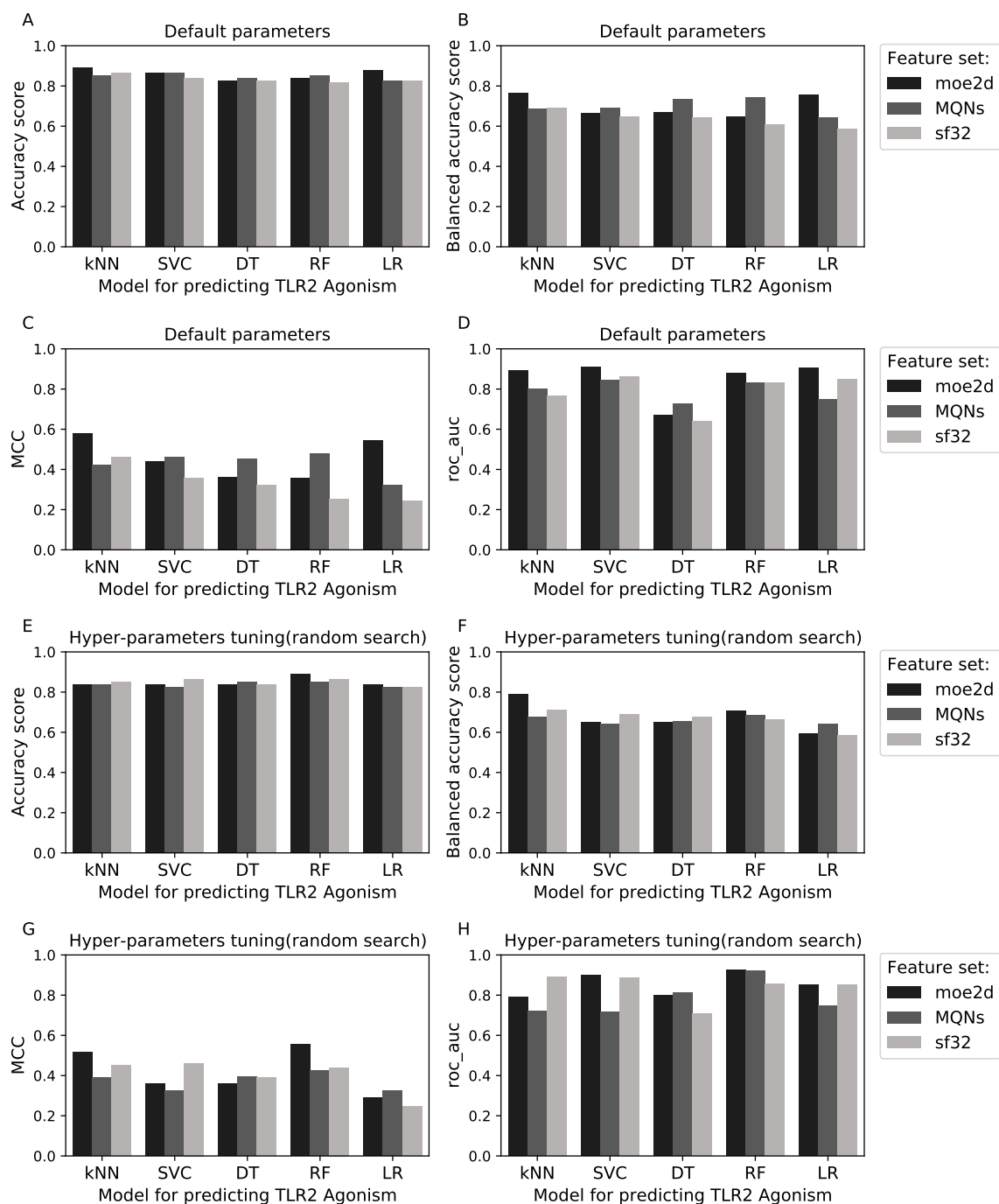
The ROC curve for the kNN and RF algorithms that use the sf32 feature set is shown in Figure 4.27. It is apparent from Figure 4.27A and Figure 4.27B that for using kNN in predicting TLR2 agonism activity, random search obtained better results than grid search. Regardless of the algorithms used, the ROC curve remains above the random curve (dashed line).

It is important to not only have an accurate model for making predictions, but also to know which features are most important in determining the forecast. Knowing which features are more important will help gain a better understanding of the model’s logic, and make the model more interpretable. This allows us to focus on these more important features when designing a new ligand. The feature importance for RF using the sf32 feature set for TLR2 agonism prediction model is shown in Figure 4.28 and Figure 4.29. From these figures we can see that the five most important features are h\_pKb (basicity), Weight (molecular weight), TPSA (topological polar surface area), a\_ICM (atom information content (mean)<sup>1</sup>) and b\_rotR (fraction of rotatable bonds) from grid search and Weight, TPSA, density<sup>2</sup>, h\_pKb and a\_ICM from random search. The physicochemical properties of a compound affect its potential to be developed into a drug. The most well-known rule of thumb is Lipinski’s rule of five (“rule of 5”) [203], which was designed to estimate oral bioavailability. However, many marketed drugs violate this rule. Tinworth and Young recently appraised the “Rule of 5” and recommended using the facts (measurements) and the patterns they reveal to establish

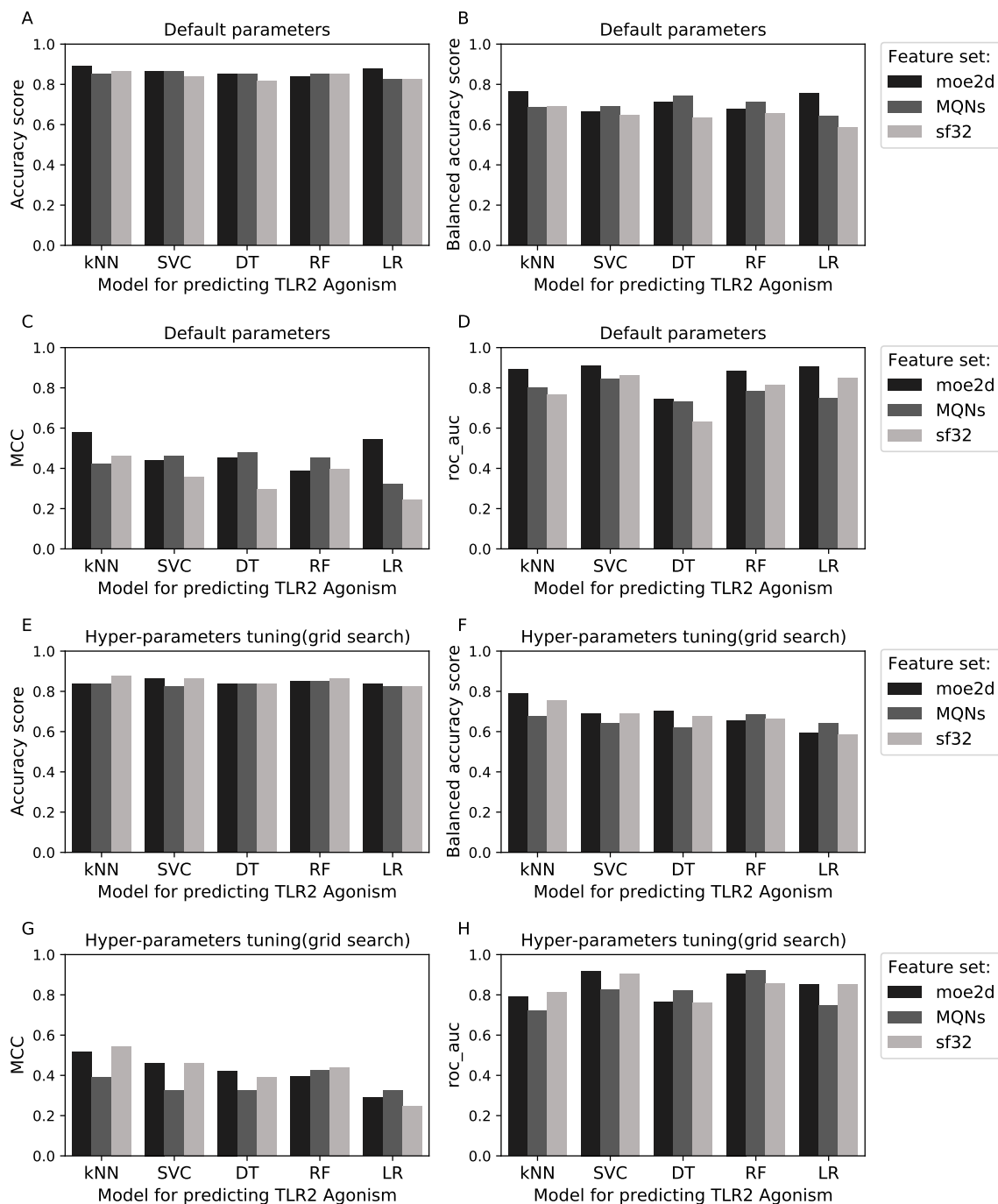
---

<sup>1</sup>This is the entropy of the element distribution in the molecule (including implicit hydrogens but not lone pair pseudo-atoms). Let  $n_i$  be the number of occurrences of atomic number  $i$  in the molecule. Let  $p_i = n_i/n$  where  $n$  is the sum of the  $n_i$ . The value of a\_ICM is the negative sum over all  $i$  of  $p_i \log p_i$ .

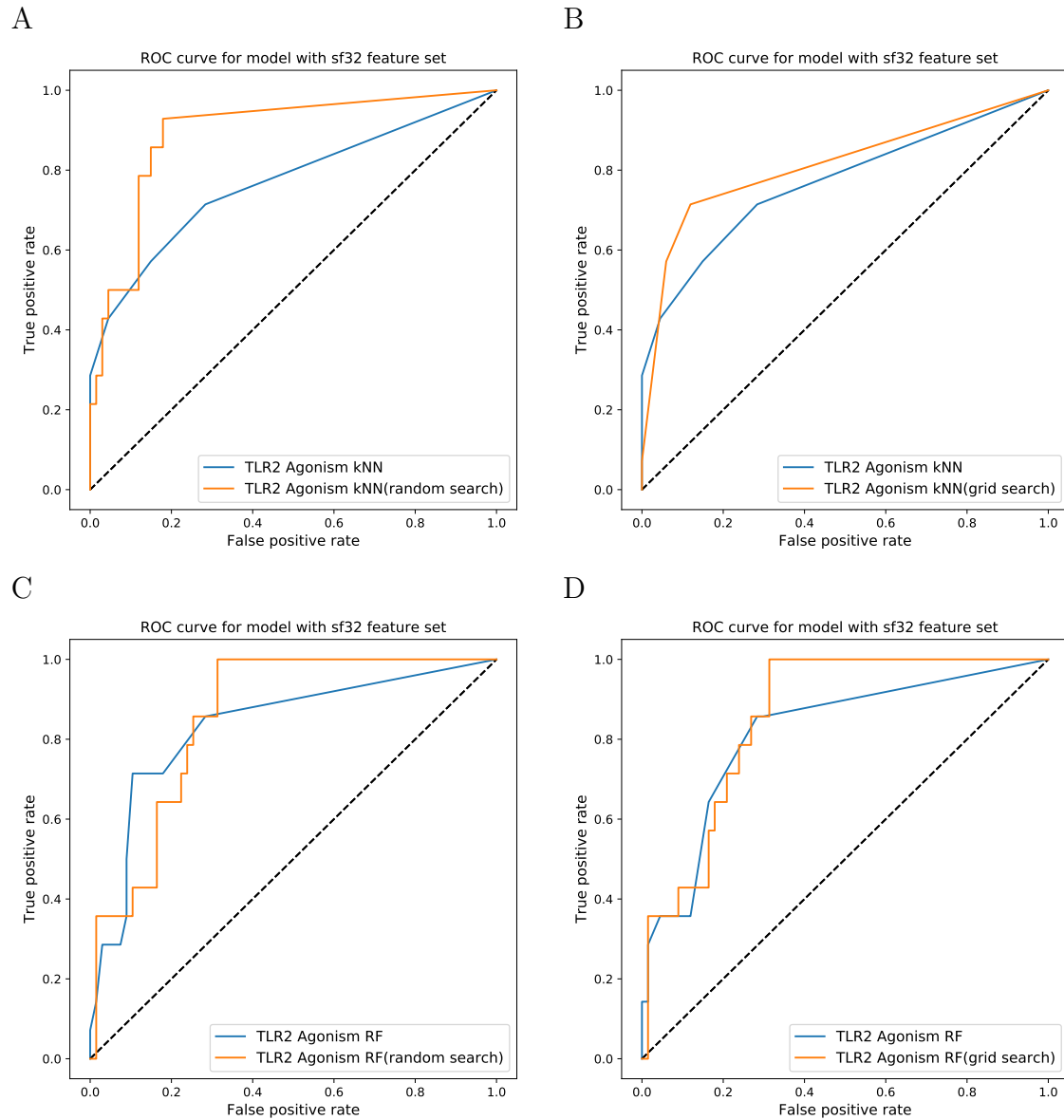
<sup>2</sup>Molecular weight divided by van der Waals volume. The van der Waals volume is calculated using a grid approximation (spacing 0.75 Å).



**Figure 4.25:** Model scores for TLR2 agonism with random search. (A) Accuracy score, (B) balanced accuracy score, (C) MCC, and (D) AUC for predicting TLR2 agonism using five different algorithms and three feature sets using default parameters for each algorithm. (E) Accuracy score, (F) balanced accuracy score, (G) MCC, and (H) AUC for predicting TLR2 agonism using five different algorithms and three feature sets. Results shown are for the best hyperparameters obtained through random search.



**Figure 4.26:** Model scores for TLR2 agonism with grid search. (A) Accuracy score, (B) balanced accuracy score, (C) MCC, and (D) AUC for predicting TLR2 agonism using five different algorithms and three feature sets using default parameters for each algorithm. (E) Accuracy score, (F) balanced accuracy score, (G) MCC, and (H) AUC for predicting TLR2 agonism using five different algorithms and three feature sets. Results shown are for the best hyperparameters obtained through grid search.



**Figure 4.27:** (A) ROC curve for TLR2 agonism prediction using the kNN algorithm and sf32 feature set with random search or (B) grid search. (C) ROC curve for TLR2 agonism prediction using the RF algorithm and sf32 feature set with random search or (D) grid search.

informative principles [204]. From the results of the RF model, the features that are important for activity prediction are MW, TPSA, h\_pKb, a\_ICM, b\_rotR and density. Typically, lipophilicity will increase with increasing MW and MW increase will also lead to the increase of log  $P$ , H-bond donor and H-bond acceptor number, but compound solubility will decrease. The flexibility and shape of a molecule will affect its permeability and solubility, and flexibility can to some extent be quantified by the number of rotatable bonds.

With these features identified as the most important ones, we plotted the distribution of these features against each other for the TLR2 agonism dataset in Figure 4.30. As shown in Figure 4.30A, the area in the upper left accumulates inactives, even though there are also inactives located in the same areas as the actives. Figure 4.30B shows that many compounds with the same MW have the same calculated TPSA. This might be due to the same core structure of shape similarity between compounds. Figure 4.30C shows that even actives show a large range of MW. The h\_pKb value of the actives are mostly in the range of 7-9. The combinations of these important features already show some level of discrimination between actives and inactives, explaining why these features possess a higher value of feature importance in our models. The scatter plots help to explore the different distributions of these features for actives and inactives, thus providing insights that can inform the design of new ligands.

Since we have imbalanced datasets for some models, we used the balanced accuracy score instead of accuracy score for evaluating the model performance. In the binary classification problem, this is defined as the arithmetic mean of sensitivity and specificity.<sup>3</sup> This will be discussed in the Discussion section.

In summary, we generated 8 prediction tasks, regarding four TLR targets (TLR2, TLR4, TLR7 and TLR8) and two effect types (agonism and antagonism). For instance, the task of predicting whether a compound is active or not for TLR2 agonism. Through our machine learning workflow, we can construct a robust model with a selected algorithm and optimized hyperparameters. The best model achieved a high level of accuracy, and can be used to aid the drug design process.

---

<sup>3</sup>Balanced-accuracy =  $\frac{1}{2}(\frac{TP}{TP+FN} + \frac{TN}{TN+FP})$



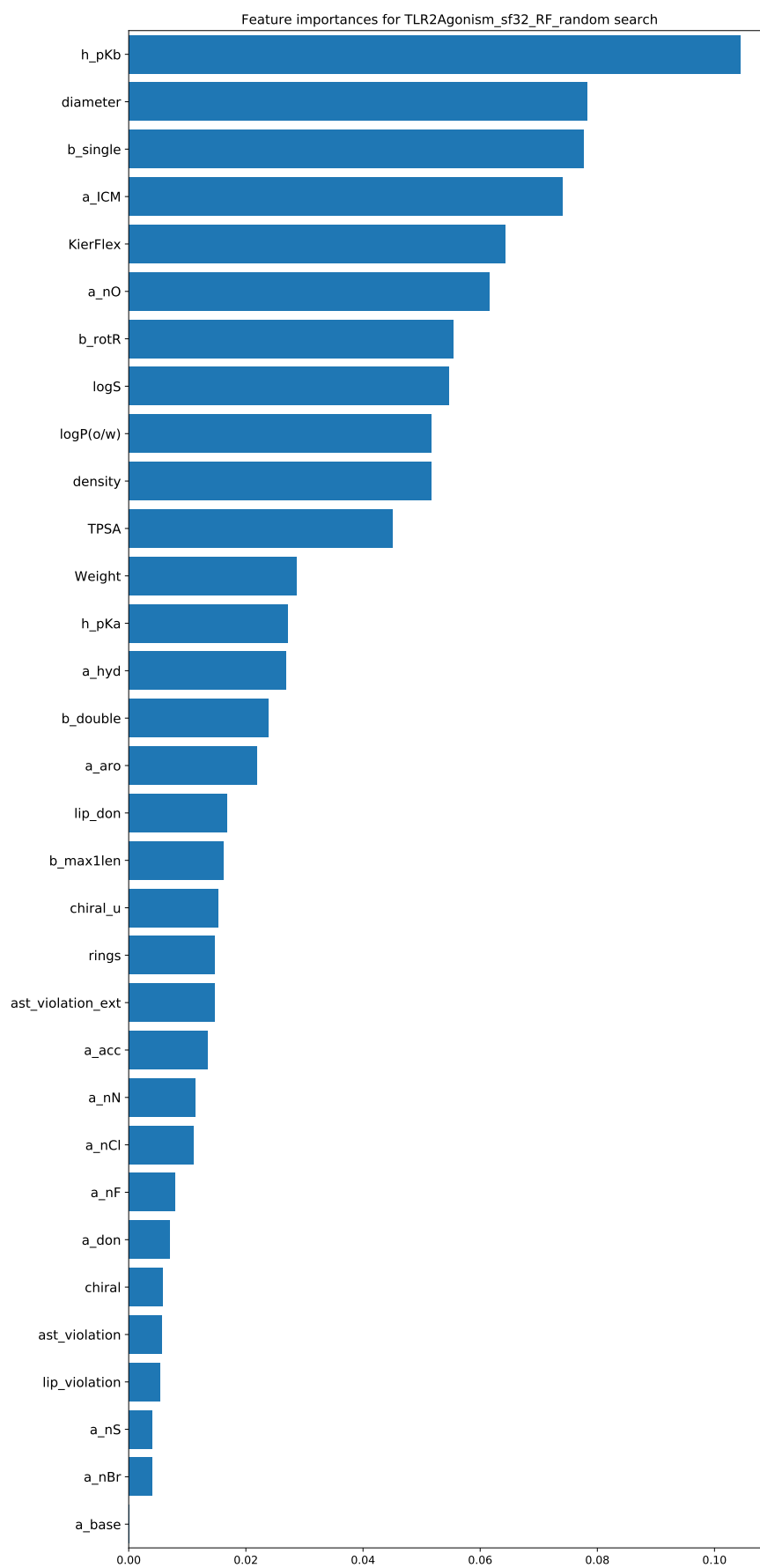


Figure 4.28: Feature importance for TLR2 agonism with random search.

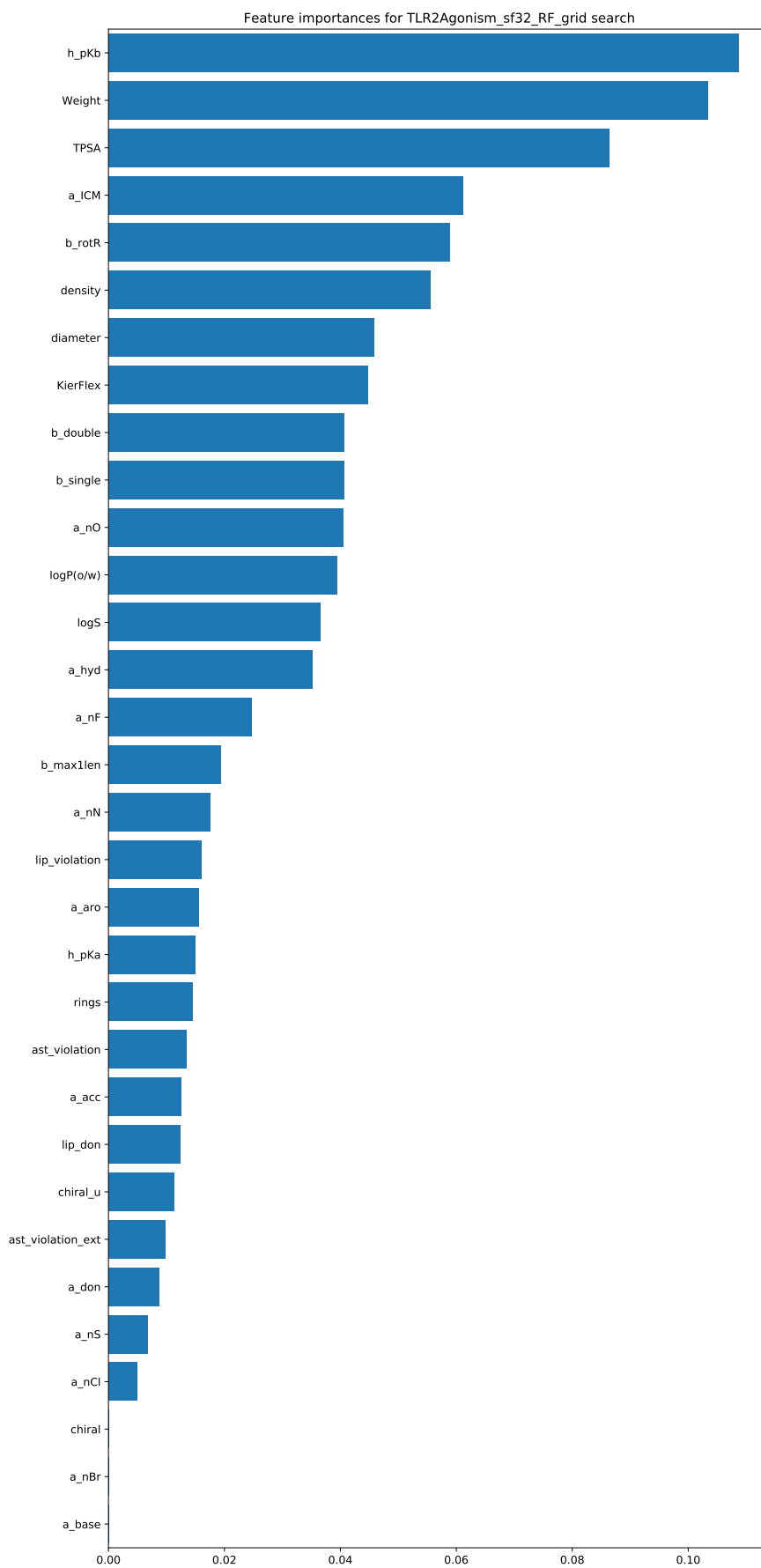
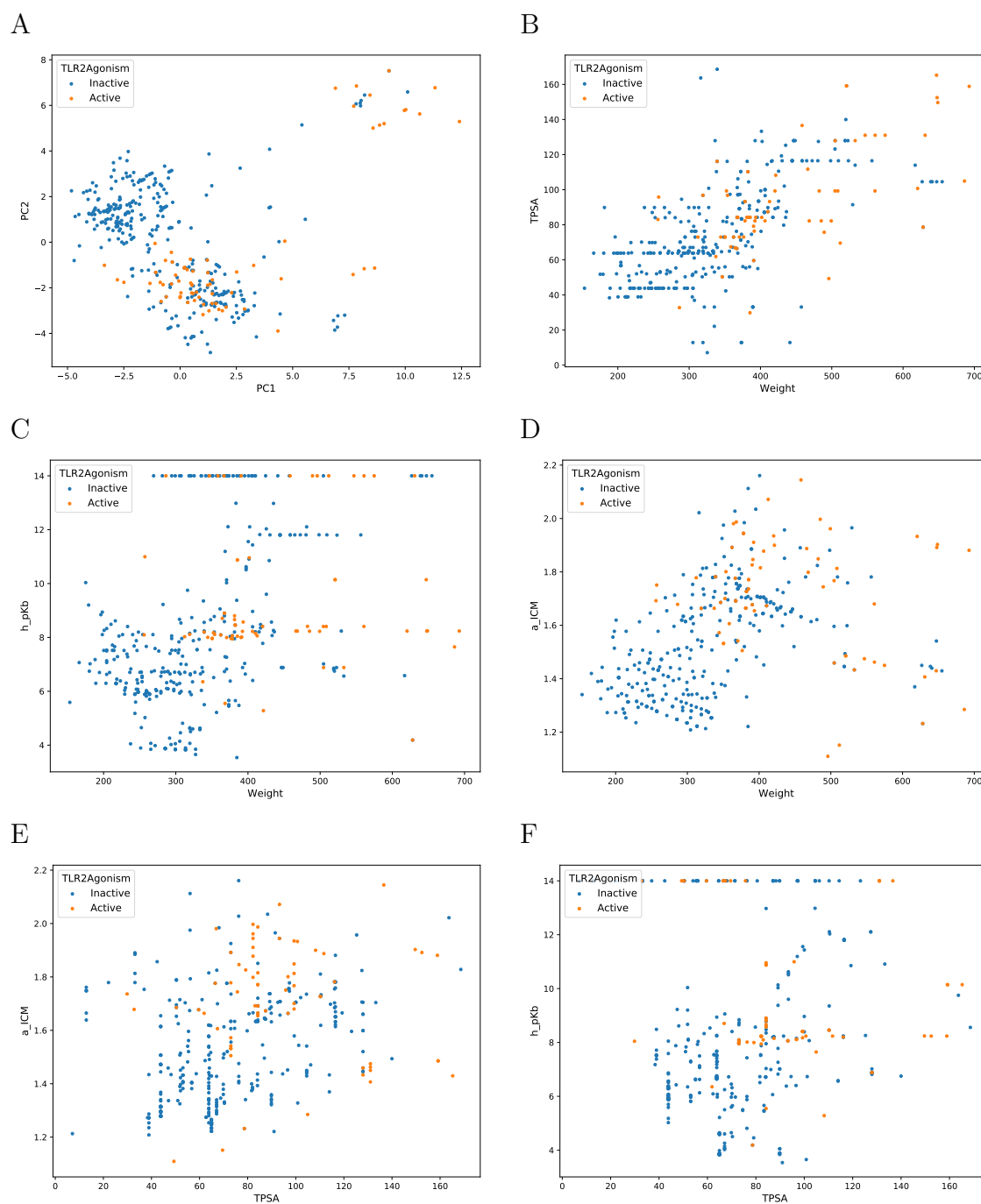


Figure 4.29: Feature importance for TLR2 agonism with grid search.



**Figure 4.30:** Scatter plots for the TLR2 agonism data set (A) PC1 against PC2 (B) MW against TPSA. (C) MW against h\_pKb. (D) MW against a\_ICM. (E) TPSA against a\_ICM. (F) TPSA against h\_pKb.



## 5. Discussion

The aim of the thesis was the construction of a small molecule TLR ligand database and the use of all available information for building machine learning models to aid in drug design. This was achieved by carefully developing a database and performing detailed data analysis. Using this curated data, prospective machine learning models for activity predictions were developed. In the following sections, the significance of the results, challenges, pitfalls and gained insights will be discussed.

### 5.1 Data compilation

Since the publications we collected come from different journals, in which the authors used different experimental conditions, such as using different concentrations of stimuli when testing antagonistic activity or using different units for compound concentrations in testing, it was essential to perform some pre-processing steps when collecting the data. This is distinct from the data cleaning process that was performed after the construction of the whole database. These pre-processing steps include validation of the compound structure using its chemical name in the publication (or checking the structure to see if it is the same as provided by the vendor); conversion of the compound concentrations to the same unit, for example,  $\mu\text{mol}$  and  $\text{mmol}$  were both converted to  $\text{nmol}$ ; removal of compounds that are cytotoxic at their effective concentration; and collection of compounds reported in supplementary materials.

It is important to record not only quantitative data, such as  $\text{IC}_{50}$ ,  $\text{EC}_{50}$ ,  $\text{K}_d$ , but also the qualitative data, such as when testing whether a compound is a selective TLR7 agonist with respect to other TLR targets (i.e. TLR8). Quite often, the publication contains information about compounds tested towards TLR7 in the form of  $\text{IC}_{50}$ ,  $\text{EC}_{50}$ , or  $\text{K}_d$  values, yet only reporting a result comparing this compound with negative/positive

control towards TLR8. This information is then used to declare that the compound is a selective TLR7 agonist. In this case, the qualitative data on the TLR8 testing result is also kept in the database, which proves useful for the overall understanding of compound selectivity.

Throughout the whole data collecting and database compiling process, using the same publication inclusion criteria, the same compound inclusion criteria and the same assay information collection method were essential for the overall consistency of the database, thus ensuring the database analysis and comparison to be comparable and reliable.

## 5.2 Data analysis and docking studies

In this work, the basic data analysis provides an overview of the compounds and assays in *TollDB*. It also provides an opportunity to explore the various features of ligands for different TLR subtypes. *TollDB* provides further insight into structure-activity relationship of TLR ligands.

### 5.2.1 Data analysis

An overview of the active compound number towards each TLR subtype is shown in Figure 4.5, giving us a first impression of the overall research status for each TLR subtype. This also indicates researchers' interests in different subtypes and the total hitherto successfully discovered number of small molecule ligands. Detailed explorations into the distribution of a specific TLR ligand (e.g., for TLR2 agonists) can provide a deeper understanding. For example, the currently discovered TLR2 agonists normally have an MW between 350-500 Da, which is on average about 100 Da heavier than TLR2 antagonists. The log  $P$  value for TLR2 agonists and antagonists is mostly in the range of 2-5, where the log  $P$  value of TLR2 antagonists is a bit smaller. TLR2 agonists have an average of three rings, six H-bond donors, two H-bond acceptors and an average TPSA of 84.22 Å<sup>2</sup>. All this information can guide the design of new TLR2 agonists as well as the prioritizing of new compounds to synthesize and test.

Comparison of the chemical space represented by the small molecules in *DrugBank* further highlights the druglike properties of ligands in *TollDB*.

### 5.2.2 Docking studies

In this thesis, docking studies were performed to explore activity cliffs between MMPs. For docking studies of TLR2 ligands, the crystal structure of the human TLR2/1 together with the ligand Pam<sub>3</sub>CSK<sub>4</sub> (PDB ID: 2Z7X) was chosen as the starting point. In 2014, Murgueitio et al. used the monomer of TLR2 from this crystal structure for docking studies of TLR2 ligands [205] while Zhong [206] and Durai [207] used the heterodimer TLR2/TLR1 structure for molecular docking. There are known TLR ligands that bind to heterodimers of TLR2/1 or TLR2/6, but as there is no compelling evidence for the direct binding of different agonists to TLR2 under physiological conditions, the exact mechanism of TLR2 interaction with its ligands remains unknown. It was widely acknowledged that the formation of ligand-TLR2 heterodimer complexes is crucial for the interaction of the TIR domain of the heterodimer, thus leading to either the activation or inhibition of the downstream signaling pathway. It was previously shown that TLR2 and TLR1 form heterodimers [208], and the cytoplasmic domain of TLR2 forms a functional pair with either TLR6 or TLR1, which is crucial for triggering the downstream signaling pathway. No matter whether this TIR domain approximation is caused by direct ligand binding to TLR2 monomer or to the surface of the heterodimer, they should all lead to a stable ligand-protein complex. Therefore, we chose this heterodimer for our docking studies.

The predicted binding mode for TLR2 agonists shown in Figure 4.21 explains the activity difference between the compound pairs: the H-bond interaction between the ligand and the protein residue PHE325A and the three hydrophobic interactions between the ligand and key interacting residues are crucial for the agonistic activity of the compounds. Apart from that, the extra H-bond between the ligand and the protein residue TYR326A dramatically enhances ligand activity. This furthers understanding of the binding mode.

## 5.3 Machine learning studies

Machine learning algorithms aim to optimize the performance of a certain task by using examples and/or experience [209]. We applied supervised learning in this work, which makes use of parts of the dataset for training and the remaining parts for validation. The

basic process for constructing a machine learning model includes data gathering, dataset pre-processing, algorithm selection, hyperparameter tuning, and model evaluation.

In the first documented application of machine learning on TLRs [123], the authors used features including count and position of motifs, the distance between the motifs and graphically derived features such as the radius of gyration and moment of inertia. These features are very specific due to the special structure of the oligodeoxyribonucleotide (ODN) ligands. Therefore, this approach is not suitable for application to a small molecule ligand dataset. Considering that our dataset only includes small molecule ligands and molecular descriptor calculation is easy and fast using MOE and RDKit, we prepared our own feature sets. The authors in the aforementioned paper [123] used downsampling due to an unbalanced dataset. We did not use this method since we value every datapoint in our dataset. However, in order to deal with the class imbalance problem, we calculated both the accuracy and the balanced accuracy.

We employed three feature sets for selection and optimization, and were able to use all the features in each feature set for building the machine learning models. However, problems occur when using all the possible features: (1) some features might be highly correlated; (2) some may not be relevant but may contribute to the noise in the model; and (3) using a large number of features may increase computational cost and even lead to overfitting. We therefore applied a series of preprocessing steps for the datasets, such as removal of constant and near-constant features, highly correlated features and features that are conceptually similar; encoding of categorical features (for the algorithms to have a better handling of the data); and double-checking the class label for the dataset.

We tried out five classical machine learning algorithms in our models, including kNN, SVC, DT, RF and LR. Each has its own advantages and disadvantages. LR is an easy, fast and simple classification method. It can be used for multiclass classifications, but is not applicable on non-linear classification problems, which is the case in many real scenarios. Additionally, collinearity and outliers tamper with the accuracy of the LR model. However, LR is the best starting point and gives an initial impression of how well machine learning models will perform in classification models. Other algorithms such as kNN, SVM and DT handle collinearity better than LR. One caveat in the use of kNN is that it is a non-parametric model and is comparatively slower. SVM



uses a kernel trick to solve complex solutions and outliers can be well handled using the soft margin constant  $C$  (a hyperparameter that decides the level of penalty over the outliers). DT is a tree-based algorithm which handles collinearity efficiently and can provide understandable explanations over the prediction. It forms no assumptions on the distribution of data and no need for data preprocessing, but the chances of overfitting are higher compared to RF, since RF is a collection of decision trees and the average/majority vote of the forest is selected as the predicted output (i.e., it gives a more generalized solution). RF is more robust and accurate than DT.

For the TLR2 agonism activity prediction model using kNN, we have a total of 402 compounds in our dataset, including 333 inactives and 69 actives. This dataset was divided into the training and testing set by a ratio of 4:1, resulting in 81 compounds in the testing set. We can see from these numbers that the dataset for TLR2 agonism is imbalanced. If we had a balanced dataset, i.e., the number of activities and inactives in the dataset is similar, we could use 50% accuracy as the baseline for evaluating the model. In other words, if our model has an accuracy of above 50% then the prediction model performs better than random guess. Since we have an imbalanced dataset, the number of actives and inactives in our training set is 57 and 264, respectively. We must consider the different number of instances for each class. If we predict all the instances as inactives, we will get an accuracy of 82%, which is far beyond 50% for balanced datasets. Predicting all the instances to the major class is a highly biased model and could not generalize well to new data. Therefore, we used the balanced accuracy score for comparison to avoid the influence of the imbalanced datasets. As we can see for this model (TLR2 agonism prediction model using kNN algorithm, sf32 feature set, with 10-fold cross-validation and random search for the best hyperparameters), the accuracy and balanced accuracy are 85.2% and 71.3%, respectively, while the MCC and AUC are 0.453 and 0.892, respectively. These values indicate that even with an imbalanced dataset, where the number of compounds in one of the classes (the actives) is quite small, the model performance is good and could be later used for prediction or for prioritizing screening hits.

The TLR2 agonism activity prediction model showed that even with an imbalanced dataset, moderate accuracy is achievable. In the case of a roughly equal number of actives and inactives as in the TLR8 agonism activity prediction model, our dataset

contains 311 inactives and 360 actives in total. Using kNN, sf32 feature set with 10-fold cross-validation and random search for searching the best hyperparameters, the accuracy score, balanced accuracy score, MCC and AUC are 82.7%, 82.3%, 0.666 and 0.926, respectively. When using RF, these values are 85.6%, 85.4%, 0.715 and 0.918, respectively. The results are much better than in the case of an imbalanced dataset.

No particular machine learning model outperforms all others, and the fact that a certain algorithm can outperform others in one model does not mean it will outweigh other algorithms in other models. When dealing with predicting different activities for TLRs, we should choose different algorithms due to the different distribution of the data. The goal is to choose the most suitable algorithm and construct the best model for a specific scenario.

Hyperparameters are critical in building robust and accurate models. They help strike the balance between bias and variance, thus preventing the model from overfitting or underfitting. To be able to adjust the hyperparameters, we need to understand what they mean and how they change a model. In our hyperparameter tuning steps, grid search and random search were both explored. Random search is much faster than grid search, and could find the best hyperparameters combination that performs nearly the same as that which was found using grid search. Thus, in our models, random search is primarily used and grid search is used when hyperparameters are narrowed down to a small set. This saved a lot of time when using more complex models such as RF or when either the hyperparameter searching space or the dataset increased in size.

When designing hyperparameter searching space, there are some rules of thumb to be taken into consideration. For example, in kNN, the hyperparameter K represents the number of neighbors to consider. It is better to be set to an odd number (in binary classification problems) to avoid ties, and the searching space for K is usually set to less than the square root of the feature numbers. It is very important to understand every hyperparameter in each algorithm in order to set a more meaningful searching space for hyperparameter tuning.

Every algorithm is different and has different requirements when applying. For example, in kNN or SVM, scaling is important in these non-tree-based algorithms. All features should be normalized to the same scale, i.e., data standardization in preprocessing is needed. Most importantly, k-nearest neighbor does not “learn”, it just

memorizes the data, so when the dataset gets bigger, computation time will increase dramatically.

For application of machine learning models in TLR activity prediction, we can see from our results that the quality of the data is important, not only data reliability and the amount of data but also a dataset balance. A balanced and large dataset results in better predictions, as we have seen for TLR2 agonist and TLR8 agonist models. For TLR2, 4, 7 and 8 we obtained robust and accurate models, even though for some targets the dataset is imbalanced. However, for other TLR targets like TLR3, 5, 9, there are not enough data to build a model. This indicates the current research state for these targets, suggesting that small molecule modulators for these targets remain to be explored. Data is the biggest limitation for developing machine learning models. With good data, an appropriately selected algorithm and careful study in hyperparameter tuning, we were able to obtain robust and accurate models and apply them to predict and prioritize possible new compounds for chemical synthesis or biological testing.



## 6. Conclusion and Outlook

TLRs play a pivotal role in the native immune response of humans. They are related to various diseases and have been proposed as a promising drug target for the treatment of these diseases. In this thesis, we aimed to build a comprehensive small molecule TLR modulator database, analyzed it for assisting rational drug design and developed predictive models using machine learning.

We carefully constructed a comprehensive database and developed a web application for information retrieval. Our work combines data analysis and machine learning. We developed a retrospective study on activity cliffs occurring in reported ligands and explained the differences using molecular docking. We developed an automated machine learning workflow that can be used for filtering hits from virtual screening campaigns, as well as prioritizing compounds for synthesis and testing.

With the rapid development of computer technology and its usage in rational drug design, more high-quality data are in need for data-oriented studies. The successful construction of *TollDB* made it possible to explore the TLR ligand space and develop prediction models through a machine learning approach. This database can also help us better understand the structure activity relationship for TLR ligands, provide a solid base for novel small TLR ligand design, and even be used as a library for drug repurposing screening.

Due to the crucial role that TLRs play in the innate and adaptive immune response and the use of TLR ligand therapeutics in various diseases, it would be hugely beneficial to develop TLR ligands into drugs. With the increase in newly discovered small molecule TLR modulators and rapidly developing artificial intelligence technology and its application in the drug discovery process, it is promising to discover novel, low toxicity and high efficiency lead compounds targeting TLRs. Furthermore, enormous amounts of data relating to chemical structures and their biological activity at TLRs

will help to gain a better understanding of the interactions between the ligand and the target, thus benefiting the study of TLR signaling mechanisms. Future research efforts can be dedicated to studying the mechanism of TLR modulation by small molecules based on our chemical space, matched molecular pair, activity cliff and machine learning study presented in this thesis.

# Appendix

## A MMPs for *TollDB*

**Table 6.1:** MMPs for *TollDB*

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
*c1ccnc2cc(O)ccc12	*Cc1ccc(O)c(C)c1	*c1ccc(O)c(C)c1	0.487	TLR8	2370.0/0.7
*c1ccnc2cc(O)ccc12	*Cc1ccc(O)c(C)c1	*c1cnc(OC)c(C)c1	0.500	TLR8	2370.0/0.7
*c1ccc2c(-c3ccc(O)c(C)c3)ccnc2c1	*c1ccccc1	*OC	0.227	TLR8	910.0/0.5
*c1ccc2c(-c3ccc(O)c(C)c3)ccnc2c1	*c1ccccc1	*O	0.209	TLR8	910.0/0.7
*c1ccnc2cc(OC)ccc12	*CCCCC	*c1ccc(O)c(C)c1	0.405	TLR8	610.0/0.5
*c1ccnc2cc(O)ccc12	*Cc1ccc(O)c(C)c1	*c1cnc(O)c(C)c1	0.487	TLR8	2370.0/2.3
	*N(CC)CCC#N	*CCC	0.240	TLR7	8390.0/8.58
*Cc1nc2c(N)nc3ccccc3c2n1CC(C)(C)O					
*c1ccc2c(-c3ccc(O)c(C)c3)ccnc2c1	*c1ccccc1	*Cl	0.209	TLR8	910.0/1.0
	*N(CC)CC	*CCC	0.208	TLR7	7710.0/8.58
*Cc1nc2c(N)nc3ccccc3c2n1CC(C)(C)O					
*c1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*OCCBr	*O	0.123	TLR2	4074.84/4.88
*c1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*OCC	*O	0.107	TLR2	3845.23/4.88

*Continued on next page*



Table 6.1 – Continued from previous page

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
	*c1cccnc1NCCOC		0.483	TLR7	125.0/0.2
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21		*c1ccc(NCCN2CCCCC2)nc1			
	*O	*N1CCCCC1	0.158	TLR7	121.0/0.2
*CCNc1ccc(Cn2c(=O)[nH]c3c(N)nc(C(F)(F)F)cc32)cn1					
	*c1cccnc1OCCN1CCCC1		0.508	TLR7	121.0/0.2
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21		*c1ccc(NCCN2CCCCC2)nc1			
*c1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*OCCCB r	*O	0.138	TLR2	2853.19/4.88
*Oc1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*CCBr	*C(=O)c1ccccc1	0.200	TLR2	4074.84/7.65
	*c1ccnc(NCCOC)c1		0.483	TLR7	105.0/0.2
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21		*c1ccc(NCCN2CCCCC2)nc1			
*Oc1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*CC	*C(=O)c1ccccc1	0.188	TLR2	3845.23/7.65
	*C#N	*F	0.106	TLR4	1600.0/3.2
*c1ccc(NS(=O)(=O)C2CCCC=C2C(=O)OCC)c(Cl)c1					

Continued on next page

Table 6.1 – *Continued from previous page*

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
*Oc1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*CCBr	*C(C)=O	0.133	TLR2	4074.84/8.83
*Oc1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*CC	*C(C)=O	0.119	TLR2	3845.23/8.83
*c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*c1ccc(C(F)(F)F)c(OCCN)c1	*c1ccc(C(F)(F)F)cc1O	0.474	TLR2	2122.13/4.88
	*c1ccnc(OCCN2CCCC2)c1		0.508	TLR7	81.9/0.2
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21		*c1ccc(NCCN2CCCC2)nc1			
*Oc1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*CCCBBr	*C(=O)c1ccccc1	0.212	TLR2	2853.19/7.65
*S(=O)(=O)Nc1ccc(F)cc1Cl	*[C@H]1CCCC=C1C(=O)OCC	*[C@@H]1CCCC=C1C(=O)OCC	0.522	TLR4	640.0/1.8
	*c1ccccc1C(=O)OC	*c1ccc(F)cc1Cl	0.417	TLR4	1100.0/3.2
*NS(=O)(=O)C1CCCC=C1C(=O)OCC					

*Continued on next page*

Table 6.1 – Continued from previous page

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
*Oc1cc(C(F)(F)F)ccc1-c1cn(- c2cc([N+](=O)[O-])ccc2NC)cn1	*CCBr	*C(=O)CCCCCCC	0.212	TLR2	4074.84/12.32
*Oc1cc(C(F)(F)F)ccc1-c1cn(- c2cc([N+](=O)[O-])ccc2NC)cn1	*CCCB	*C(C)=O	0.148	TLR2	2853.19/8.83
*c1c(C)nc(N)nc1NCCCC	*C#C[Si](C)(C)C	*CCCN1Cc2ccccc2C1	0.455	TLR8	9240.0/29.0
*CC- CCc1cccc2nc(N)c(CCCCC)cc12	*NC(=N)N	*CN	0.174	TLR8	2862.0/9.0
*Oc1cc(C(F)(F)F)ccc1-c1cn(- c2cc([N+](=O)[O-])ccc2NC)cn1	*CC	*C(=O)CCCCCCC	0.200	TLR2	3845.23/12.32
*c1ccc(Cn2c(=O)[nH]c3c(N)nc(C(F)(F)F)cc32)cn1	*NCCO	*N(C)CCN1CCN(C)CC1	0.288	TLR7	121.0/0.4
*CCc1cccc2nc(N)c(CCCCC)cc12	*C(N)=O	*CCN	0.182	TLR8	2181.0/9.0
*c1cn(-c2cc([N+](=O)[O-] )ccc2NC)cn1	*c1ccc(C(F)(F)F)c(OCCN)c1	*c1ccc(C(F)(F)F)cc1OC(C)=O	0.500	TLR2	2122.13/8.83

Continued on next page

Table 6.1 – *Continued from previous page*

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
	*NCC	*CCC	0.174	TLR7	2000.0/8.58
*Cc1nc2c(N)nc3ccccc3c2n1CC(C)(C)O					
*Oc1cc(C(F)(F)F)ccc1-c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*CCCBBr	*C(=O)CCCCCCC	0.224	TLR2	2853.19/12.32
*c1ccnc2ccccc12	*C(=O)c1ccc(OC)cc1	*c1cc(C)c(O)c(C)c1	0.538	TLR8	5570.0/25.5
*S(=O)(=O)Nc1ccc(F)cc1Cl	*[C@H]1CCCC=C1C(=O)OCC	*C1CCCC=C1C(=O)OCC	0.522	TLR4	640.0/3.2
*c1c(C)nc(N)nc1NCCCC	*C#CCCC	*CCCN1Cc2ccccc2C1	0.442	TLR8	5000.0/29.0
*c1nc2ccc(OCCCN3CCCC3)cc2o1	*c1cccc(NC(=O)c2ccccc2)c1	*c1ccc(OCCCN2CCCC2)cc1	0.485	TLR9	2560.0/15.0
*c1c(C)nc(N)nc1NCCCC	*C#C	*CCCN1Cc2ccccc2C1	0.400	TLR8	4900.0/29.0
*c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*c1ccc(C(F)(F)F)cc1OCCBr	*c1ccc(OC(F)(F)F)cc1	0.474	TLR2	4074.84/24.87
	*N1CCOCC1	*NCCN1CCCCC1	0.288	TLR7	32.1/0.2
*c1ccc(Cn2c(=O)[nH]c3c(N)nc(C(F)(F)F)cc32)cn1					

*Continued on next page*

Table 6.1 – Continued from previous page

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
	*c1ccccc1	*CN1CCCCC1	0.246	TLR7	31.0/0.2
*CNc1ccc(Cn2c(=O)[nH]c3c(N)nc(C(F)(F)F)cc32)cn1					
*c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*c1ccc(C(F)(F)F)cc1OCC	*c1ccc(OC(F)(F)F)cc1	0.464	TLR2	3845.23/24.87
*c1ccnc2cc(OC)ccc12	*CCCCC	*c1enc(O)c(C)c1	0.405	TLR8	610.0/4.2
*c1ccnc2ccccc12	*C(=O)c1ccc(OC)cc1	*c1ccc(O)c(C)c1	0.526	TLR8	5570.0/38.6
	*OC	*N1CCCCC1	0.172	TLR7	28.6/0.2
*CCNc1ccc(Cn2c(=O)[nH]c3c(N)nc(C(F)(F)F)cc32)cn1					
	*c1ccc(C)nn1	*c1ccc(N2CCN(C)CC2)nc1	0.423	TLR7	411.0/3.0
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21					
	*OCCN(C)C	*NCCN1CCCCC1	0.288	TLR7	26.9/0.2
*c1ccc(Cn2c(=O)[nH]c3c(N)nc(C(F)(F)F)cc32)cn1					
	*c1ccc(C#N)cc1Cl	*c1ccccc1Cl	0.391	TLR4	1600.0/12.0
*NS(=O)(=O)C1CCCC=C1C(=O)OCC					

Continued on next page

Table 6.1 – *Continued from previous page*

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
*S(=O)(=O)Nc1ccc(F)cc1Cl	*C1CCCCC=C1C(=O)OCC	*[C@@H]1CCCC=C1C(=O)OCC	0.532	TLR4	240.0/1.8
*c1ccnc2cc(OC)ccc12	*CCCCC	*c1cnc(OC)c(C)c1	0.421	TLR8	610.0/4.6
*c1c(C)nc(N)nc1NCCCC	*C#CCN1CCCCC1	*CCCN1Cc2ccccc2C1	0.489	TLR8	3790.0/29.0
	*c1ccc(Cl)cc1	*c1ccc(F)cc1Cl	0.378	TLR4	400.0/3.2
*NS(=O)(=O)C1CCCC=C1C(=O)OCC	*c1cccnc1NCCOC		0.483	TLR7	125.0/1.0
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21		*c1ccc(N(C)CCN(CC)CC)nc1			
	*NCCO	*N(C)CCN(CC)CC	0.263	TLR7	121.0/1.0
*c1ccc(Cn2c(=O)[nH]c3c(N)nc(C(F)(F)F)cc32)cn1	*c1cccnc1OCCN1CCCC1		0.508	TLR7	121.0/1.0
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21		*c1ccc(N(C)CCN(CC)CC)nc1			
*c1cn(-c2cc([N+](=O)[O-])ccc2NC)cn1	*c1ccc(C(F)(F)F)cc1OCCCBBr	*c1ccc(OC(F)(F)F)cc1	0.483	TLR2	2853.19/24.87
*c1c(C)nc(N)nc1NCCCC	*C	*CCCN1Cc2ccccc2C1	0.385	TLR8	3280.0/29.0

*Continued on next page*

Table 6.1 – Continued from previous page

Common_Substructure	Transformation1	Transformation2	MCS <sup>a</sup>	Target	Activity Compare <sup>b</sup>
*Cc1cccc2nc(N)c(CCCCC)cc12	*c1cccc1CN	*CCCCN	0.319	TLR8	1000.0/9.0
	*c1cccc(C)n1	*c1ccc(N2CCN(C)CC2)nc1	0.423	TLR7	328.0/3.0
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21					
*CCc1c(C)nc(N)nc1NCCCC	*C(C)(C)O	*CN1Cc2cccc2C1	0.364	TLR8	3140.0/29.0
*CC-	*NC(=N)N	*N	0.156	TLR8	2862.0/27.0
CCc1cccc2nc(N)c(CCCCC)cc12					
	*c1ccnc(NCCOC)c1		0.483	TLR7	105.0/1.0
*Cn1c(=O)[nH]c2c(N)nc(C(F)(F)F)cc21		*c1ccc(N(C)CCN(CC)CC)nc1			

<sup>a</sup>: MCS distances.

<sup>b</sup>: the activity of molecule with common substructure and Transformation1 divided by the activity of molecule with common substructure and Transformation2.

**B Machine learning results for random search (Table 6.2) and grid search (Table 6.3)**



**Table 6.2:** Machine learning study for *ToxDB* with random search

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR2Agonism	kNN	moe2d	0.889	0.763	0.581	0.896	0.840	0.790	0.518	0.790
TLR2Agonism	SVC	moe2d	0.864	0.664	0.440	0.913	0.840	0.649	0.358	0.899
TLR2Agonism	DT	moe2d	0.827	0.670	0.361	0.670	0.840	0.649	0.358	0.800
TLR2Agonism	RF	moe2d	0.840	0.649	0.358	0.878	0.889	0.707	0.557	0.924
TLR2Agonism	LR	moe2d	0.877	0.756	0.545	0.907	0.840	0.592	0.290	0.853
TLR2Agonism	kNN	MQNs	0.852	0.684	0.424	0.801	0.840	0.677	0.391	0.720
TLR2Agonism	SVC	MQNs	0.864	0.692	0.462	0.844	0.827	0.641	0.325	0.717
TLR2Agonism	DT	MQNs	0.840	0.733	0.455	0.730	0.852	0.656	0.396	0.813
TLR2Agonism	RF	MQNs	0.852	0.741	0.482	0.832	0.852	0.684	0.424	0.920
TLR2Agonism	LR	MQNs	0.827	0.641	0.325	0.748	0.827	0.641	0.325	0.748
TLR2Agonism	kNN	sf32	0.864	0.692	0.462	0.769	0.852	0.713	0.453	0.892
TLR2Agonism	SVC	sf32	0.840	0.649	0.358	0.861	0.864	0.692	0.462	0.885
TLR2Agonism	DT	sf32	0.827	0.641	0.325	0.641	0.840	0.677	0.391	0.706
TLR2Agonism	RF	sf32	0.815	0.606	0.254	0.832	0.864	0.664	0.440	0.856
TLR2Agonism	LR	sf32	0.827	0.585	0.245	0.851	0.827	0.585	0.245	0.851

*Continued on next page*

Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR2Antagonism	kNN	moe2d	0.727	0.735	0.462	0.791	0.773	0.774	0.540	0.774
TLR2Antagonism	SVC	moe2d	0.727	0.752	0.504	0.812	0.636	0.658	0.316	0.778
TLR2Antagonism	DT	moe2d	0.682	0.714	0.437	0.714	0.591	0.637	0.302	0.722
TLR2Antagonism	RF	moe2d	0.727	0.735	0.462	0.718	0.636	0.641	0.277	0.744
TLR2Antagonism	LR	moe2d	0.727	0.735	0.462	0.752	0.682	0.697	0.388	0.786
TLR2Antagonism	kNN	MQNs	0.636	0.624	0.248	0.748	0.682	0.662	0.332	0.756
TLR2Antagonism	SVC	MQNs	0.727	0.718	0.436	0.795	0.773	0.774	0.540	0.812
TLR2Antagonism	DT	MQNs	0.682	0.679	0.354	0.679	0.727	0.718	0.436	0.752
TLR2Antagonism	RF	MQNs	0.773	0.791	0.574	0.769	0.818	0.812	0.624	0.795
TLR2Antagonism	LR	MQNs	0.773	0.774	0.540	0.838	0.773	0.774	0.540	0.838
TLR2Antagonism	kNN	sf32	0.727	0.735	0.462	0.650	0.591	0.568	0.140	0.667
TLR2Antagonism	SVC	sf32	0.682	0.679	0.354	0.684	0.773	0.791	0.574	0.786
TLR2Antagonism	DT	sf32	0.455	0.487	-0.027	0.487	0.591	0.620	0.245	0.620
TLR2Antagonism	RF	sf32	0.682	0.679	0.354	0.748	0.727	0.752	0.504	0.778
TLR2Antagonism	LR	sf32	0.682	0.679	0.354	0.778	0.773	0.722	0.567	0.769

*Continued on next page*

Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR4Agonism	kNN	moe2d	0.842	0.755	0.526	0.882	0.833	0.793	0.550	0.855
TLR4Agonism	SVC	moe2d	0.789	0.693	0.385	0.839	0.807	0.718	0.436	0.862
TLR4Agonism	DT	moe2d	0.798	0.727	0.436	0.721	0.763	0.690	0.357	0.715
TLR4Agonism	RF	moe2d	0.798	0.727	0.436	0.810	0.825	0.729	0.473	0.858
TLR4Agonism	LR	moe2d	0.807	0.689	0.404	0.858	0.807	0.689	0.404	0.858
TLR4Agonism	kNN	MQNs	0.833	0.778	0.535	0.858	0.825	0.801	0.550	0.801
TLR4Agonism	SVC	MQNs	0.825	0.729	0.473	0.842	0.825	0.744	0.488	0.874
TLR4Agonism	DT	MQNs	0.860	0.795	0.590	0.788	0.781	0.701	0.387	0.842
TLR4Agonism	RF	MQNs	0.877	0.806	0.632	0.873	0.825	0.686	0.433	0.854
TLR4Agonism	LR	MQNs	0.807	0.704	0.420	0.840	0.842	0.741	0.514	0.841
TLR4Agonism	kNN	sf32	0.807	0.776	0.503	0.859	0.833	0.793	0.550	0.873
TLR4Agonism	SVC	sf32	0.825	0.729	0.473	0.833	0.781	0.629	0.294	0.751
TLR4Agonism	DT	sf32	0.754	0.670	0.323	0.669	0.789	0.678	0.367	0.789
TLR4Agonism	RF	sf32	0.860	0.766	0.568	0.845	0.816	0.666	0.396	0.853
TLR4Agonism	LR	sf32	0.789	0.635	0.314	0.819	0.789	0.635	0.314	0.819

*Continued on next page*

Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR4Antagonism	kNN	moe2d	0.822	0.812	0.604	0.891	0.844	0.828	0.645	0.919
TLR4Antagonism	SVC	moe2d	0.822	0.793	0.585	0.926	0.844	0.828	0.645	0.853
TLR4Antagonism	DT	moe2d	0.756	0.685	0.400	0.685	0.733	0.709	0.403	0.760
TLR4Antagonism	RF	moe2d	0.867	0.825	0.680	0.876	0.778	0.682	0.441	0.906
TLR4Antagonism	LR	moe2d	0.800	0.757	0.525	0.843	0.800	0.757	0.525	0.843
TLR4Antagonism	kNN	MQNs	0.733	0.709	0.403	0.810	0.800	0.757	0.525	0.829
TLR4Antagonism	SVC	MQNs	0.800	0.757	0.525	0.843	0.733	0.709	0.403	0.737
TLR4Antagonism	DT	MQNs	0.711	0.673	0.339	0.673	0.756	0.744	0.466	0.795
TLR4Antagonism	RF	MQNs	0.822	0.793	0.585	0.927	0.844	0.789	0.623	0.931
TLR4Antagonism	LR	MQNs	0.689	0.657	0.303	0.806	0.689	0.657	0.303	0.806
TLR4Antagonism	kNN	sf32	0.756	0.725	0.441	0.838	0.756	0.725	0.441	0.844
TLR4Antagonism	SVC	sf32	0.822	0.773	0.572	0.919	0.756	0.744	0.466	0.776
TLR4Antagonism	DT	sf32	0.778	0.741	0.482	0.741	0.711	0.712	0.397	0.729
TLR4Antagonism	RF	sf32	0.778	0.721	0.463	0.877	0.756	0.666	0.384	0.882
TLR4Antagonism	LR	sf32	0.822	0.793	0.585	0.816	0.822	0.793	0.585	0.816

*Continued on next page*

Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR7Agonism	kNN	moe2d	0.835	0.830	0.677	0.883	0.857	0.853	0.718	0.905
TLR7Agonism	SVC	moe2d	0.835	0.834	0.669	0.909	0.813	0.811	0.625	0.909
TLR7Agonism	DT	moe2d	0.857	0.855	0.714	0.855	0.824	0.824	0.648	0.881
TLR7Agonism	RF	moe2d	0.830	0.829	0.658	0.907	0.835	0.835	0.669	0.918
TLR7Agonism	LR	moe2d	0.802	0.800	0.603	0.891	0.802	0.800	0.603	0.891
TLR7Agonism	kNN	MQNs	0.791	0.786	0.585	0.869	0.813	0.810	0.627	0.870
TLR7Agonism	SVC	MQNs	0.802	0.799	0.605	0.857	0.775	0.772	0.548	0.828
TLR7Agonism	DT	MQNs	0.808	0.806	0.614	0.806	0.813	0.808	0.630	0.850
TLR7Agonism	RF	MQNs	0.819	0.816	0.637	0.890	0.797	0.795	0.592	0.885
TLR7Agonism	LR	MQNs	0.725	0.723	0.448	0.807	0.753	0.750	0.503	0.783
TLR7Agonism	kNN	sf32	0.797	0.793	0.593	0.879	0.808	0.806	0.614	0.888
TLR7Agonism	SVC	sf32	0.802	0.800	0.603	0.877	0.791	0.788	0.581	0.854
TLR7Agonism	DT	sf32	0.742	0.739	0.481	0.739	0.775	0.770	0.550	0.817
TLR7Agonism	RF	sf32	0.824	0.825	0.649	0.893	0.808	0.806	0.614	0.905
TLR7Agonism	LR	sf32	0.747	0.745	0.492	0.787	0.742	0.739	0.481	0.780

*Continued on next page*

Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR7Antagonism	kNN	moe2d	0.810	0.792	0.612	0.838	0.762	0.736	0.517	0.880
TLR7Antagonism	SVC	moe2d	0.667	0.639	0.304	0.796	0.667	0.653	0.311	0.648
TLR7Antagonism	DT	moe2d	0.714	0.681	0.420	0.681	0.714	0.722	0.440	0.718
TLR7Antagonism	RF	moe2d	0.762	0.764	0.523	0.866	0.667	0.653	0.311	0.806
TLR7Antagonism	LR	moe2d	0.667	0.653	0.311	0.704	0.667	0.653	0.311	0.704
TLR7Antagonism	kNN	MQNs	0.857	0.833	0.730	0.861	0.667	0.639	0.304	0.741
TLR7Antagonism	SVC	MQNs	0.810	0.778	0.645	0.898	0.714	0.694	0.408	0.722
TLR7Antagonism	DT	MQNs	0.714	0.681	0.420	0.681	0.667	0.639	0.304	0.755
TLR7Antagonism	RF	MQNs	0.571	0.556	0.113	0.653	0.714	0.681	0.420	0.806
TLR7Antagonism	LR	MQNs	0.714	0.694	0.408	0.722	0.714	0.694	0.408	0.722
TLR7Antagonism	kNN	sf32	0.762	0.736	0.517	0.838	0.714	0.708	0.417	0.708
TLR7Antagonism	SVC	sf32	0.714	0.681	0.420	0.861	0.810	0.792	0.612	0.787
TLR7Antagonism	DT	sf32	0.667	0.639	0.304	0.639	0.714	0.681	0.420	0.667
TLR7Antagonism	RF	sf32	0.667	0.653	0.311	0.731	0.714	0.694	0.408	0.806
TLR7Antagonism	LR	sf32	0.810	0.792	0.612	0.731	0.810	0.792	0.612	0.731

*Continued on next page*

Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR8Agonism	kNN	moe2d	0.813	0.809	0.636	0.901	0.835	0.830	0.682	0.909
TLR8Agonism	SVC	moe2d	0.827	0.825	0.659	0.898	0.863	0.863	0.726	0.909
TLR8Agonism	DT	moe2d	0.806	0.806	0.612	0.806	0.734	0.729	0.477	0.810
TLR8Agonism	RF	moe2d	0.856	0.854	0.714	0.924	0.871	0.869	0.743	0.929
TLR8Agonism	LR	moe2d	0.748	0.748	0.496	0.852	0.763	0.760	0.527	0.822
TLR8Agonism	kNN	MQNs	0.763	0.759	0.531	0.872	0.820	0.817	0.646	0.906
TLR8Agonism	SVC	MQNs	0.784	0.783	0.568	0.892	0.791	0.792	0.585	0.870
TLR8Agonism	DT	MQNs	0.849	0.845	0.712	0.843	0.835	0.834	0.669	0.907
TLR8Agonism	RF	MQNs	0.842	0.839	0.686	0.910	0.820	0.818	0.644	0.924
TLR8Agonism	LR	MQNs	0.791	0.791	0.582	0.845	0.791	0.791	0.582	0.845
TLR8Agonism	kNN	sf32	0.791	0.789	0.584	0.875	0.827	0.823	0.666	0.926
TLR8Agonism	SVC	sf32	0.799	0.797	0.598	0.898	0.777	0.776	0.553	0.850
TLR8Agonism	DT	sf32	0.770	0.766	0.544	0.766	0.842	0.841	0.683	0.887
TLR8Agonism	RF	sf32	0.856	0.855	0.712	0.923	0.856	0.854	0.715	0.918
TLR8Agonism	LR	sf32	0.734	0.733	0.467	0.827	0.712	0.711	0.423	0.825

*Continued on next page*

Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR8Antagonism	kNN	moe2d	0.684	0.681	0.365	0.808	0.658	0.653	0.312	0.744
TLR8Antagonism	SVC	moe2d	0.737	0.739	0.478	0.797	0.711	0.711	0.422	0.808
TLR8Antagonism	DT	moe2d	0.632	0.639	0.288	0.639	0.658	0.667	0.351	0.649
TLR8Antagonism	RF	moe2d	0.632	0.636	0.275	0.731	0.737	0.733	0.472	0.794
TLR8Antagonism	LR	moe2d	0.711	0.714	0.430	0.822	0.711	0.711	0.422	0.797
TLR8Antagonism	kNN	MQNs	0.684	0.675	0.376	0.661	0.605	0.603	0.206	0.603
TLR8Antagonism	SVC	MQNs	0.605	0.594	0.208	0.647	0.658	0.661	0.324	0.678
TLR8Antagonism	DT	MQNs	0.553	0.561	0.129	0.597	0.658	0.664	0.335	0.640
TLR8Antagonism	RF	MQNs	0.553	0.558	0.119	0.624	0.605	0.608	0.218	0.647
TLR8Antagonism	LR	MQNs	0.632	0.631	0.261	0.681	0.632	0.631	0.261	0.681
TLR8Antagonism	kNN	sf32	0.763	0.767	0.536	0.756	0.658	0.658	0.316	0.678
TLR8Antagonism	SVC	sf32	0.711	0.714	0.430	0.781	0.737	0.739	0.478	0.781
TLR8Antagonism	DT	sf32	0.632	0.633	0.267	0.633	0.737	0.744	0.506	0.783
TLR8Antagonism	RF	sf32	0.632	0.633	0.267	0.693	0.737	0.739	0.478	0.794
TLR8Antagonism	LR	sf32	0.711	0.711	0.422	0.742	0.684	0.689	0.382	0.739

*Continued on next page*



Table 6.2 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				random search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc

<sup>a</sup>: balanced accuracy.

**Table 6.3:** Machine learning study for *ToIIB* with grid search

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy	MCC	roc_auc	accuracy	b_accuracy	MCC	roc_auc
TLR2Agonism	kNN	moe2d	0.889	0.763	0.581	0.896	0.840	0.790	0.518	0.790
TLR2Agonism	SVC	moe2d	0.864	0.664	0.440	0.913	0.864	0.692	0.462	0.919
TLR2Agonism	DT	moe2d	0.852	0.713	0.453	0.746	0.840	0.705	0.423	0.763
TLR2Agonism	RF	moe2d	0.840	0.677	0.391	0.886	0.852	0.656	0.396	0.906
TLR2Agonism	LR	moe2d	0.877	0.756	0.545	0.907	0.840	0.592	0.290	0.853
TLR2Agonism	kNN	MQNs	0.852	0.684	0.424	0.801	0.840	0.677	0.391	0.720
TLR2Agonism	SVC	MQNs	0.864	0.692	0.462	0.844	0.827	0.641	0.325	0.827
TLR2Agonism	DT	MQNs	0.852	0.741	0.482	0.735	0.840	0.620	0.324	0.819
TLR2Agonism	RF	MQNs	0.852	0.713	0.453	0.784	0.852	0.684	0.424	0.920
TLR2Agonism	LR	MQNs	0.827	0.641	0.325	0.748	0.827	0.641	0.325	0.748
TLR2Agonism	kNN	sf32	0.864	0.692	0.462	0.769	0.877	0.756	0.545	0.812
TLR2Agonism	SVC	sf32	0.840	0.649	0.358	0.861	0.864	0.692	0.462	0.906
TLR2Agonism	DT	sf32	0.815	0.634	0.295	0.634	0.840	0.677	0.391	0.761
TLR2Agonism	RF	sf32	0.852	0.656	0.396	0.814	0.864	0.664	0.440	0.856
TLR2Agonism	LR	sf32	0.827	0.585	0.245	0.851	0.827	0.585	0.245	0.851

*Continued on next page*

Table 6.3 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR2Antagonism	kNN	moe2d	0.727	0.735	0.462	0.791	0.773	0.774	0.540	0.774
TLR2Antagonism	SVC	moe2d	0.727	0.752	0.504	0.812	0.636	0.658	0.316	0.778
TLR2Antagonism	DT	moe2d	0.636	0.658	0.316	0.658	0.545	0.564	0.128	0.470
TLR2Antagonism	RF	moe2d	0.773	0.808	0.629	0.718	0.682	0.697	0.388	0.726
TLR2Antagonism	LR	moe2d	0.727	0.735	0.462	0.752	0.682	0.697	0.388	0.786
TLR2Antagonism	kNN	MQNs	0.636	0.624	0.248	0.748	0.636	0.624	0.248	0.765
TLR2Antagonism	SVC	MQNs	0.727	0.718	0.436	0.795	0.773	0.774	0.540	0.838
TLR2Antagonism	DT	MQNs	0.682	0.679	0.354	0.679	0.773	0.756	0.524	0.795
TLR2Antagonism	RF	MQNs	0.682	0.679	0.354	0.705	0.818	0.812	0.624	0.821
TLR2Antagonism	LR	MQNs	0.773	0.774	0.540	0.838	0.773	0.774	0.540	0.838
TLR2Antagonism	kNN	sf32	0.727	0.735	0.462	0.650	0.682	0.679	0.354	0.731
TLR2Antagonism	SVC	sf32	0.682	0.679	0.354	0.684	0.773	0.791	0.574	0.786
TLR2Antagonism	DT	sf32	0.500	0.543	0.094	0.543	0.636	0.675	0.370	0.585
TLR2Antagonism	RF	sf32	0.636	0.641	0.277	0.705	0.682	0.714	0.437	0.786
TLR2Antagonism	LR	sf32	0.682	0.679	0.354	0.778	0.773	0.722	0.567	0.769

*Continued on next page*

Table 6.3 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR4Agonism	kNN	moe2d	0.842	0.755	0.526	0.882	0.833	0.764	0.520	0.878
TLR4Agonism	SVC	moe2d	0.789	0.693	0.385	0.839	0.842	0.741	0.514	0.869
TLR4Agonism	DT	moe2d	0.825	0.787	0.534	0.785	0.763	0.618	0.257	0.749
TLR4Agonism	RF	moe2d	0.842	0.741	0.514	0.834	0.868	0.772	0.592	0.877
TLR4Agonism	LR	moe2d	0.807	0.689	0.404	0.858	0.807	0.689	0.404	0.858
TLR4Agonism	kNN	MQNs	0.833	0.778	0.535	0.858	0.825	0.801	0.550	0.801
TLR4Agonism	SVC	MQNs	0.825	0.729	0.473	0.842	0.807	0.718	0.436	0.830
TLR4Agonism	DT	MQNs	0.851	0.789	0.571	0.804	0.789	0.693	0.385	0.823
TLR4Agonism	RF	MQNs	0.851	0.775	0.558	0.858	0.842	0.769	0.539	0.880
TLR4Agonism	LR	MQNs	0.807	0.704	0.420	0.840	0.842	0.741	0.514	0.841
TLR4Agonism	kNN	sf32	0.807	0.776	0.503	0.859	0.833	0.793	0.550	0.873
TLR4Agonism	SVC	sf32	0.825	0.729	0.473	0.833	0.746	0.593	0.201	0.822
TLR4Agonism	DT	sf32	0.781	0.716	0.405	0.716	0.807	0.704	0.420	0.839
TLR4Agonism	RF	sf32	0.842	0.755	0.526	0.881	0.860	0.752	0.559	0.874
TLR4Agonism	LR	sf32	0.789	0.635	0.314	0.819	0.789	0.635	0.314	0.819

*Continued on next page*

Table 6.3 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR4Antagonism	kNN	moe2d	0.822	0.812	0.604	0.891	0.822	0.812	0.604	0.862
TLR4Antagonism	SVC	moe2d	0.822	0.793	0.585	0.926	0.733	0.709	0.403	0.836
TLR4Antagonism	DT	moe2d	0.800	0.718	0.504	0.718	0.711	0.692	0.367	0.740
TLR4Antagonism	RF	moe2d	0.756	0.666	0.384	0.843	0.800	0.737	0.511	0.917
TLR4Antagonism	LR	moe2d	0.800	0.757	0.525	0.843	0.800	0.757	0.525	0.843
TLR4Antagonism	kNN	MQNs	0.733	0.709	0.403	0.810	0.778	0.741	0.482	0.803
TLR4Antagonism	SVC	MQNs	0.800	0.757	0.525	0.843	0.756	0.725	0.441	0.712
TLR4Antagonism	DT	MQNs	0.733	0.630	0.315	0.630	0.800	0.796	0.565	0.847
TLR4Antagonism	RF	MQNs	0.867	0.825	0.680	0.894	0.889	0.841	0.735	0.942
TLR4Antagonism	LR	MQNs	0.689	0.657	0.303	0.806	0.689	0.657	0.303	0.806
TLR4Antagonism	kNN	sf32	0.756	0.725	0.441	0.838	0.800	0.737	0.511	0.737
TLR4Antagonism	SVC	sf32	0.822	0.773	0.572	0.918	0.800	0.796	0.565	0.788
TLR4Antagonism	DT	sf32	0.711	0.614	0.264	0.614	0.756	0.764	0.495	0.767
TLR4Antagonism	RF	sf32	0.800	0.757	0.525	0.829	0.800	0.757	0.525	0.857
TLR4Antagonism	LR	sf32	0.822	0.793	0.585	0.816	0.822	0.793	0.585	0.816

*Continued on next page*

Table 6.3 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR7Agonism	kNN	moe2d	0.835	0.830	0.677	0.883	0.857	0.853	0.718	0.905
TLR7Agonism	SVC	moe2d	0.835	0.834	0.669	0.909	0.857	0.855	0.714	0.923
TLR7Agonism	DT	moe2d	0.835	0.833	0.670	0.833	0.813	0.811	0.625	0.890
TLR7Agonism	RF	moe2d	0.863	0.864	0.726	0.925	0.846	0.845	0.691	0.923
TLR7Agonism	LR	moe2d	0.802	0.800	0.603	0.891	0.802	0.800	0.603	0.891
TLR7Agonism	kNN	MQNs	0.791	0.786	0.585	0.869	0.830	0.827	0.659	0.863
TLR7Agonism	SVC	MQNs	0.802	0.799	0.605	0.857	0.824	0.821	0.649	0.862
TLR7Agonism	DT	MQNs	0.808	0.806	0.614	0.806	0.824	0.820	0.651	0.872
TLR7Agonism	RF	MQNs	0.819	0.818	0.636	0.889	0.791	0.788	0.581	0.880
TLR7Agonism	LR	MQNs	0.725	0.723	0.448	0.807	0.753	0.750	0.503	0.783
TLR7Agonism	kNN	sf32	0.797	0.793	0.593	0.879	0.808	0.806	0.614	0.888
TLR7Agonism	SVC	sf32	0.802	0.800	0.603	0.877	0.775	0.772	0.548	0.844
TLR7Agonism	DT	sf32	0.769	0.768	0.537	0.767	0.775	0.770	0.552	0.804
TLR7Agonism	RF	sf32	0.852	0.851	0.702	0.901	0.835	0.833	0.669	0.907
TLR7Agonism	LR	sf32	0.747	0.745	0.492	0.787	0.742	0.739	0.481	0.780

*Continued on next page*

Table 6.3 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR7Antagonism	kNN	moe2d	0.810	0.792	0.612	0.838	0.714	0.694	0.408	0.778
TLR7Antagonism	SVC	moe2d	0.667	0.639	0.304	0.796	0.667	0.653	0.311	0.676
TLR7Antagonism	DT	moe2d	0.714	0.681	0.420	0.681	0.810	0.792	0.612	0.769
TLR7Antagonism	RF	moe2d	0.714	0.708	0.417	0.810	0.762	0.750	0.510	0.833
TLR7Antagonism	LR	moe2d	0.667	0.653	0.311	0.704	0.667	0.653	0.311	0.704
TLR7Antagonism	kNN	MQNs	0.857	0.833	0.730	0.861	0.667	0.639	0.304	0.741
TLR7Antagonism	SVC	MQNs	0.810	0.778	0.645	0.898	0.714	0.694	0.408	0.722
TLR7Antagonism	DT	MQNs	0.762	0.736	0.517	0.736	0.714	0.694	0.408	0.819
TLR7Antagonism	RF	MQNs	0.667	0.653	0.311	0.796	0.714	0.681	0.420	0.806
TLR7Antagonism	LR	MQNs	0.714	0.694	0.408	0.722	0.714	0.694	0.408	0.722
TLR7Antagonism	kNN	sf32	0.762	0.736	0.517	0.838	0.810	0.778	0.645	0.815
TLR7Antagonism	SVC	sf32	0.714	0.681	0.420	0.861	0.762	0.750	0.510	0.750
TLR7Antagonism	DT	sf32	0.619	0.597	0.204	0.597	0.762	0.750	0.510	0.796
TLR7Antagonism	RF	sf32	0.714	0.708	0.417	0.801	0.762	0.750	0.510	0.833
TLR7Antagonism	LR	sf32	0.810	0.792	0.612	0.731	0.810	0.792	0.612	0.731

*Continued on next page*

Table 6.3 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR8Agonism	kNN	moe2d	0.813	0.809	0.636	0.901	0.835	0.830	0.682	0.909
TLR8Agonism	SVC	moe2d	0.827	0.825	0.659	0.898	0.842	0.841	0.683	0.902
TLR8Agonism	DT	moe2d	0.820	0.821	0.641	0.821	0.842	0.841	0.683	0.872
TLR8Agonism	RF	moe2d	0.835	0.834	0.669	0.918	0.871	0.868	0.745	0.933
TLR8Agonism	LR	moe2d	0.748	0.748	0.496	0.852	0.763	0.760	0.527	0.822
TLR8Agonism	kNN	MQNs	0.763	0.759	0.531	0.872	0.849	0.846	0.705	0.911
TLR8Agonism	SVC	MQNs	0.784	0.783	0.568	0.892	0.799	0.797	0.597	0.878
TLR8Agonism	DT	MQNs	0.878	0.875	0.763	0.879	0.856	0.855	0.712	0.884
TLR8Agonism	RF	MQNs	0.863	0.860	0.734	0.935	0.813	0.810	0.630	0.916
TLR8Agonism	LR	MQNs	0.791	0.791	0.582	0.845	0.791	0.791	0.582	0.845
TLR8Agonism	kNN	sf32	0.791	0.789	0.584	0.875	0.827	0.823	0.666	0.926
TLR8Agonism	SVC	sf32	0.799	0.797	0.598	0.898	0.799	0.797	0.597	0.898
TLR8Agonism	DT	sf32	0.791	0.789	0.585	0.789	0.784	0.783	0.568	0.854
TLR8Agonism	RF	sf32	0.856	0.855	0.712	0.925	0.849	0.847	0.700	0.934
TLR8Agonism	LR	sf32	0.734	0.733	0.467	0.827	0.712	0.711	0.423	0.825

*Continued on next page*



Table 6.3 – *Continued from previous page*

prediction_task	algorithm	descriptors	default parameter				grid search			
			accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc	accuracy	b_accuracy <sup>a</sup>	MCC	roc_auc
TLR8Antagonism	kNN	moe2d	0.684	0.681	0.365	0.808	0.658	0.653	0.312	0.744
TLR8Antagonism	SVC	moe2d	0.737	0.739	0.478	0.797	0.711	0.711	0.422	0.806
TLR8Antagonism	DT	moe2d	0.658	0.667	0.351	0.667	0.658	0.667	0.351	0.633
TLR8Antagonism	RF	moe2d	0.632	0.633	0.267	0.668	0.763	0.764	0.527	0.761
TLR8Antagonism	LR	moe2d	0.711	0.714	0.430	0.822	0.711	0.711	0.422	0.797
TLR8Antagonism	kNN	MQNs	0.684	0.675	0.376	0.661	0.605	0.603	0.206	0.603
TLR8Antagonism	SVC	MQNs	0.605	0.594	0.208	0.647	0.658	0.658	0.316	0.694
TLR8Antagonism	DT	MQNs	0.553	0.561	0.129	0.597	0.605	0.608	0.218	0.550
TLR8Antagonism	RF	MQNs	0.553	0.561	0.129	0.712	0.632	0.633	0.267	0.631
TLR8Antagonism	LR	MQNs	0.632	0.631	0.261	0.681	0.632	0.631	0.261	0.681
TLR8Antagonism	kNN	sf32	0.763	0.767	0.536	0.756	0.658	0.658	0.316	0.678
TLR8Antagonism	SVC	sf32	0.711	0.714	0.430	0.781	0.737	0.739	0.478	0.775
TLR8Antagonism	DT	sf32	0.684	0.689	0.382	0.689	0.684	0.683	0.367	0.686
TLR8Antagonism	RF	sf32	0.737	0.744	0.506	0.781	0.711	0.711	0.422	0.758
TLR8Antagonism	LR	sf32	0.711	0.711	0.422	0.742	0.684	0.689	0.382	0.739



# Bibliography

- [1] Hansson, G. K. and Edfeldt, K. Toll to be paid at the gateway to the vessel wall. *Arterioscler., Thromb., Vasc. Biol.*, 25(6):1085–1087, 2005.
- [2] Kawai, T. and Akira, S. Toll-like receptors and their crosstalk with other innate receptors in infection and immunity. *Immunity*, 34(5):637–650, 2011.
- [3] Kawasaki, T. and Kawai, T. Toll-like receptor signaling pathways. *Front. Immunol.*, 5:461–461, 2014.
- [4] Kagan, J. C., Su, T., Horng, T., Chow, A., Akira, S., and Medzhitov, R. TRAM couples endocytosis of Toll-like receptor 4 to the induction of interferon- $\beta$ . *Nat. Immunol.*, 9(4):361–368, 2008.
- [5] Gao, W., Xiong, Y., Li, Q., and Yang, H. Inhibition of toll-like receptor signaling as a promising therapy for inflammatory diseases: a journey from molecular to nano therapeutics. *Front. Physiol.*, 8(508):508, 2017.
- [6] Hasan, U., Chaffois, C., Gaillard, C., Saulnier, V., Merck, E., Tancredi, S., Guiet, C., Brière, F., Vlach, J., Lebecque, S., et al. Human TLR10 is a functional receptor, expressed by B cells and plasmacytoid dendritic cells, which activates gene transcription through MyD88. *J. Immunol.*, 174(5):2942–2950, 2005.
- [7] Vijay, K. Toll-like receptors in immunity and inflammatory diseases: Past, present, and future. *Int. Immunopharmacol.*, 59:391–412, 2018.
- [8] Anwar, M. A., Shah, M., Kim, J., and Choi, S. Recent clinical trends in Toll-like receptor targeting therapeutics. *Med. Res. Rev.*, 39(3):1053–1090, 2019.
- [9] Murgueitio, M. S., Rakers, C., Frank, A., and Wolber, G. Balancing inflammation:

- computational design of small-molecule toll-like receptor modulators. *Trends Pharmacol. Sci.*, 38(2):155–168, 2017.
- [10] Akira, S., Takeda, K., and Kaisho, T. Toll-like receptors: critical proteins linking innate and acquired immunity. *Nat. Immunol.*, 2:675–680, 2001.
- [11] Bukau, B. and Horwich, A. L. The HSP70 and HSP60 chaperone machines. *Cell*, 92(3):351–366, 1998.
- [12] Zhu, G., Xu, Y., Cen, X., Nandakumar, K. S., Liu, S., and Cheng, K. Targeting pattern-recognition receptors to discover new small molecule immune modulators. *Eur. J. Med. Chem.*, 144:82–92, 2018.
- [13] David, C. M. Latest advances in small molecule TLR 7/8 agonist drug research. *Curr. Top. Med. Chem. (Sharjah, United Arab Emirates)*, 19(24):2228–2238, 2019.
- [14] Sriskandan, S. and Altmann, D. M. The immunology of sepsis. *J. Pathol.*, 214(2):211–223, 2008.
- [15] Dasu, M. R., Ramirez, S., and Isseroff, R. R. Toll-like receptors and diabetes: a therapeutic perspective. *Clin. Sci.*, 122(5):203–214, 2012.
- [16] Jialal, I., Huet, B. A., Kaur, H., Chien, A., and Devaraj, S. Increased toll-like receptor activity in patients with metabolic syndrome. *Diabetes Care*, 35(4):900–904, 2012.
- [17] Persing, D. H., Coler, R. N., Lacy, M. J., Johnson, D. A., Baldrige, J. R., Hershberg, R. M., and Reed, S. G. Taking toll: lipid A mimetics as adjuvants and immunomodulators. *Trends Microbiol.*, 10(10):s32–s37, 2002.
- [18] Hussein, W. M., Liu, T.-Y., Skwarczynski, M., and Toth, I. Toll-like receptor agonists: a patent review (2011-2013). *Expert Opin. Ther. Pat.*, 24(4):453–470, 2014.
- [19] Hennessy, E. J., Parker, A. E., and O’Neill, L. A. J. Targeting Toll-like receptors: emerging therapeutics? *Nat. Rev. Drug Discovery*, 9(4):293–307, 2010.
- [20] Hadley, G., Derry, S., and Moore, R. A. Imiquimod for actinic keratosis: systematic review and meta-analysis. *J. Invest. Dermatol.*, 126(6):1251–1255, 2006.

- [21] Tyring, S. K., Arany, I., Stanley, M. A., Tomai, M. A., Miller, R. L., Smith, M. H., McDermott, D. J., and Slade, H. B. A randomized, controlled, molecular study of condylomata acuminata clearance during treatment with imiquimod. *J. Infect. Dis.*, 178(2):551–555, 1998.
- [22] Geisse, J., Caro, I., Lindholm, J., Golitz, L., Stampone, P., and Owens, M. Imiquimod 5% cream for the treatment of superficial basal cell carcinoma: results from two phase III, randomized, vehicle-controlled studies. *J. Am. Acad. Dermatol.*, 50(5):722–733, 2004.
- [23] Rook, A. H., Gelfand, J. M., Wysocka, M., Troxel, A. B., Benoit, B., Surber, C., Elenitsas, R., Buchanan, M. A., Leahy, D. S., Watanabe, R., Kirsch, I. R., Kim, E. J., and Clark, R. A. Topical resiquimod can induce disease regression and enhance T-cell effector functions in cutaneous T-cell lymphoma. *Blood*, 126(12):1452–1461, 2015.
- [24] European Medicines Agency: EMA/COMP/240708/2016, EU designation number EU/3/16/1653. Public summary of opinion on orphan designation: Resiquimod for the treatment of cutaneous T-cell lymphoma, 2016. URL <https://www.ema.europa.eu/en/medicines/human/orphan-designations/eu3161653>. Accessed: 2019-09-30.
- [25] Smith, M., Haney, E., McDonagh, M., and et al. Treatment of myalgic encephalomyelitis/chronic fatigue syndrome: a systematic review for a National Institutes of Health Pathways to Prevention Workshop. *Ann. Intern. Med.*, 162(12):841–850, 2015.
- [26] Gao, D. and Li, W. Structures and recognition modes of toll-like receptors. *Proteins*, 85(1):3–9, 2017.
- [27] Botos, I., Segal, D., and Davies, D. The structural biology of Toll-like receptors. *Structure*, 19(4):447–459, 2011.
- [28] Choe, J., Kelker, M. S., and Wilson, I. A. Crystal structure of human toll-like receptor 3 (TLR3) ectodomain. *Science*, 309(5734):581–585, 2005.

- [29] Bell, J. K., Botos, I., Hall, P. R., Askins, J., Shiloach, J., Segal, D. M., and Davies, D. R. The molecular structure of the Toll-like receptor 3 ligand-binding domain. *Proc. Natl. Acad. Sci. U. S. A.*, 102(31):10976–10980, 2005.
- [30] Jin, M. S. and Lee, J.-O. Structures of TLR–ligand complexes. *Curr. Opin. Immunol.*, 20(4):414–419, 2008.
- [31] Kang, J. Y. and Lee, J.-O. Structural biology of the Toll-like receptor family. *Annu. Rev. Biochem.*, 80(1):917–941, 2011.
- [32] O’Neill, L. A., Golenbock, D., and Bowie, A. G. The history of Toll-like receptors—redefining innate immunity. *Nat. Rev. Immunol.*, 13(6):453–460, 2013.
- [33] Lin, S.-C., Lo, Y.-C., and Wu, H. Helical assembly in the MyD88–IRAK4–IRAK2 complex in TLR/IL-1R signalling. *Nature*, 465(7300):885–890, 2010.
- [34] Vidya, M. K., Kumar, V. G., Sejian, V., Bagath, M., Krishnan, G., and Bhatta, R. Toll-like receptors: significance, ligands, signaling pathways, and functions in mammals. *Int. Rev. Immunol.*, 37(1):20–36, 2018.
- [35] Hornung, V., Rothenfusser, S., Britsch, S., Krug, A., Jahrsdörfer, B., Giese, T., Endres, S., and Hartmann, G. Quantitative expression of toll-like receptor 1–10 mRNA in cellular subsets of human peripheral blood mononuclear cells and sensitivity to CpG oligodeoxynucleotides. *J. Immunol.*, 168(9):4531–4537, 2002.
- [36] Mulder, W. J. M., Ochando, J., Joosten, L. A. B., Fayad, Z. A., and Netea, M. G. Therapeutic targeting of trained immunity. *Nat. Rev. Drug Discovery*, 18(7):553–566, 2019.
- [37] Cook, D. N., Pisetsky, D. S., and Schwartz, D. A. Toll-like receptors in the pathogenesis of human disease. *Nat. Immunol.*, 5(10):975–979, 2004.
- [38] Kanzler, H., Barrat, F. J., Hessel, E. M., and Coffman, R. L. Therapeutic targeting of innate immunity with Toll-like receptor agonists and antagonists. *Nat. Med. (N. Y., NY, U. S.)*, 13(5):552–559, 2007.
- [39] Romagne, F. Current and future drugs targeting one class of innate immunity receptors: the Toll-like receptors. *Drug Discovery Today*, 12(1):80–87, 2007.

- [40] Neill, L. A. J., Bryant, C. E., and Doyle, S. L. Therapeutic targeting of Toll-like receptors for infectious and inflammatory diseases and cancer. *Pharmacol. Rev.*, 61(2):177–197, 2009.
- [41] Wang, X., Smith, C., and Yin, H. Targeting Toll-like receptors with small molecule agents. *Chem. Soc. Rev.*, 42(12):4859–4866, 2013.
- [42] Shukla, N. M., Chan, M., Hayashi, T., Carson, D. A., and Cottam, H. B. Recent advances and perspectives in small-molecule TLR ligands and their modulators. *ACS Med. Chem. Lett.*, 9(12):1156–1159, 2018.
- [43] Makkouk, A. and Abdelnoor, A. M. The potential use of Toll-like receptor (TLR) agonists and antagonists as prophylactic and/or therapeutic agents. *Immunopharmacol. Immunotoxicol.*, 31(3):331–338, 2009.
- [44] Cen, X., Liu, S., and Cheng, K. The role of toll-like receptor in inflammation and tumor immunity. *Front. Pharmacol.*, 9(878):878, 2018.
- [45] Narayanankutty, A. Toll-like receptors as a novel therapeutic target for natural products against chronic diseases. *Curr. Drug Targets*, 20(10):1068–1080, 2019.
- [46] Matsumoto, M., Takeda, Y., and Seya, T. Targeting Toll-like receptor 3 in dendritic cells for cancer immunotherapy. *Expert Opin. Biol. Ther.*, pages 1–10, 2020.
- [47] Aryan, Z. and Rezaei, N. Toll-like receptors as targets for allergen immunotherapy. *Curr. Opin. Allergy Clin. Immunol.*, 15(6):568–574, 2015.
- [48] Schmidt, M., Hagner, N., Marco, A., König-Merediz, S. A., Schroff, M., and Wittig, B. Design and structural requirements of the potent and safe TLR-9 agonistic immunomodulator MGN1703. *Nucleic Acid Ther.*, 25(3):130–140, 2015.
- [49] Shah, M., Anwar, M. A., Kim, J. H., and Choi, S. Advances in antiviral therapies targeting Toll-like receptors. *Expert Opin. Invest. Drugs*, 25(4):437–453, 2016.
- [50] Embrechts, W., Herschke, F., Pauwels, F., Stoops, B., Last, S., Pieters, S., Pande, V., Pille, G., Amssoms, K., Smyej, I., Dhuyvetter, D., Scholliers, A., Mostmans, W., Van Dijck, K., Van Schoubroeck, B., Thone, T., De Pooter, D.,

- Fanning, G., Jonckers, T. H. M., Horton, H., Raboisson, P., and McGowan, D. 2,4-diaminoquinazolines as dual toll-like receptor (tlr) 7/8 modulators for the treatment of hepatitis b virus. *Journal of Medicinal Chemistry*, 61(14):6236–6246, 2018.
- [51] Martinsen, J. T., Gunst, J. D., Højen, J. F., Tolstrup, M., and Søgaard, O. S. The use of Toll-like receptor agonists in HIV-1 cure strategies. *Front. Immunol.*, 11, 2020.
- [52] Kay, E., Scotland, R. S., and Whiteford, J. R. Toll-like receptors: Role in inflammation and therapeutic potential. *Biofactors*, 40(3):284–294, 2014.
- [53] Gao, W., Xiong, Y., Li, Q., and Yang, H. Inhibition of toll-like receptor signaling as a promising therapy for inflammatory diseases: a journey from molecular to nano therapeutics. *Front. Physiol.*, 8:508, 2017.
- [54] Wietzorrek, G., Drexel, M., Trieb, M., and Santos-Sierra, S. Anti-inflammatory activity of small-molecule antagonists of Toll-like receptor 2 (TLR2) in mice. *Immunobiology*, 224(1):1–9, 2019.
- [55] Hutchinson, M. R., Zhang, Y., Brown, K., Coats, B. D., Shridhar, M., Sholar, P. W., Patel, S. J., Crysedale, N. Y., Harrison, J. A., Maier, S. F., Rice, K. C., and Watkins, L. R. Non-stereoselective reversal of neuropathic pain by naloxone and naltrexone: involvement of toll-like receptor 4 (TLR4). *Eur. J. Neurosci.*, 28(1):20–29, 2008.
- [56] Wang, X., Grace, P. M., Pham, M. N., Cheng, K., Strand, K. A., Smith, C., Li, J., Watkins, L. R., and Yin, H. Rifampin inhibits toll-like receptor 4 signaling by targeting myeloid differentiation protein 2 and attenuates neuropathic pain. *FASEB J.*, 27(7):2713–2722, 2013.
- [57] Frasca, L. and Lande, R. Toll-like receptors in mediating pathogenesis in systemic sclerosis. *Clin. Exp. Immunol.*, 2020.
- [58] Liu, Y., Yin, H., Zhao, M., and Lu, Q. TLR2 and TLR4 in autoimmune diseases: a comprehensive review. *Clin. Rev. Allergy Immunol.*, 47(2):136–147, 2014.



- [59] Duffy, L. and O'Reilly, S. C. Toll-like receptors in the pathogenesis of autoimmune diseases: recent and emerging translational developments. *ImmunoTargets Ther.*, 5:69–80, 2016.
- [60] Basith, S., Manavalan, B., Yoo, T. H., Kim, S. G., and Choi, S. J. A. o. P. R. Roles of toll-like receptors in cancer: A double-edged sword for defense and offense. *Arch. Pharmacol Res.*, 35(8):1297–1316, 2012.
- [61] Pradere, J. P., Dapito, D. H., and Schwabe, R. F. The Yin and Yang of Toll-like receptors in cancer. *Oncogene*, 33(27):3485–3495, 2014.
- [62] Wang, R. F., Miyahara, Y., and Wang, H. Y. Toll-like receptors and immune regulation: implications for cancer therapy. *Oncogene*, 27(2):181–189, 2008.
- [63] Cheng, B., Yuan, W.-E., Su, J., Liu, Y., and Chen, J. Recent advances in small molecule based cancer immunotherapy. *Eur. J. Med. Chem.*, 157:582–598, 2018.
- [64] Lowy, D. R., Solomon, D., Hildesheim, A., Schiller, J. T., and Schiffman, M. Human papillomavirus infection and the primary and secondary prevention of cervical cancer. *Cancer*, 113(S7):1980–1993, 2008.
- [65] Vacchelli, E., Galluzzi, L., Eggermont, A., Fridman, W. H., Galon, J., Sautès-Fridman, C., Tartour, E., Zitvogel, L., and Kroemer, G. Trial watch: FDA-approved Toll-like receptor agonists for cancer therapy. *OncoImmunology*, 1(6): 894–907, 2012.
- [66] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT01294293>. Identifier NCT01294293, TLR8 Agonist VTX-2337 and Pegylated Liposomal Doxorubicin Hydrochloride or Paclitaxel in Treating Patients With Recurrent or Persistent Ovarian Epithelial, Fallopian Tube, or Peritoneal Cavity Cancer; 2011 Feb 11 [cited 2020 Oct 15].
- [67] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT03906526>. Identifier NCT03906526, A Study to Evaluate Immune Biomarker Modulation in Response

to VTX-2337 in Combination With an Anti- PD-1 Inhibitor in Head and Neck Cancer; 2019 Apr 8 [cited 2020 Oct 15].

- [68] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT02035657>. Identifier NCT02035657, A Proof-of-Concept Trial of GLA-SE in Patients With Merkel Cell Carcinoma; 2014 Jan 14 [cited 2020 Oct 15].
- [69] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT02320305>. Identifier NCT02320305, MART-1 Antigen With or Without TLR4 Agonist GLA-SE in Treating Patients With Stage II-IV Melanoma That Has Been Removed by Surgery; 2014 Dec 19 [cited 2020 Oct 15].
- [70] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT02092909>. Identifier NCT02092909, Phase 1/2 Dose Escalation Study in Patients With Relapsed or Refractory Waldenstrom's Macroglobulinemia (8400-401); 2014 Mar 20 [cited 2020 Oct 15].
- [71] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT04126876>. Identifier NCT04126876, A Randomized Controlled Phase II Trial With Intradermal IMO-2125 in Pathological Tumor Stage (p) T3-4 cN0M0 Melanoma (INTRIM); 2019 Oct 15 [cited 2020 Oct 20].
- [72] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT02927964>. Identifier NCT02927964, TLR9 Agonist SD-101, Ibrutinib, and Radiation Therapy in Treating Patients With Relapsed or Refractory Grade 1-3A Follicular Lymphoma; 2016 Oct 7 [cited 2020 Oct 20].
- [73] Vacchelli, E., Eggermont, A., Sautès-Fridman, C., Galon, J., Zitvogel, L., Kroemer, G., and Galluzzi, L. Trial watch: Oncolytic viruses for cancer therapy. *OncoImmunology*, 2(6):e24612, 2013.

- [74] Smith, M., García-Martínez, E., Pitter, M. R., Fucikova, J., Spisek, R., Zitvogel, L., Kroemer, G., and Galluzzi, L. Trial watch: Toll-like receptor agonists in cancer immunotherapy. *OncoImmunology*, 7(12):e1526250, 2018.
- [75] Kubo, T., Morita, H., Sugita, K., and Akdis, C. A. *Chapter 1 - Introduction to Mechanisms of Allergic Diseases*, pages 1–27. Elsevier, 2017. ISBN 978-0-323-37579-5. doi: 10.1016/B978-0-323-37579-5.00001-5.
- [76] Ebbo, M., Crinier, A., Vely, F., and Vivier, E. Innate lymphoid cells: major players in inflammatory diseases. *Nat. Rev. Immunol.*, 17(11):665, 2017.
- [77] Barnes, P. J. Targeting cytokines to treat asthma and chronic obstructive pulmonary disease. *Nat. Rev. Immunol.*, 18(7):454–466, 2018.
- [78] Lambrecht, B. N., Hammad, H., and Fahy, J. V. The cytokines of asthma. *Immunity*, 50(4):975–991, 2019.
- [79] Bousquet, J., Khaltaev, N., Cruz, A. A., Denburg, J., Fokkens, W., Togias, A., Zuberbier, T., Baena-Cagnani, C. E., Canonica, G., Van Weel, C., et al. Allergic rhinitis and its impact on asthma (ARIA) 2008. *Allergy*, 63:8–160, 2008.
- [80] Leynaert, B., Neukirch, C., Liard, R., Bousquet, J., and Neukirch, F. Quality of life in allergic rhinitis and asthma: a population-based study of young adults. *Am. J. Respir. Crit. Care Med.*, 162(4):1391–1396, 2000.
- [81] Horak, F. VTX-1463, a novel TLR8 agonist for the treatment of allergic rhinitis. *Expert Opin. Invest. Drugs*, 20(7):981–986, 2011.
- [82] Aryan, Z., Holgate, S. T., Radzioch, D., and Rezaei, N. A new era of targeting the ancient gatekeepers of the immune system: Toll-like agonists in the treatment of allergic rhinitis and asthma. *Int. Arch. Allergy Immunol.*, 164(1):46–63, 2014.
- [83] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT02833974>. Identifier NCT02833974, Effect of the GSK2245035 on the Allergen-induced Asthmatic Response; 2016 Jul 15 [cited 2020 Oct 20].

- [84] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT03707678>. Identifier NCT03707678, A Safety and Efficacy Study of Intranasal GSK2245035 in Adults With Allergic Asthma; 2018 Oct 16 [cited 2020 Oct 20].
- [85] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT01480271>. Identifier NCT01480271, An Investigation of the Safety, Tolerability, Pharmacokinetics and Pharmacodynamics of GSK2245035 in Healthy Volunteers and Allergic Rhinitics; 2011 Nov 28 [cited 2020 Oct 20].
- [86] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT01607372>. Identifier NCT01607372, A Study to Investigate the Safety and Pharmacodynamics of Repeat Intranasal Administration of the TLR7 Agonist GSK2245035 in Subjects With Respiratory Allergies; 2012 May 30 [cited 2020 Oct 20].
- [87] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT01788813>. Identifier NCT01788813, To Investigate the Safety, Pharmacodynamics and Effect on Allergic Reactivity of the Toll-like Receptor 7 (TLR7) Agonist GSK2245035 in Subjects With Respiratory Allergies; 2013 Feb 11 [cited 2020 Oct 20].
- [88] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT02087644>. Identifier NCT02087644, CYT003 in Patients With Mild to Moderate Allergic Asthma Not Sufficiently Controlled on Inhaled Glucocorticosteroids; 2014 Mar 14 [cited 2020 Oct 20].
- [89] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. . URL <https://clinicaltrials.gov/ct2/show/NCT01673672>. Identifier NCT01673672, CYT003-QbG10, a TLR9-agonist, for Treatment of Uncontrolled Moderate to Severe Allergic Asthma; 2012 Aug 28 [cited 2020 Oct 20].
- [90] Archer, K. A. and Roy, C. R. MyD88-dependent responses involving toll-like

- receptor 2 are important for protection and clearance of legionella pneumophila in a mouse model of Legionnaires' disease. *Infect. Immun.*, 74(6):3325–3333, 2006.
- [91] Mifsud, E., Tan, A., and Jackson, D. TLR agonists as modulators of the innate immune response and their potential as agents against infectious disease. *Front. Immunol.*, 5(79):79, 2014.
- [92] Hedayat, M., Netea, M. G., and Rezaei, N. Targeting of Toll-like receptors: a decade of progress in combating infectious diseases. *Lancet Infect. Dis.*, 11(9):702–712, 2011.
- [93] Velavan, T. P. and Meyer, C. G. The COVID-19 epidemic. *Trop. Med. Int. Health*, 25(3):278–280, 2020.
- [94] Muthuraman, S., Al Haziati, M., et al. COVID-19 impact on health, social & economy. *Int. J. Nutr., Pharmacol., Neurol. Dis.*, 10(3):159–160, 2020.
- [95] Fernandes, N. Economic effects of coronavirus outbreak (COVID-19) on the world economy, 2020. URL <http://dx.doi.org/10.2139/ssrn.3557504>. March 22, 2020. IESE Business School Working Paper No. WP-1240-E.
- [96] Li, G., Fan, Y., Lai, Y., Han, T., Li, Z., Zhou, P., Pan, P., Wang, W., Hu, D., Liu, X., et al. Coronavirus infections and immune responses. *J. Med. Virol.*, 92(4):424–432, 2020.
- [97] Patra, M. C. and Choi, S. Recent progress in the development of Toll-like receptor (TLR) antagonists. *Expert Opin. Ther. Pat.*, 26(6):719–730, 2016.
- [98] Keely, S., Talley, N. J., and Hansbro, P. M. Pulmonary-intestinal cross-talk in mucosal inflammatory disease. *Mucosal Immunol.*, 5(1):7–18, 2012.
- [99] Sidletskaya, K., Vitkina, T., and Denisenko, Y. The role of Toll-like receptors 2 and 4 in the pathogenesis of chronic obstructive pulmonary disease. *Int. J. Chronic Obstruct. Pulm. Dis.*, 15:1481, 2020.
- [100] Shi, H., Hua, X., Kong, D., Stein, D., and Hua, F. Role of Toll-like receptor mediated signaling in traumatic brain injury. *Neuropharmacology*, 145:259–267, 2019.

- [101] Wong, S. K., Chin, K.-Y., and Ima-Nirwana, S. Toll-like receptor as a molecular link between metabolic syndrome and inflammation: a review. *Curr. Drug Targets*, 20(12):1264–1280, 2019.
- [102] Mills, K. H. G. TLR-dependent T cell activation in autoimmunity. *Nat. Rev. Immunol.*, 11(12):807–822, 2011.
- [103] ClinicalTrials.gov [Internet]. Bethesda (MD): National Library of Medicine (US). 2000-. URL <https://clinicaltrials.gov/ct2/show/NCT01899729>. Identifier NCT01899729, A 12-week Dose-Ranging Trial in Patients With Moderate to Severe Plaque Psoriasis (8400-201); 2012 Aug 28 [cited 2020 Oct 20].
- [104] Kain, S. R. and Ganguly, S. Overview of genetic reporter systems. *Curr. Protoc. Mol. Biol.*, 68(1):9.6.1–9.6.12, 2004.
- [105] Giulietti, A., Overbergh, L., Valckx, D., Decallonne, B., Bouillon, R., and Mathieu, C. An overview of real-time quantitative PCR: applications to quantify cytokine gene expression. *Methods*, 25(4):386–401, 2001.
- [106] Leng, S. X., McElhaney, J. E., Walston, J. D., Xie, D., Fedarko, N. S., and Kuchel, G. A. Elisa and multiplex technologies for cytokine measurement in inflammation and aging research. *J. Gerontol. A Biol. Sci. Med. Sci.*, 63(8):879–884, 2008.
- [107] Nolan, T., Hands, R. E., and Bustin, S. A. Quantification of mRNA using real-time RT-PCR. *Nat. Protoc.*, 1(3):1559–1582, 2006.
- [108] Wong, M. L. and Medrano, J. F. Real-time PCR for mRNA quantitation. *Biotechniques*, 39(1):75–85, 2005.
- [109] Tian, B., Nowak, D. E., Jamaluddin, M., Wang, S., and Brasier, A. R. Identification of direct genomic targets downstream of the nuclear factor- $\kappa$ b transcription factor mediating tumor necrosis factor signaling. *J. Biol. Chem.*, 280(17):17435–17448, 2005.
- [110] Zhang, Y., Ma, F., Tang, B., and Zhang, C.-y. Recent advances in transcription factor assays in vitro. *Chem. Commun.*, 52(26):4739–4748, 2016.

- [111] Adan, A., Kiraz, Y., and Baran, Y. Cell proliferation and cytotoxicity assays. *Curr. Pharm. Biotechnol.*, 17(14):1213–1221, 2016.
- [112] Koch, K.-W. *Surface Plasmon Resonance*, pages 1832–1835. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-29623-2. doi: 10.1007/3-540-29623-9\_4880. URL [https://doi.org/10.1007/3-540-29623-9\\_4880](https://doi.org/10.1007/3-540-29623-9_4880).
- [113] Reinisch, T. and Hinz, H.-J. *Isothermal Titration Calorimetry*, pages 919–925. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-29623-2. doi: 10.1007/3-540-29623-9\_5770. URL [https://doi.org/10.1007/3-540-29623-9\\_5770](https://doi.org/10.1007/3-540-29623-9_5770).
- [114] Lakowicz, J. R. *Principles of fluorescence spectroscopy*. Springer science & business media, 2013.
- [115] Wienken, C. J., Baaske, P., Rothbauer, U., Braun, D., and Duhr, S. Protein-binding assays in biological liquids using microscale thermophoresis. *Nat. Commun.*, 1:100, 2010.
- [116] Ganten, D., Ruckpaul, K., Birchmeier, W., Epplen, J. T., Genser, K., Gossen, M., Kersten, B., Lehrach, H., Oschkinat, H., Ruiz, P., Schmieder, P., Wanker, E., and Nolte, C., editors. *Primary Cells*, pages 1459–1459. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-29623-2. doi: 10.1007/3-540-29623-9\_8327. URL [https://doi.org/10.1007/3-540-29623-9\\_8327](https://doi.org/10.1007/3-540-29623-9_8327).
- [117] Fundamental techniques in cell culture laboratory handbook, ECACC Laboratory Handbook 4th Edition, 2011. URL <https://www.sigmaaldrich.com/technical-documents/protocols/biology/cell-types-culture.html>. Accessed on 09 August 2020.
- [118] Ramakrishnan, G. *Drug Discovery*, pages 3–28. Springer Netherlands, Dordrecht, 2017. ISBN 978-94-024-1045-7. doi: 10.1007/978-94-024-1045-7\_1.
- [119] Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., and Zhao, S. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discovery*, 18(6): 463–477, 2019.

- [120] Yang, X., Wang, Y., Byrne, R., Schneider, G., and Yang, S. Concepts of artificial intelligence for computer-assisted drug discovery. *Chem. Rev. (Washington, DC, U. S.)*, 119(18):10520–10594, 2019.
- [121] Boulesteix, A.-L., Janitza, S., Kruppa, J., and König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, 2(6):493–507, 2012.
- [122] Biau, G. and Scornet, E. A random forest guided tour. *Test*, 25(2):197–227, 2016.
- [123] Khanna, V., Li, L., Fung, J., Ranganathan, S., and Petrovsky, N. Prediction of novel mouse TLR9 agonists using a random forest approach. *BMC Mol. Cell Biol.*, 20(2):56, 2019.
- [124] Samuel, A. L. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3(3):210–229, 1959.
- [125] Harnad, S. The annotation game: On turing (1950) on computing, machinery, and intelligence (published version bowdlerized). In Epstein, R., Roberts, G., and Beber, G., editors, *Parsing the Turing Test: Philosophical and Methodological Issues in the Quest for the Thinking Computer*, pages 23–66. Springer, 2008. URL <https://eprints.soton.ac.uk/262954/>. Chapter: 3 Commentary On: Turing, A.M. (1950) Computing Machinery and Intelligence. *Mind* 49 433-460 Address: Amsterdam.
- [126] Géron, A. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O’Reilly Media, 2019.
- [127] Crone, S. F., Lessmann, S., and Stahlbock, R. The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing. *Eur. J. Oper. Res.*, 173(3):781–800, 2006.
- [128] Dasu, T. and Johnson, T. *Exploratory data mining and data cleaning*, volume 479. John Wiley & Sons, 2003.



- [129] Mohd Nawi, N., Atomia, W. H., and Rehman, M. Z. The effect of data pre-processing on optimized training of artificial neural networks. In: 4th International Conference on Electrical Engineering and Informatics (ICEEI 2013, 24-25 June), Universiti Kebangsaan Malaysia, 2013.
- [130] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.*, 12:2825–2830, 2011.
- [131] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pages 108–122, 2013.
- [132] Feurer, M. and Hutter, F. *Hyperparameter Optimization*, pages 3–33. Springer International Publishing, 2019.
- [133] Bergstra, J. and Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.*, 13(1):281–305, 2012.
- [134] Sokolova, M. and Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manage.*, 45(4):427–437, 2009.
- [135] Sammut, C. and Webb, G. I. *Encyclopedia of machine learning*. Springer Science & Business Media, 2011.
- [136] Sammut, C. and Webb, G. I., editors. *Accuracy*, pages 9–10. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8\_3. URL [https://doi.org/10.1007/978-0-387-30164-8\\_3](https://doi.org/10.1007/978-0-387-30164-8_3).
- [137] Chawla, N. V. Data mining for imbalanced datasets: An overview. In *Data mining and knowledge discovery handbook*, pages 875–886. Springer, 2009.
- [138] Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8): 861–874, 2006.

- [139] Myerson, J., Green, L., and Warusawitharana, M. Area under the curve as a measure of discounting. *J. Exp. Anal. Behav.*, 76(2):235–243, 2001.
- [140] Bradley, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognit.*, 30(7):1145–1159, 1997.
- [141] Chicco, D. and Jurman, G. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.
- [142] Instant JChem was used for structure database management, search and prediction. Instant JChem version 18.1.0, 2018. ChemAxon (<http://www.chemaxon.com>).
- [143] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.
- [144] Weininger, D., Weininger, A., and Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.*, 29(2):97–101, 1989.
- [145] Daylight SMILES theory, 2019. URL <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html>.
- [146] Molecular operating environment (MOE), 20190102. Chemical Computing Group Inc., 2019. <https://www.chemcomp.com/>.
- [147] O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open Babel: An open chemical toolbox. *J. Cheminf.*, 3(1):33, 2011.
- [148] The Open Babel Package, version 2.4.1, Oct 2016. URL <http://openbabel.org>.
- [149] Nguyen, K. T., Blum, L. C., van Deursen, R., and Reymond, J.-L. Classification of organic molecules by molecular quantum numbers. *ChemMedChem*, 4(11):1803–1805, 2009.
- [150] Landrum, G., Tosco, P., Kelley, B., sriniker, gedec, NadineSchneider, Vianello, R., Dalke, A., Cole, B., AlexanderSavelyev, Turk, S., Ric, Swain, M., Vaucher, A.,

- N, D., Wójcikowski, M., Pahl, A., JP, Berenger, F., strets123, JLVarjo, O'Boyle, N., Cosgrove, D., Fuller, P., Jensen, J. H., Sforza, G., DoliathGavid, Leswing, K., Nowotka, M., and van Santen, J. rdkit/rdkit: 2019\_03\_3 (q1 2019) release, June 2019. URL <https://doi.org/10.5281/zenodo.3256322>.
- [151] MySQL Workbench 8.0 CE, version 8.0.12, 2018. URL <https://www.mysql.com/products/workbench/>.
- [152] Bootstrap, an open source toolkit for developing with HTML, CSS, and JS, 2019. URL <https://getbootstrap.com/>.
- [153] Marvin JS provides quick and convenient ways to draw and modify standard and advanced chemical structures., 2019. Marvin JS (<https://chemaxon.com/products/marvin-js>).
- [154] Probst, D. and Reymond, J.-L. Smilesdrawer: parsing and drawing SMILES-encoded molecular structures using client-side javascript. *J. Chem. Inf. Model.*, 58(1):1–7, 2018.
- [155] Python Software Foundation. Python Language Reference, version 3.6, 2019. URL <http://www.python.org>. Available at <http://www.python.org>.
- [156] Wold, S., Esbensen, K., and Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.*, 2(1-3):37–52, 1987.
- [157] Martin, Y. C., Kofron, J. L., and Traphagen, L. M. Do structurally similar molecules have similar biological activity? *Journal of medicinal chemistry*, 45(19):4350–4358, 2002.
- [158] Wassermann, A. M. and Bajorath, J. Chemical substitutions that introduce activity cliffs across different compound classes and biological targets. *J. Chem. Inf. Model.*, 50(7):1248–1256, 2010.
- [159] Leach, A. G., Jones, H. D., Cosgrove, D. A., Kenny, P. W., Ruston, L., MacFaul, P., Wood, J. M., Colclough, N., and Law, B. Matched molecular pairs as a guide in the optimization of pharmaceutical properties; a study of aqueous solubility, plasma protein binding and oral exposure. *J. Med. Chem.*, 49(23):6672–6682, 2006.

- [160] Griffen, E., Leach, A. G., Robb, G. R., and Warner, D. J. Matched molecular pairs as a medicinal chemistry tool: miniperspective. *J. Med. Chem.*, 54(22): 7739–7750, 2011.
- [161] Hussain, J. and Rea, C. Computationally efficient algorithm to identify matched molecular pairs (MMPs) in large data sets. *J. Chem. Inf. Model.*, 50(3):339–348, 2010.
- [162] Dalke, A., Hert, J., and Kramer, C. mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *J. Chem. Inf. Model.*, 58(5): 902–910, 2018.
- [163] Stewart, K. D., Shiroda, M., and James, C. A. Drug Guru: a computer software program for drug design using medicinal chemistry rules. *Bioorg. Med. Chem.*, 14(20):7011–7022, 2006.
- [164] Warner, D. J., Griffen, E. J., and St-Gallay, S. A. WizePairZ: A novel algorithm to identify, encode, and exploit matched molecular pairs with unspecified cores in medicinal chemistry. *J. Chem. Inf. Model.*, 50(8):1350–1357, 2010.
- [165] Weber, J., Achenbach, J., Moser, D., and Proschak, E. VAMMPIRE: a matched molecular pairs database for structure-based drug design and optimization. *J. Med. Chem.*, 56(12):5203–5207, 2013.
- [166] Tyrchan, C. and Evertsson, E. Matched molecular pair analysis in short: algorithms, applications and limitations. *Comput. Struct. Biotechnol. J.*, 15:86–90, 2016.
- [167] Wassermann, A. M., Dimova, D., Iyer, P., and Bajorath, J. Advances in computational medicinal chemistry: matched molecular pair analysis. *Drug Dev. Res.*, 73(8):518–527, 2012.
- [168] O’Boyle, N. M., Boström, J., Sayle, R. A., and Gill, A. Using matched molecular series as a predictive tool to optimize biological activity. *J. Med. Chem.*, 57(6): 2704–2713, 2014.
- [169] Wawer, M. and Bajorath, J. Local structural changes, global data views: graphical substructure-activity relationship trailing. *J. Med. Chem.*, 54(8):2944–2951, 2011.

- [170] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Sieb, C., Thiel, K., and Wiswedel, B. KNIME: The Konstanz Information Miner. In *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007. ISBN 978-3-540-78239-1.
- [171] Berthold, M. R., Cebron, N., Dill, F., Gabriel, T. R., Kötter, T., Meinl, T., Ohl, P., Thiel, K., and Wiswedel, B. KNIME-the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor*, 11(1):26–31, 2009.
- [172] Raymond, J. W., Watson, I. A., and Mahoui, A. Rationalizing lead optimization by associating quantitative relevance with molecular structure modification. *J. Chem. Inf. Model.*, 49(8):1952–1962, 2009.
- [173] Sheridan, R. P., Hunt, P., and Culberson, J. C. Molecular transformations as a way of finding and exploiting consistent local QSAR. *J. Chem. Inf. Model.*, 46(1):180–192, 2006.
- [174] Wagener, M. and Lommerse, J. P. M. The quest for bioisosteric replacements. *J. Chem. Inf. Model.*, 46(2):677–685, 2006.
- [175] Papadatos, G., Alkarouri, M., Gillet, V. J., Willett, P., Kadirkamanathan, V., Luscombe, C. N., Bravi, G., Richmond, N. J., Pickett, S. D., Hussain, J., Pritchard, J. M., Cooper, A. W. J., and Macdonald, S. J. F. Lead optimization using matched molecular pairs: inclusion of contextual information for enhanced prediction of hERG inhibition, solubility, and lipophilicity. *J. Chem. Inf. Model.*, 50(10):1872–1886, 2010.
- [176] Jones, G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.*, 267(3):727–748, 1997.
- [177] Jin, M. S., Kim, S. E., Heo, J. Y., Lee, M. E., Kim, H. M., Paik, S.-G., Lee, H., and Lee, J.-O. Crystal structure of the TLR1-TLR2 heterodimer induced by binding of a tri-acylated lipopeptide. *Cell*, 130(6):1071–1082, 2007.
- [178] Tanji, H., Ohto, U., Shibata, T., Miyake, K., and Shimizu, T. Structural

- reorganization of the Toll-like receptor 8 dimer induced by agonistic ligands. *Science*, 339(6126):1426–1429, 2013.
- [179] Zhang, S., Hu, Z., Tanji, H., Jiang, S., Das, N., Li, J., Sakaniwa, K., Jin, J., Bian, Y., Ohto, U., et al. Small-molecule inhibition of TLR8 through stabilization of its resting state. *Nat. Chem. Biol.*, 14(1):58–64, 2018.
- [180] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *J. Comput. Chem.*, 17(5-6):490–519, 1996.
- [181] Wolber, G. and Langer, T. LigandScout: 3-D pharmacophores derived from protein-bound ligands and their use as virtual screening filters. *J. Chem. Inf. Model.*, 45(1):160–169, 2005.
- [182] Wolber, G., Dornhofer, A. A., and Langer, T. Efficient overlay of small organic molecules using 3D pharmacophores. *J. Comput.-Aided Mol. Des.*, 20(12):773–788, 2006.
- [183] Hall, L. H. and Kier, L. B. *The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling*, pages 367–422. John Wiley & Sons, Ltd, 1991. ISBN 9780470125793. doi: <https://doi.org/10.1002/9780470125793.ch9>. URL <https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470125793.ch9>.
- [184] Kier, L., LB, K., and LH, H. The nature of structure-activity relationships and their relation to molecular connectivity. *Eur. J. Med. Chem.*, 12(4):307–312, 1977.
- [185] Sanders, H. and Saxe, J. Garbage in, garbage out: How purportedly great ML models can be screwed up by bad data. In *Proceedings of Blackhat 2017. July 26-July 27, 2017, Las Vegas, USA*, 2017.
- [186] Buuren, S. v. and Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.*, pages 1–68, 2010.
- [187] Little, R. J. and Rubin, D. B. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

- [188] Willett, P., Barnard, J. M., and Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.*, 38(6):983–996, 1998.
- [189] Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L. J., Cibrián-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magariños, M. P., Overington, J. P., Papadatos, G., Smit, I., and Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res.*, 45(D1):D945–D954, 2017.
- [190] Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.*, 47(D1):D930–D940, 2019.
- [191] Gilson, M. K., Liu, T., Baitaluk, M., Nicola, G., Hwang, L., and Chong, J. BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.*, 44(D1):D1045–D1053, 2016.
- [192] Lata, S. and Raghava, G. P. S. PRRDB: a comprehensive database of pattern-recognition receptors and their ligands. *BMC Genomics*, 9(1):180, 2008.
- [193] Kaur, D., Patiyal, S., Sharma, N., Usmani, S. S., and Raghava, G. P. S. PRRDB 2.0: a comprehensive database of pattern-recognition receptors and their ligands. *Database*, 2019, 2019.
- [194] Chatterjee, D., Kaur, G., Muradia, S., Singh, B., and Agrewala, J. N. Imm-torLig\_DB: repertoire of virtually screened small molecules against immune receptors to bolster host immunity. *Sci. Rep.*, 9(3092):1–13, 2019.
- [195] Alexander, S. P., Fabbro, D., Kelly, E., Mathie, A., Peters, J. A., Veale, E. L., Armstrong, J. F., Faccenda, E., Harding, S. D., Pawson, A. J., et al. The concise guide to pharmacology 2019/20: catalytic receptors. *British journal of pharmacology*, 176:S247–S296, 2019.
- [196] Davies, M., Nowotka, M., Papadatos, G., Dedman, N., Gaulton, A., Atkinson, F.,

- Bellis, L., and Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.*, 43(W1):W612–W620, 2015.
- [197] van De Waterbeemd, H., Camenisch, G., Folkers, G., and Raevsky, O. A. Estimation of Caco-2 cell permeability using calculated molecular descriptors. *Quant. Struct.-Act. Relat.*, 15(6):480–490, 1996.
- [198] Ertl, P., Rohde, B., and Selzer, P. Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties. *J. Med. Chem.*, 43(20):3714–3717, 2000.
- [199] Fernandes, J. and Gattass, C. R. Topological polar surface area defines substrate transport by multidrug resistance associated protein 1 (MRP1/ABCC1). *J. Med. Chem.*, 52(4):1214–1218, 2009.
- [200] Veber, D. F., Johnson, S. R., Cheng, H.-Y., Smith, B. R., Ward, K. W., and Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.*, 45(12):2615–2623, 2002.
- [201] Labute, P. A widely applicable set of descriptors. *J. Mol. Graphics Modell.*, 18(4):464–477, 2000.
- [202] Wishart, D. S., Feunang, Y. D., Guo, A. C., Lo, E. J., Marcu, A., Grant, J. R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., et al. DrugBank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Res.*, 46(D1):D1074–D1082, 2018.
- [203] Lipinski, C. A., Lombardo, F., Dominy, B. W., and Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.*, 23(1):3–25, 1997.
- [204] Tinworth, C. P. and Young, R. J. Facts, patterns, and principles in drug discovery: Appraising the rule of 5 with measured physicochemical data. *J. Med. Chem.*, 63(18):10091–10108, 2020.
- [205] Murgueitio, M. S., Henneke, P., Glossmann, H., Santos-Sierra, S., and Wolber, G. Prospective virtual screening in a sparse data scenario: design of small-molecule TLR2 antagonists. *ChemMedChem*, 9(4):813–822, 2014.



- [206] Zhong, Z., Liu, L.-J., Dong, Z.-Q., Lu, L., Wang, M., Leung, C.-H., Ma, D.-L., and Wang, Y. Structure-based discovery of an immunomodulatory inhibitor of TLR1–TLR2 heterodimerization from a natural product-like database. *Chem. Commun.*, 51(56):11178–11181, 2015.
- [207] Durai, P., Shin, H. J., Achek, A., Kwon, H. K., Govindaraj, R. G., Panneerselvam, S., Yesudhas, D., Choi, J., No, K. T., and Choi, S. Toll-like receptor 2 antagonists identified through virtual screening and experimental validation. *FEBS J.*, 284(14):2264–2283, 2017.
- [208] Ozinsky, A., Underhill, D. M., Fontenot, J. D., Hajjar, A. M., Smith, K. D., Wilson, C. B., Schroeder, L., and Aderem, A. The repertoire for pattern recognition of pathogens by the innate immune system is defined by cooperation between toll-like receptors. *PNAS*, 97(25):13766–13771, 2000.
- [209] Alpaydin, E. *Introduction to machine learning*. MIT press, 2020.



# List of Figures

1.1	MyD88-dependent and MyD88-independent signaling pathways for TLRs.	3
1.2	Drug discovery process.	11
1.3	A labeled training set for supervised learning.	12
3.1	Compound selection and labeling criteria.	22
3.2	KNIME workflow applied for searching activity cliffs between MMPs.	25
3.3	Machine Learning workflow applied for a specific prediction task.	26
4.1	Database schema for <i>TollDB</i> .	32
4.2	Screenshot of example query for Simple Search.	34
4.3	Screenshot of example query for Advanced Search.	35
4.4	Screenshot of example query for Structure Search.	36
4.5	Number of active molecules targeting each TLR subtype.	38
4.6	Assay condition number count for the same assay readout.	39
4.7	Composition of <i>TollDB</i> with comparison to <i>ChEMBL</i> .	40
4.8	Box plot for MW and log <i>P</i> .	41
4.9	Box plot of MW and log <i>P</i> for TLR subtypes	42
4.10	Compound count by chiral atoms for each TLR subtypes.	43
4.11	Distribution of TPSA for TLR subtypes.	44
4.12	Ring number distribution for ligands of different TLR subtypes.	45
4.13	Distribution of H-bond acceptors for TLR subtypes.	45
4.14	Distribution of H-bond donors for TLR subtypes.	46
4.15	Distribution of rotatable bonds for TLR subtypes.	47
4.16	Distribution of ASA for TLR subtypes.	47
4.17	Distribution of VSA for TLR subtypes.	48
4.18	Chemical space of <i>TollDB</i> vs <i>DrugBank</i> .	49

4.19	Activity cliff examples for TLR2 agonists. . . . .	53
4.20	2D representation of ligand-protein interactions for TLR2 agonists. . .	55
4.21	Predicted binding poses for TLR2 agonists. . . . .	55
4.22	Activity cliff examples for TLR8 agonists. . . . .	56
4.23	Predicted binding pose for TLR8 agonists. . . . .	57
4.24	Predicted binding pose for TLR8 agonists. . . . .	58
4.25	Model scores for TLR2 agonism with random search. . . . .	61
4.26	Model scores for TLR2 agonism with grid search. . . . .	62
4.27	ROC curve for TLR2 agonism prediction models. . . . .	63
4.28	Feature importance for TLR2 agonism with random search. . . . .	65
4.29	Feature importance for TLR2 agonism with grid search. . . . .	66
4.30	Scatter plots for the TLR2 agonism data set. . . . .	67

# List of Tables

1.1	TLR family as potential drug targets . . . . .	5
1.2	Confusion matrix . . . . .	14
3.1	Data collection and curation . . . . .	20
4.1	Loadings of the first two components resulting from PCA analysis. . . . .	50
4.2	Examples of activity cliffs between MMPs. . . . .	52
6.1	MMPs for <i>TollDB</i> . . . . .	80
6.2	Machine learning study for <i>TollDB</i> with random search . . . . .	89
6.3	Machine learning study for <i>TollDB</i> with grid search . . . . .	98



# Publications

## Peer-reviewed articles

1. Schaller, D., Šribar, D., Noonan, T., Deng, L., et al. Next generation 3D pharmacophore modeling. *Wiley Interdiscip. Rev. Comput. Mol. Sci.*, e1468, 2020.
2. Omieczynski, C., Nguyen, T. N., Šribar, D., Deng, L., et al. BiasDB: A Comprehensive Database for Biased GPCR Ligands, first published August 24, 2019. doi: 10.1101/742643. [In preparation]
3. Deng, L., et al. TollDB: A comprehensive database for Toll-like receptor modulators. [In preparation]

## Poster presentations

1. Deng, L., et al. Rational design and optimization of small molecule Toll-like receptor modulators. Vienna Summer School on Drug Design (*Sep 2017*), Department of Pharmaceutical Chemistry, University of Vienna, 1090 Vienna, Austria.