ARTIFICIAL LEARNING-BASED ANALYSIS OF MOLECULAR, CLINICAL TRIALS AND PATENT DATA FOR IMPROVED DRUG DEVELOPMENT

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF INFORMATICS OF
THE MIDDLE EAST TECHNICAL UNIVERSITY
BY

FULYA ÇIRAY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE
OF DOCTOR OF PHILOSOPHY
IN
MEDICAL INFORMATICS

AUGUST 2022

Approval of the thesis:

ARTIFICIAL LEARNING-BASED ANALYSIS OF MOLECULAR, CLINICAL TRIALS AND PATENT DATA FOR IMPROVED DRUG DEVELOPMENT

Submitted by Fulya Çıray in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Health Informatics Department, Middle East Technical University** by,

Prof. Dr. Deniz Zeyrek Bozşahin
Dean, **Graduate School of Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Head of Department, **Health Informatics** _____

Assoc. Prof. Dr. Yeşim Aydın Son
Supervisor, **Health Informatics, METU** _____

Assoc. Prof. Dr. Tunca Doğan
Co-Supervisor, **Computer Engineering Dept., Hacettepe University** _____

Examining Committee Members:

Prof. Dr. Erden Banoğlu
Pharmaceutical Chemistry, Gazi University _____

Assoc. Prof. Dr. Yeşim Aydın Son
Health Informatics, METU _____

Asst. Prof. Dr. Aybar Can Acar
Health Informatics, METU _____

Asst. Prof. Dr. Burçak Otlu Sarıtaş
Health Informatics, METU _____

Asst. Prof. Dr. İdil Yet
Department of Bioinformatics, Hacettepe University _____

Date: *31.08.2022*

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last name : Fulya ÇIRAY

Signature : _____

ABSTRACT

ARTIFICIAL LEARNING-BASED ANALYSIS OF MOLECULAR, CLINICAL TRIALS AND PATENT DATA FOR IMPROVED DRUG DEVELOPMENT

Çıray, Fulya

Ph.D., Department of Health Informatics

Supervisor: Assoc. Prof. Dr. Yeşim Aydın Son
Co-Supervisor: Assoc. Prof. Dr. Tunca Doğan

August 2022, 146 pages

Drug development is a costly process, especially in terms of the required time and money. Many promising drug candidates are eliminated at late development stages, e.g., phase II or III of clinical trials, due to insufficient efficacy or unexpected adverse health related affects. Lately, pharmaceutical companies are evaluating computational approaches, to increase the efficiency of this process. In this thesis study, we investigated the computational prediction of the approval of drug candidate compounds by regulatory bodies (i.e., approved for an official use to treat the indicated disease) while the trial process is still continuing, using relevant information from previous discovery and development stages and machine learning. As a preliminary analysis, we examined drug substructures to observe whether the presence of specific molecular structures in drug candidates lead to undesirable outcomes (i.e., unapproved). In the main part of the study, we employed a wider and more heterogeneous set of features including molecular and physicochemical properties of drugs, together with clinical trial and patent related features, to represent each drug-indication pair as a heterogeneous numerical vector. Following data gathering, manual curation and imputation procedures, our finalized feature vectors are processed by random forest (RF) classifiers to train independent drug approval prediction models for 14 different disease groups. We achieved high prediction scores in our cross validation-based performance evaluation, varying in ranges of; accuracy: 0.67-0.81, precision: 0.77-0.82, recall: 0.77-0.96, F1-score: 0.77-0.88 and MCC: 0.45-0.62. Furthermore, by conducting a temporal analysis, we showed that our method is also capable of producing successful results in a prospective manner. We also carried out a performance comparison against a baseline model and a state-of-the-art method from literature, the results of which indicated both robustness and the generalization capability of our approach. Additionally, we identified the most important features for accurately predicting drug approvals, which heavily includes clinical trial and patent related features. Within a use-case study, we showed that our method can successfully

predict regulatorily approved (phase IV) drugs that are later withdrawn from the market due to severe side effects. Finally, we used pre-trained models to predict the approval of drug candidates that are currently in clinical trial phases I/II/III and presented prediction results. We hope that the results of our study and the computational tool we presented will contribute to the literature in terms of evaluating and improving the drug development process. All of the datasets, source code, results and pre-trained models of this study are freely available at https://github.com/HUBioDataLab/DrugApp.

Keywords: Approval of drugs, clinical trials, drug patents, machine learning, predictive modeling.

ÖZ

İYİLEŞTİRİLMİŞ İLAÇ GELİŞTİRME İÇİN MOLEKÜLER, KLİNİK ÇALIŞMALAR VE PATENT VERİLERİNİN YAPAY ÖĞRENME TEMELLİ ANALİZİ

Çıray, Fulya

Doktora, Sağlık Bilişimi Bölümü

Tez Yöneticisi: Doç. Dr. Yeşim Aydın Son
Tez Eş Yöneticisi: Doç. Dr. Tunca Doğan

Ağustos 2022, 146 sayfa

İlaç geliştirme, özellikle gereken zaman ve maliyetler açısından oldukça masraflı bir süreçtir. Pek çok umut vadeden ilaç adayı molekül, yetersiz etkinlik veya sağlıkla ilgili beklenmeyen olumsuz etkilere yol açması nedeniyle klinik araştırmaların ikinci veya üçüncü fazları gibi aşamalarda geliştirme sürecinden elenmektedir. Son zamanlarda ilaç firmaları bu sürecin verimliliğini artırmak adına hesaplamalı yaklaşımları değerlendirmektedir. Bu tez çalışmasında, önceki ilaç keşfi ve geliştirme aşamalarından elde edilen özellikler ve makine öğrenmesi kullanılarak, ilaç adaylarına, belirtilen hastalıkların tedavisi için kullanım onayı verilip verilmemesinin otomatik ve hesaplamalı biçimde tahmin edilmesi araştırılmıştır. Çalışmanın başlangıcında, bir ön analiz olarak, ilaç adaylarında spesifik moleküler yapıların varlığının istenmeyen sonuçlara (onay alamamaya) yol açıp açmadığını gözlemlemek için ilaç alt yapılarını inceledik. Çalışmanın ana bölümünde, her bir ilaç endikasyon çiftini heterojen bir nümerik vektör olarak temsil etmek amacıyla ilaçların moleküler ve fizikokimyasal özelliklerini içeren bir dizi öznitelik ile, klinik araştırma ve patentlerle ilgili özellikleri kullandık. Veri toplama, manuel kürasyon ve imputasyon prosedürlerini takiben, nihai hale getirilmiş olan öznitelik vektörlerimiz, 14 farklı hastalık grubunun her biri için bir ilaç onay tahmin modeli eğitmek üzere rastgele orman (RF) sınıflandırıcıları tarafından işlendi. Çapraz doğrulamaya dayalı performans değerlendirmemizde aşağıda verilen aralıklarda değişen yüksek tahmin puanları elde ettik; doğruluk: 0,67-0,81, kesinlik: 0,77-0,82, duyarlılık: 0,77-0,96, F1-skoru: 0,77-0,88 ve MCC: 0,45-0,62. Ayrıca zamansal analizler yaparak yöntemimizin ileriye dönük olarak da başarılı sonuçlar üretebildiğini gösterdik. Bunun yanında, temel bir model ve literatürde yer alan yeni bir yönteme karşı gerçekleştirilen bir performans karşılaştırma analizi sonucunda yaklaşımımızın sağlamlığını ve veriyi genelleme kabiliyetini sergiledik. Ek olarak, ilaç onaylarını doğru bir şekilde tahmin etmek için önemli olarak nitelendirilen özellikleri belirledik ve tartıştık. Bir kullanım örneği çalışması kapsamında, yöntemimizin, önce yasal olarak onaylanan (faz IV),

fakat sonrasında ciddi yan etkiler nedeniyle piyasadan çekilen ilaçları başarılı bir şekilde tahmin edebildiğini gösterdik. Son olarak, şu anda klinik araştırmalar faz I/II/III aşamalarında olan ilaç adaylarının onaylarını tahmin etmek amacıyla önceden eğitilmiş modellerimizi kullandık ve tahmin sonuçlarını sunduk. Çalışmamızın sonuçlarının ve sunduğumuz hesaplamalı aracın ilaç geliştirme sürecinin değerlendirilmesi ve iyileştirilmesi açısından literatüre katkıda bulunacağını umuyoruz. Bu çalışmanın tüm veri kümeleri, kaynak kodu, sonuçları ve önceden eğitilmiş modelleri https://github.com/HUBioDataLab/DrugApp adresinde açık kaynaklı olarak paylaşılmıştır.

Anahtar Sözcükler: İlaç onayları, klinik araştırmalar, ilaç patentleri, makine öğrenmesi, tahmine dayalı modelleme.

To My Family

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Assoc. Prof. Dr. Tunca Doğan and Assoc. Prof. Dr. Yeşim Aydın Son, whose positive and invaluable guidance and support made this work possible and carried me through all the stages of this thesis process.

I would also like to thank Prof. Dr. Rengül Çetin Atalay, Prof. Dr. Erden Banoğlu and Asst. Prof. Dr. Aybar Can Acar, for their brilliant suggestions and for letting my presentations be fruitful moments.

To my friends, Heval Ataş, Fatma Cankara, Esra Nalbat and Kübra Kılıç, this would have been a much more difficult process without you, thank you for your countless support.

I am extremely grateful to my mother, Sevim Çıray, for her continuous support and understanding during my research and writing my thesis.

My very special thanks go to a very special person, to my daughter, Defne Duygu, who continuously adds color and gives meaning to my life.

I would also like to acknowledge Science Fellowships and Grant Program (TÜBİTAK-BIDEB) for granting under TUBITAK 2211-A General Domestic PhD Scholarship Program during my research study.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ADME	Absorption, Distribution, Metabolism, Excretion
CPC	Cooperative Patent Classification
ECFP	Extended Connectivity Fingerprint
EPO	European Patent Office
FDA	U.S. Food and Drug Administration
HBA	Hydrogen Bond Acceptor
HBD	Hydrogen Bond Donor
IPC	International Patent Classification
KNN	K-Nearest Neighbors
MACCS	Molecular ACCess System
MCC	Matthews Correlation Coefficient
MDI	Mean Decrease Impurity
MW	Molecular Weight
NDA	New Drug Application
NIH	National Institutes of Health
RF	Random Forest
PCT	Patent Cooperation Treaty
PFI	Permutation Feature Importance
PSA	Polar Surface Area
RO5	Lipinski's Rule of Five
USPTO	The United States Patent and Trademark Office
VS	Virtual Screening
WIPO	World Intellectual Property Organization
XGBoost	Extreme Gradient Boosting

CHAPTER 1

INTRODUCTION & MOTIVATION

1.1. Drug Development Process

Drug development is a long-term, multifaceted process that starts from the target identification and extends to the human trials of the drug. Since a drug's successful entry into the market requires various factors, including significant resources, well-equipped research facilities, and advanced project management, this process is highly costly (around $900 million to $2 billion) and takes approximately 12-15 years (Deore et al., 2019).

The drug development process includes the following steps (Figure 1.1.):

(i) Identification and validation of the target,

(ii) Identification and optimization of the lead,

(iii) Characterization of the product,

(iv) Formulation,

(v) *In vitro* studies,

(vi) *In vivo* studies,

(vii) Clinical trials

This process also covers the patenting process, which usually starts after identifying a valuable chemical in treating an indication and lasts around 30 months until national entry in the case of international PCT (Patent Cooperation Treaty) applications (WIPO, 2022). The duration of the patenting process (from the application date until the date of a grant) varies from country to country, depending on the national legislation. The protection of a particular patent lasts 20 years and is intended to last sometime after the drug's marketing.

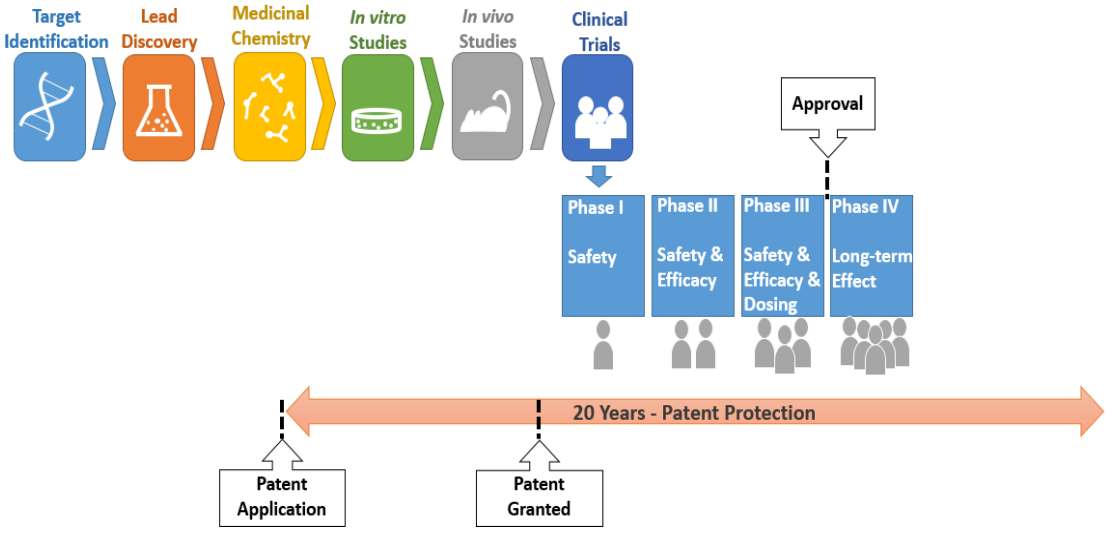


Figure 1.1. The Stages of Drug Development Process.

1.1.1. Computational Approaches in Drug Development

As drug development involves expensive, time-consuming, and laborious processes such as high-throughput screening and clinical trials, computational methods have been developed and used to increase the efficiency of this process (Rifaioglu et al., 2018). The employment of data-centric approaches in data mining and artificial learning has been gaining popularity in computational drug discovery and development. The majority of these studies focus on the prediction of novel drug candidate compounds that would be bio-active against selected targets and diseases or on the prediction of ADME and toxicological properties of these compounds, the structural and functional properties of targets (Doğan et al., 2021; Rifaioglu et al., 2020; Yang et al., 2019, Doğan et al., 2018). However, many good drug candidates are eliminated at late development phases due to both health-related and unrelated (i.e., administrative, legal, etc.) factors. As every drug candidate needs to undergo clinical trials followed by the approval by the responsible authority before taking a place in the market, developers must determine drugs with high chance of approval as early as possible to save time, money, and effort. In this context, the idea of estimating the outcome of clinical trials stands out as a non-trivial endeavor.

1.1.2. Factors Affecting Drug Development

In this part of the thesis, the factors mainly investigated throughout this study regarding drug development are explained in detail below. These factors consist of molecular structures and physicochemical properties of drugs and clinical trial and patenting processes for drugs.

1.1.2.1. Molecular Structures

Molecular structures of drugs contain important hints for drug discovery. Computational drug discovery methods regarding molecular structures, called structure-based virtual screening (VS), utilize structure information of both target and candidate compounds. In order to be used for virtual screening methods, compounds must be quantized by molecular descriptors.

Molecular descriptors can be defined as representative numerical vectors for compounds, constructed by algorithms depending on the various features of compounds, including physiochemical, structural, and geometrical properties, using line notations such as SMILES as input. One well-known sub-group of molecular descriptors consists of binary vectors called fingerprints. Each dimension of the binary vector corresponds to 1: presence or 2: absence of a specific feature to define compounds in terms of their structural elements, chemical bonds, connectivity pathways, and functional groups. Different molecular descriptors can represent different compound features. Molecular descriptors can be categorized into the following groups (Rifaioglu et al., 2018):

(i) 0D molecular descriptors (molecular weight, atom number, etc.(Todechini & Consonni, 2008));

(ii) 1D molecular descriptors (substituent atoms, functional groups, etc. (Todechini & Consonni, 2009));

(iii) 2D molecular descriptors (topological, graph invariants, etc.);

- Path-based (e.g. DayLight (Sastry, et al., 2010)),
- Substructure key-based (e.g., MACCS (Duan et al., 2010)),
- Circular (e.g., ECFPs (Rogers and Hahn, 2010)),

(iv) 3D molecular descriptors (geometrical, binding site, etc.);

- Pharmacophore (e.g., hydrophobicity (Wood et al., 2012)),

- Geometrical (e.g. triangular (Todechini & Consonni, 2009)),

(v) Non-structure-based molecular descriptors (substring occurrence in SMILES; etc.);

- LINGO (Vidal et al., 2005).

1.1.2.2. Physicochemical Drug Properties

The physicochemical properties of drugs, examples of which include solubility, lipophilicity, number of hydrogen bond acceptors and donors, polar surface area, and molecular weight, will affect the conversion of biologically active drugs to therapeutically active compounds in our body and the pharmacokinetics of the drugs (Kerns & Di, 2008). For a biologically active compound to pass all the drug development processes until the marketing of a drug requires it to have drug-like properties. Several guidelines exist, such as "rule of five (RO5)" regarding drug-likeness to decrease attrition in the drug development process. According to one of the most widely used guidelines, RO5, the following physicochemical properties are suggested to be consistent with the following criteria (Lipinski, 2001):

(i) LogP (1-octanol–water partition coefficient) <5,

(ii) Molecular weight <500 Da,

(iii) The number of hydrogen-bond donors (HBD) <5,

(iv) The number of hydrogen-bond acceptors (HBA) <10.

Among the aforementioned physicochemical properties, LogP is the measure of the lipophilicity of a drug and can be considered one of the most important drug-like properties (Leeson and Springthorpe, 2007).

In addition to the physicochemical properties mentioned in the "rule of five," polar surface area (PSA), bioavailability, and the number of rotatable bonds have also been revealed to have roles in reducing the attrition rates in the drug development process (Hou et al., 2007; Linnankoski et al., 2006; Veber et al., 2002; Lu et al., 2004). While PSA relates to the count of O + N atoms together with LogD (1-octanol–water coefficient at different pH), bioavailability is the portion of an oral dose that enters into the bloodstream and based on first-pass metabolism, dissolution, and gut transit time (Leeson and Springthorpe, 2007). Additionally, aromatic ring counts that are higher than three can lead to a higher risk of attrition in the drug development process (Ritchie et al., 2009). Furthermore, specific 3D pharmacophoric and structural properties comprising polarizability will have a role in the binding affinity for protein

active sites regarding drug development (Leeson and Springthorpe, 2007). Finally, Ghose (1999) states that refractivity measures above 40-130 can correlate with poorer drug development processes.

1.1.2.3. Clinical Trials

Clinical trials constitute one of the most important parts of the drug development process, performed on human volunteers to answer drug safety, efficacy, and dosing questions. Clinical trials mainly consist of four phases.

During phase I, a safety assessment is conducted using 20-30 healthy volunteers, as shown in Figure 1.1. Patients are used if healthy people cannot tolerate the drug's action mechanism. The dosage regimen is used according to the animal studies conducted. The side effects of the drugs are determined. Around 70% of drugs can proceed with phase II trials (Deore et al., 2019).

During phase II, several hundred people conduct safety and efficacy assessments. By providing additional safety data, this phase allows researchers to prepare new Phase III protocols. Around 33% of drugs can proceed with phase III trials (Deore et al., 2019).

Phase III trials are conducted with 300 to 3,000 volunteers. During this phase, safety, efficacy, and dosing assessments are performed. The majority of the safety data comes from the phase III trials. As these trials are performed using a large number of people and longer duration, log-term or uncommon side effects can also be determined. Around 25-30% of drugs can proceed to the next phase (Deore et al., 2019).

After providing the data showing that the drug is safe and effective, the FDA review team carefully inspects and decides to approve a drug (Friedhoff, 2009). For a new drug application (NDA), along with all preclinical and clinical trial data until phase III, patent information is also required (Lal, 2015).

After the FDA approval of a drug, phase IV trials are conducted to determine the long-term effect of the drug while the drug is on the market. Based on the complication reports, decisions regarding the addition of precautions, such as dosage information, can be made, as well as withdrawal of the drug from the market for much more adverse drug reactions. (Suvarna, 2010).

1.1.2.4. Patenting Process

A patent is defined as a right granted for an invention. Patents provide a/n new/alternative technical way of doing something or a/n new/alternative technical way to solve a problem. In order to obtain patent protection, all the technical information required to implement the invention must be disclosed to the public (WIPO, 2022).

Patents can be classified according to the Cooperative Patent Classification (CPC) system managed by EPO and USPTO, depending on the particular field of the invention (EPO, 2022). CPC is an extension of the International Patent Classification (IPC) and is divided into nine main parts (A-H and Y) and various sub-groups. Patents regarding medical preparations belong to the "A61K" class.

Patenting process of the drugs constitutes one of the important parts of the drug development process (Figure 1.1). This process usually starts after identifying a valuable drug chemical in treating an indication and involves mainly formal control, as well as search and examination procedures regarding novelty, inventive steps, and industrial applicability of a drug. The duration from the application date until the grant date can show variations from country to country. Additionally, factors such as inadequate experimental support and clarity objections can extend the duration of patent procedures.

The protection of a particular patent lasts 20 years. However, the drug development process takes so much time until the drug is presented to the market, and in order to protect the balance between new drug innovation and generic drug competition, the FDA grants certain exclusivities after the approval of a drug. Exclusivities vary from six months to seven years (FDA, 2022).

A patent application is composed of the following main parts (EPO, 2022):

(i) Abstract, where the aim and the content of the invention are briefly mentioned,

(ii) Description, where the field of the invention, the aim of the invention, prior art, the detailed explanation of the invention, and figures are present,

(iii) Drawings, non-mandatory part of the patent application,

(iv) Claims, where the parts of the invention that are desired to be protected are present.

(v) Sequence listing, if present.

The "Claims" section can be considered the most critical part of a patent application as this section constitutes the part where the actual protection for the drug is requested.

This section represents small molecule drugs as Markush structures (Figure 1.2). An example Markush claim can be written as “an alcohol, shown as R-OH, wherein R is chosen from CH3-, CH3CH2- and (CH3)2CH-” (Gardner and Vinter, 2010). Markush claims usually include combinations of different groups of substituents around a common core molecule, which in the end, let the patent cover multiple compounds with common characteristics or similar physical and chemical properties (Valance, 1961).

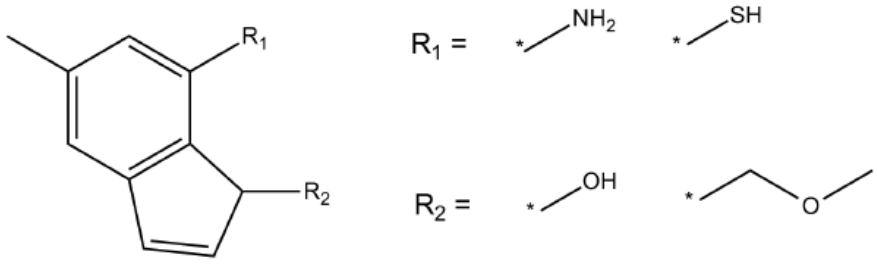


Figure 1.2. Example of a Markush structure that corresponds to four specific compounds. R1 and R2 indicate potential radical substituents (Gardner and Vinter, 2010).

1.2. Problem Statement

Despite previous efforts, drug approval prediction is still understudied, a significant problem that should be solved and translated into real-life pipelines to effectively decrease the overall costs of developing new drugs. This may be possible by increasing the scope and size of the data used in training these predictive models. Different layers of this data can be found on open access resources; however, the main difficulty is integrating and using this data due to differing formats, poor cross-referencing, and the highly unstructured nature of data in these resources.

1.3. Scope and Objectives

The main aim of this thesis study is to investigate the early prediction of the regulatory approval status of drug candidates using the information previously gathered during the process of drug development via data-centric computational approaches, primarily focusing on the utilization of multiple types of input/source data and its effects on the obtained results.

Our first objective was to determine drug substructures, which could be significant for getting approval, using molecular fingerprints of drugs (as input) and the random forest (RF) algorithm in the framework of a preliminary analysis. Our second objective was to propose a new methodology and implement it in a "DrugApp" tool. The application utilizes available clinical trial- and patent-related data together with physicochemical and molecular features of drug candidate compounds within a random forest classifier to predict their potential of getting regulatory approval (i.e., phase IV transition) as drugs targeted for specific indications. The main contribution of this methodology is utilizing multiple data resources to enrich the data at hand in terms of both the size and scope by both automated and manual curation processes, thus presenting a valuable benchmark dataset, carefully conducting the modeling process by pre-processing and ML model training, obtaining high predictive performance in the end, and finally sharing the source code, datasets and the results of the study in full open access.

Patenting can be considered one of the most important drug development processes since patent documents may hold clues about the approval status that comes much later in the development pipeline. However, utilizing patent data in a model is difficult due to the primarily unstructured representation of information and variations in both the availability and formatting of patent documents in different countries. In order to address these issues, here, we combined information from various databases to obtain specific patent-related features for each drug. To our knowledge, these patent features are used here for the first time. We constructed an extensive dataset of 14 different disease groups using multiple types of features, including molecular fingerprints, physicochemical compound features (e.g., lipophilicity, molecular weight, polar surface area, etc.), and clinical trial-based features (e.g., intervention model of the trial, gender/age of participants, etc.), on top of the patent related ones (e.g., duration of the patent process, the number of claims of the patent application, etc.), all extracted from non-proprietary databases. We hypothesize that all these features contain necessary signals regarding the outcomes of clinical trials and the subsequent approval.

We used the random forest (RF) algorithm within a supervised binary classification framework (i.e., approved vs. unapproved) and employed mean imputation in order to deal with the missing data. We performed a predictive performance analysis using various evaluation metrics including accuracy, precision, recall, F1-score and MCC, in the course of an ablation study. In order to reveal the temporal effects on approval prediction success, we conducted an analysis over six different time frames. Our finalized models were employed to predict the outcomes of drugs withdrawn from the market after regulatory approval, to observe whether the model is successful in detecting these critically important cases. Finally, we analyzed a large set of experimental drugs (i.e., drug candidates) that are currently in different phases of the drug development process.

We hope that our study will shed light on the types of features that are important for drug approval, which can be utilized by both commercial and non-commercial entities to increase the efficiency of drug development procedures. Also, we believe that our newly constructed datasets will be useful for further computational studies as benchmarks. Furthermore, we showed that our method has the potential to reveal drugs that were withdrawn from the market sometime after their approval within a certain level of confidence, which indicates that it may be possible to act upon those cases in the first place to prevent critical losses regarding human health potentially.

1.4. Structure of the Thesis

This thesis consists of five main chapters. In the first chapter, background information regarding drug development and the factors affecting this process is given to provide a clear picture of the problem. The main problem statement and the scope and objectives of this thesis are also presented in this chapter.

In Chapter 2, we provided a literature review about clinical trial outcome predictions. We also presented the type of artificial learning techniques used in the literature and compared the performances of previous models for predicting clinical trial outcomes.

In Chapter 3, we explained our methodology from dataset construction to modeling. Additionally, we presented a temporal analysis approach and methods for determining essential features for predicting drug regulatory approvals.

In Chapter 4, we presented and discussed the results for models belonging to 14 different disease groups and important features for predicting drug approvals. We also provided a case study regarding drugs withdrawn from the market after approval. Moreover, we presented regulatory approval predictions for ongoing clinical trials.

In the final chapter, we provided concluding remarks and discussed how the findings of this thesis might be utilized and further developed in future studies.

CHAPTER 2

LITERATURE REVIEW

2.1. Artificial Learning Applications in Clinical Trial Outcome Predictions

In this part of the thesis, we presented previous works regarding clinical trial outcome predictions, including the prediction of drug regulatory approvals after phase III of clinical trials.

Table 2.1 summarizes some recent notable works for predicting clinical trial outcomes. Most recent clinical trial outcome prediction studies utilize the random forest algorithm. Therefore, in the following sections, we will analyze previous models for predicting phase transitions of clinical trials depending on the type of utilized artificial learning technique.

Table 2.1. The list of studies from the drug approval / clinical trial outcome prediction literature.

Year	Reference	Method/Technique	Aim of the Study	Findings
2014	Malik et al.	Fisher's exact and chi-square tests	Prediction of drug approval from phase I results	A higher number of partial responses and longer response durations predict approval.
2015	DiMasi et al.	Logistic regression, random forest classifier	Prediction of anticancer drug approval after phase II	Activity, number of patients, and prevalence-related measures are essential for predicting drug approval.
2016	Gayvert et al.	Random forest classifier	Prediction of toxicity regarding drug approvals	Drug target network connectivity, expression levels, and MW are essential for adverse clinical events.
2016	Artemov et al.	Deep neural networks	Prediction of clinical trial outcomes after phase I/II	Metabolic pathways regarding oxidation and endoplasmic reticulum are among the important predictors for failures.

Table 2.1. (cont.)				
2017	Jardim et al.	Chi-square test	Determination of potential risk factors for failure at phase III for anticancer drug candidates	Potential risk factors include lack of a biomarker-driven strategy and failure to attain proof of concept in phase II.
2018	Lo et al.	Random forest classifier	Prediction of drug approval after phase II/III	The most important features for predicting approval include trial outcomes, trial status, trial accrual rates, duration, prior approval for another indication, and sponsor track records.
2019	Beinse et al.	Cox regression	Prediction of anticancer drug approval after phase I	Drug characteristics, trial design, cytotoxic chemotherapy targets, or categories have important roles in the prediction.
2020	Feijoo et al.	Random forest classifier	Prediction of phase transition after phase II/III	The number of endpoints and the complexity of the eligibility criteria is associated with success.
2020	Zhavoronkov et al.	Deep learning-based proprietary AI tool	Prediction of clinical trial outcome after phase I/II/III	-
2022	Fu et al.	Graph neural networks	Clinical trial outcome prediction after phase I/II/III	Using multi-modal data sources increases the success of models. The lack of sufficient data (e.g., diseases with low prevalence) hinders the performance.

Below, we grouped each model based on the machine learning algorithm utilized in the study.

2.1.1. Logistic Regression Models

Logistic regression is a classification algorithm that implements a nonlinear sigmoidal function to linear feature combinations, and the basic version of this method models binary outcomes (Cox, 1958). Logistic regression is frequently used as a benchmark for clinical prediction models (Christodoulou et al., 2019). Previous studies used logistic regression to predict anti-cancer drug approval after phase I/II (DiMasi et al., 2015; Beinse et al., 2019).

DiMasi et al. (2015) used logistic regression to predict new anti-cancer drug approval after phase II. Clinical trial features were utilized in this study. As a result of logistic regression analysis linking various factors to the probability of regulatory approval of a drug (success), the following factors provided an important basis for predictions: pivotal trial size, activity, number of patients treated, and the duration of phase II. The most significant impact was obtained from the activity variable.

2.1.2. Cox Regression Models

Cox regression is used to assess the link between the rate of survival and variables. Risk ratio (RR) can be defined as the measure of each variable risk. If RR equals 1, the risk is the same for each participant. On the other hand, if RR is higher than 1, RR is representative of increased risk, and vice versa. For example, if RR equals 5.5, the patients with a variable are 5.5 times more likely to face the studied outcome (Wendy, 2007).

Beinse et al. (2019) studied anticancer drug approval predictions after phase I trials using lasso penalized cox regression. A multivariable Cox model with lasso penalization was trained to predict the approval-free survival for each antineoplastic agent. PubMed abstracts belonging to phase I clinical trials regarding antineoplastic agents were utilized with pharmacologic data from the DrugBank database to model the time to regulatory approval (approval-free survival) since the first phase I publication. Performance analysis was conducted using a weighted concordance index (IPCW), and an IPCW of 0.89 on the independent test set was obtained.

2.1.3. Random Forest Models

Random forest (RF) is a machine learning algorithm that build classification or regression models consisting of a combination of decision trees and employs bagging and feature randomness during the construction of an individual tree (Breiman, 2001). Random forest algorithm, frequently used in clinical trial outcome predictions, was also used in this study and explained more in detail in Section 3.4.

Gayvert et al. (2016) used drug-likeness (RO5), drug physicochemical features and drug target-based features in a random forest model to predict whether the drug will fail clinical trials because of toxicity reasons. As important features for adverse clinical trial events; molecular weight (MW) together with drug target network connectivity and expression levels were determined. They used 10-fold cross-validation on a set of 784 regulatorily approved drugs with known targets and the drugs associated with 100 failed toxic clinical trial drugs having at least one target and known structure. They determined the model's predictive performance with an area under the curve (AUC) of 0.8263.

Lo et al. (2018) used multiple types of data (mostly about clinical trials) obtained from proprietary databases to predict the outcome of clinical trials for various disease groups. They utilized random forest classifiers to predict drug approvals using clinical trial data across 15 different disease groups. Drug-indication pairs were used in this study. The predictors achieved performances of 0.78 AUC for predicting transitions from phase 2 to regulatory approval and 0.81 AUC for predicting transitions from phase 3 to regulatory approval. They determined the most important features for the prediction of drug regulatory approval as trial outcomes, trial status, rates of trial accrual, trial duration, earlier drug approval for another indication, and sponsor track records.

The model from Feijoo et al. (2020) was also based on clinical trial features. They used random forest classifier for the prediction of phase transition after phase II/III of clinical trials. The predictor achieved an average accuracy of 0.80. Moreover, they also determined that the complexity of the eligibility criteria and the number of endpoints constituted common protocol characteristics across different therapeutic fields that are linked to drug approvals.

2.1.4. Deep Learning Applications

Nowadays, deep learning methods have intensively been used in various fields of technology. Multi-layer perceptrons constitute the simplest form of neural networks, which is unable to account for temporal dependencies. On the other hand, recurrent neural networks utilize feedback connections in order to model temporal behavior. Deep learning techniques have strong performances; however, there exists lack of interpretability in their decision-making process (Shamout et al., 2021).

Artemov et al. (2016) developed deep learning techniques and biology analytical tools which utilize drug side effect information to predict the drug approval for phase I/II clinical trials. First of all, they performed the predictions of the side effects of a drug by deep neural networks and drug-induced pathway activation estimation. Next, they used the probabilities of the predicted side effects and the pathway activation scores

as an input to train a classifier to predict the clinical trial outcomes. The predictor of this study achieved an accuracy of 0.83.

Zhavoronkov et al. (2020) used wide range of features for the prediction of reaching phase II and phase III, including molecule descriptors; ADMETox data, biomedical documents, clinical trials, and drug/target omic features. However, the output of this study is a proprietary method/tool/system, and researchers cannot obtain either the source code, datasets, or detailed results.

In a recent work, Fu et al. (2022) presented the Hierarchical Interaction Network (HINT) method, in which drug molecule, target disease, and trial eligibility criteria embeddings are obtained together with drug pharmacokinetic and trial data-based knowledge-embeddings. Then these features are connected by a hierarchical interaction graph to capture their interactions and predict the outcomes of phase I, II, and III clinical trials.

These studies have shown that it is possible to predict drug approvals, or clinical trial outcomes, using machine learning and data science, and revealed important predictors thereof. Overall, they pave the way for developing more advanced and systematic methods/tools to be used in actual drug development pipelines.

CHAPTER 3

METHODS

3.1. Preliminary Data Analysis

In this part of the thesis, we constructed random forest models for 14 disease groups to determine important drug substructures using PubChem fingerprints as input features to the model.

3.1.1. Dataset Construction

For the construction of datasets of regulatorily approved and unapproved drug classes, please see the dataset construction part of the main model (Section 3.2.1). The only difference with the datasets of the main model resides in the removal of SMILES up to length 6. Here, we did not remove SMILES data up to length 6 because they could also contain important structural information regarding approved and unapproved drugs.

3.1.2. The Use of PubChem Substructure Fingerprints as Features

PubChem provides structured data about fragments of molecules, called PubChem substructure fingerprints. A fingerprint consists of an ordered list of binary bits, each corresponding to a presence (1) or absence (0) of an element, a ring system type, atom environment as nearest neighbors etc., in a compound structure. PubChem fingerprints (881 bits) correspond to 881 substructures, divided into seven sections, as seen in Table 3.1 (Bethesha, 2009). Therefore, we decided to use these fingerprints as input features in a machine learning model and then determine important features (substructures). PubChem fingerprints (881 bits) were generated using the PubChemPy package.

Table 3.1. The main sections' explanations and examples that are available in the PubChem substructure fingerprints. Abbreviations; FPS: fingerprints.

PubChem FPS Section Number	PubChem FPS Section Name	Examples to PubChem FPS
1	Hierarchic Element Counts	>= 4 H, >= 1 Fe
2	Rings in a canonic Extended Smallest Set of Smallest Rings (ESSSR) ring set	>= 1 any ring size 3, >= 1 hetero-aromatic ring
3	Simple atom pairs	Li-H, O-Na
4	Simple atom nearest neighbors	C(~Br)(~C), N(~C)(~C)
5	Detailed atom neighborhoods	C(-N)(=C), N-C:N:C
6	Simple SMARTS patterns	O=C-C:C, S=C-N-C
7	Complex SMARTS patterns	Cc1ccc(Cl)cc1, NC1CC(N)CCC1

3.1.3. Modeling for the Determination of Important Drug Substructures

We used the random forest classifier as a machine learning algorithm. For the modeling methodology, please see the modeling part for the main model (Section 3.2.3). We extracted the top twenty most informative substructures using MDI-based feature importance and the built-in function of the RF classifier, which was implemented in the Python scikit-learn package (please see Section 3.2.5).

3.2. Construction and Evaluation of the Drug Approval Prediction Model

3.2.1. Dataset Construction

Figure 3.6 summarizes the overall system of DrugApp, including the data gathering procedure, which is also explained below. All the data used in this study are extracted from public databases that are listed as follows:

(i) Clinicaltrials.gov (Zarin et al., 2011), the most extensive database comprising clinical trial information,
(ii) DrugBank (Wishart et al., 2006), a database containing information about approved and investigational drugs,
(iii) ChEMBL (Gaulton et al., 2017), a chemistry database comprising small molecule compound information and their bioactivities against targets,

(iv) SureChEMBL (Papadatos et al., 2016), a chemically annotated patent document database including text and image-based compound information,
(v) PatentsView (USPTO, 2019) is a database containing patent information from the United States Patent and Trademark Office.

In each clinical trial record, a drug candidate is tested against one or more disease(s); as a result, in our study, we evaluated a drug's approval status considering the reported indication in the respective trial record. Clinicaltrials.gov also provides 25 condition groups in which these indications are classified. However, most of them have a small sample size. Due to this reason, we decided to merge certain groups that could have common properties; for example, we constructed an "Infective diseases" group, which is obtained by merging "Bacterial and Fungal Diseases" with "Viral Diseases." Additionally, condition groups such as "Occupational Diseases" were discarded as we considered diseases in these groups would not have common characteristics. This way, we constructed datasets of drug-indication pairs that cover 14 disease groups as; rare diseases, nervous system, alimentary tract and metabolism, cancers and other neoplasms, dermatological, urinary tract/ sexual organs and pregnancy conditions, heart and blood diseases, immune system diseases, infective (bacterial, fungal and viral), respiratory tract (lung and bronchial) diseases, gland and hormone-related diseases, blood and lymph conditions, musculoskeletal diseases and sensory (eye, ear, nose, and throat). Table 3.2 displays these disease groups and the number of drug-indication pairs under each group. The Clinicaltrials.gov database originally contained 25 condition groups.

After that, we downloaded SMILES notations and names for each drug/compound obtained in the previous step from DrugBank and ChEMBL databases. In order to construct the dataset of regulatorily approved drugs, we searched for drug names in "phase IV" clinical trial records in the Clinicaltrials.gov database. During this search, we ignored trials in which drug combinations were utilized. To exclude the small auxiliary ions, compounds with SMILES notations up to the length of six characters were also removed from the dataset. We also added approved drug information from the DrugBank database for the corresponding indications on top of the ones extracted from the Clinicaltrials.gov database. In the end, we obtained 14 datasets containing approved drug-indication pairs, one for each disease group.

For constructing the unapproved drug datasets, we searched for drugs in clinical trial records with "Suspended," "Terminated," or "Withdrawn" status and place them into the unapproved dataset of the corresponding disease group, similar to the previous studies (Lo et al. 2018) (Figure 3.6). We also added unapproved drug data from the DrugBank database by considering the aforementioned status for the corresponding indications. Again, drug combinations were ignored, and SMILES up to the length of six characters were removed to finalize the disease group-based unapproved drug datasets.

Table 3.2. The number of regulatorily approved and unapproved drugs and drug-indication pairs in each disease group.

Condition/Disease group	# of approved drug-indication pairs	# of unapproved drug-indication pairs	# of unique approved drugs	# of unique unapproved drugs
All Diseases	18,427	2,480	1,995	1,127
Rare Diseases	3,886	1,768	1,133	835
Nervous System Diseases	6,397	808	1,132	555
Alimentary tract and Metabolism Diseases	4,832	861	1,075	555
Cancers and Other Neoplasms	2,450	1,530	869	704
Dermatological Diseases	3,246	802	930	521
Urinary Tract, Sexual Organs, and Pregnancy Conditions	3,215	818	914	506
Heart and Blood Diseases	4,372	516	883	365
Immune System Diseases	2,494	848	837	495
Infective (Bacterial, Fungal and Viral) Diseases	3,766	582	906	397
Respiratory Tract (Lung and Bronchial) Diseases	2,147	688	699	470
Gland and Hormone Related Diseases	2 ,050	648	631	374
Blood and Lymph Conditions	1,038	775	463	403
Musculoskeletal Diseases	1,483	240	516	202
Sensory (Eye, Ear, Nose and Throat) Diseases	1,378	306	451	218

As shown in Table 3.2, the overall dataset consists of 3,122 unique small molecules, 8,324 unique indications, and 20,907 unique drug-indication pairs. The disease group "Rare Diseases" makes up the largest subgroup. For most subgroups, approved drug data outnumber the ones in the unapproved category.

A drug may have multiple disease group tags. For instance, a drug may exist both in "Cancers & Other Neoplasms" and "Immune System Diseases" in the Clinicaltrials.gov database. Moreover, one disease group may contain a particular drug multiple times as the drug can be used for multiple indications belonging to the same disease category. On the other hand, a drug labeled as "unapproved" in one disease group may be categorized as an approved drug for another group. We allowed these cases as this does not cause any contradictions.

3.2.2. Featurization

We used four types of features (i.e., molecular descriptors of drugs, physicochemical properties of drugs, clinical trial-related features, and patent-related features) to build vectors representing each drug-indication pair data point to be given as input to our machine learning models that predict drug approval. Details about these features are given in both Table 3.3 and Figure 3.6. We utilized extended connectivity fingerprints with the diameter value of two (ECFP4) with 128-bits size (Rogers and Hahn, 2010) using the publicly available tool RDKit (Landrum et al., 2006). ECFPs are among the most widely used similarity search tools in drug discovery and contain information regarding the environments of each atom (Carracedo-Reboredo et al., 2021). The physicochemical properties of molecules (i.e., logP, MW, PSA, HBA, HBD, number of rings, polarizability, refractivity, number of rotatable bonds, bioavailability) were extracted from the DrugBank database. These features are evaluated early in the drug development process for assessing drug-like properties and are generally used for rules regarding drug-like properties such as RO5. The clinical trial-related features were obtained from the Clinicaltrials.gov database, including the gender and age of participants, allocation of trial to assign participants to an arm of a clinical study (non-randomized, randomized, etc.), intervention model of the trial (single group assignment, parallel assignment, etc.), funder of the trial (Industry, NIH, etc.) and locations by country, that were also utilized by previous studies (Beinse et al., 2019; Feijoo et al., 2019; Lo et al., 2018). Finally, we used patent features comprising CPC (Cooperative Patent Classification), number of related patents, number of claims, presence of international application (PCT: Patent Cooperation Treaty), and duration of patenting process. CPC classifies patents with respect to various aspects such as depending on whether a drug contains active organic ingredients such as hydrocarbons or phenols, thereby giving clues regarding the structural properties of drugs. For some related patents with the presence of international applications, we speculated that if a drug is highly promising, drug developers may tend to apply for more patents in more

countries. Claim numbers mainly determine the number of properties to be protected by a patent. A high number of claims may reflect an alternative high number of side groups of the primary drug, extensive methodology for obtaining that drug, or additional properties, etc. For the duration of patenting process, this duration can extend depending on several factors. For example, if the patent application is evaluated to lack some properties such as inventive step, patent applicant/s can raise objections, thereby leading to re-examination as well as the extension of the duration from the patent application until the grant of a patent.

Table 3.3. Types, names, descriptions, and formats of features used in our machine learning-based drug approval predictor.

Type and name of the feature	Description	Format
Molecular features of drugs		
ECFP4 (128 bits)	Molecular descriptor; Extended-connectivity fingerprint	Categorical
Physicochemical features of drugs		
logP	Lipophilicity	Numerical
MW	Molecular weight	Numerical
PSA	Polar surface area	Numerical
HBA	Hydrogen bond acceptor	Numerical
HBD	Hydrogen bond donor	Numerical
Rings	Number of rings	Numerical
Polarizability	Polarizability	Numerical
Refractivity	Refractivity	Numerical
Bonds	Rotatable bond count	Numerical
Bioavailability	The likelihood that the drug will exhibit oral bioavailability	Categorical
Clinical trial features		
Gender	Gender of participants	Categorical
Age	Age of participants	Categorical
Allocation type	Allocation of trial to assign participants to an arm of a clinical study (Non-Randomized, Randomized, etc.)	Categorical
Intervention model	Intervention model of the trial (Single Group Assignment, Parallel Assignment, etc.)	Categorical
Funder type	Funder of the trial (Industry, NIH, etc.)	Categorical
Locations	Locations by country	Categorical
Patent features		
CPC	Cooperative Patent Classification, patent classification of the patent application	Categorical
Number of related patents	Patent numbers containing drug information in the part of the claims before the regulatory approval date of the drug	Numerical
Claim numbers	The number of claims of the patent application.	Numerical
PCT	The presence of the international Patent Cooperation Treaty application	Categorical
Duration	Duration of the patent process in months (from application date to grant date)	Numerical

The drug patent data's usability is lower than the other types of drug features in the literature. One of the reasons for this situation is the specific demonstration of drugs in patent documents, called the "Markush structure" which covers all possible formulations comprising the actual drug to expand the scope of legal protection. Markush claims usually include combinations of different groups of substituents around a common core molecule, which in the end, let the patent cover multiple compounds with common characteristics or similar physical and chemical properties (Valance, 1961). Thus, finding the drug/compound of interest among all these claims is difficult. Other reasons that reduce the usability of drug patent data are non-standardized ways of accessing patent data in different resources and differences in patent document formats. USPTO has provided bulk patent data to increase readability and requires applicants to submit chemical structures as MDL molfiles and ChemDraw CDX binary files since 2001. However, for the majority of patent authorities around the world, patent data can only be manually extracted in the form of unstructured text documents (Papadatos, 2016). In order to generate our patent features by addressing these issues, we gathered patent information by combining data from multiple databases: USPTO database, SureChEMBL, and DrugBank following these steps:

(i) The patent ID number (USPTO Patent ID) for a drug/compound was extracted from the DrugBank database by only taking the earliest patent application of that drug into account (to avoid subsequent patent applications that may be processed during or after the clinical trial procedures).

(ii) For the drugs that did not have a patent ID in DrugBank, the SureChEMBL database was utilized. In order to extract patent data from SureChEMBL, first, USPTO patents (and if not present in USPTO, patents from other countries) containing the drug of interest in the "claims" part of the documents were gathered. The reason behind searching the "claims" part is that this is the main section of patent documents where the actual drug to be protected shall be given. Among the retrieved patents for a particular drug, the earliest patent (by publication date) was assumed to be the document of interest for a given drug.

(iii) After obtaining all patent IDs, patent features were gathered from DrugBank, SureChEMBL, USPTO, or Google Patents (the latter is only used when the required data is not found in formerly listed databases).

- For the "number of related patents" feature, all obtained patents whose publication date is before the regulatory approval date of the drug of interest were taken into account.

- The values of the "number of claims" feature were obtained from SureChEMBL and Google Patents.

- The values of “duration”, “PCT”, and “CPC” features were extracted from the USPTO database and Google Patents.

The final feature vector comprises the combination of ECFP4, physicochemical, clinical trial- and patent-related features, respectively, 128, 10, 6, and 5 dimensions in size. Thus, the size of the finalized feature vector was 149 dimensions.

A drug may exist in multiple drug-indication data points in a disease group’s dataset (each time paired with a different indication). In these cases, molecular, physicochemical, and patent features are replicated for the drug of interest, and only clinical trial features differ (since each drug-indication pair comes from a different clinical trial). This situation could lead to over-optimistic performance results due to the high similarity between training and test samples. To avoid this, we discarded duplicate data points of drugs under each disease group by randomly filtering all but one data point for each drug. This way, we ensured that our performance evaluation was not biased regarding data leakage from training to test datasets.

In our dataset, the data types of features are heterogeneous. For instance, clinical trial features are categorical, whereas molecular and physicochemical drug features are numerical. Finally patent features are a combination of numerical and categorical values. We converted categorical values into one-hot encoded features to distinguish between these different data types.

In order to deal with missing values in our dataset, mean and median imputation methods were employed. Figures 3.1-3.3 summarize the visual representation of missing data. Mean imputation was used for logP, MW, PSA, polarizability, refractivity, and the duration of the patent process, while median imputation was used for HBA, HBD, number of rings, number of claims, number of rotatable bonds, and the number of patents. Figure 3.1 shows patterns of missing data for physicochemical drug features. 26% of physicochemical features were missing from the approved dataset and 25% from the unapproved dataset.

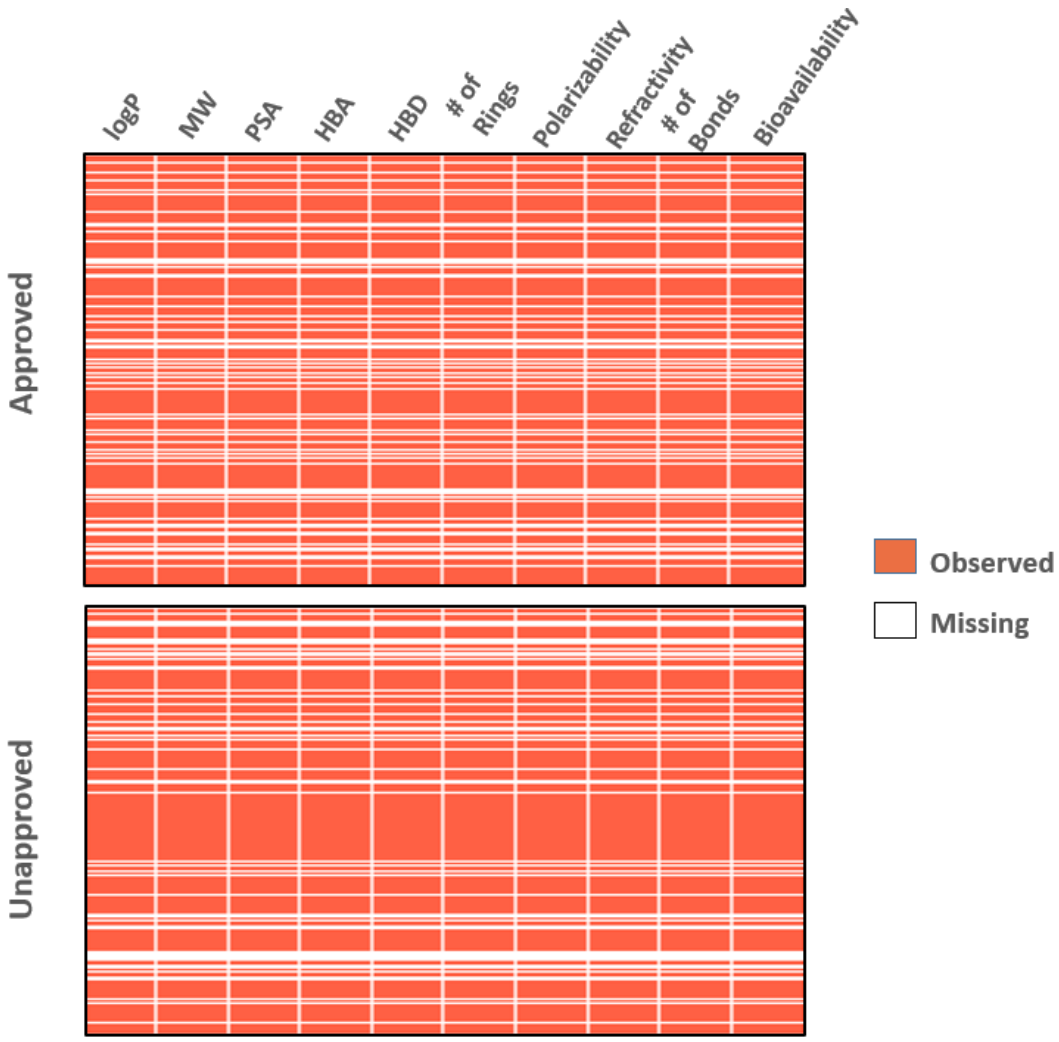


Figure 3.1. Patterns of missing data for physicochemical drug features. Abbreviations; #: number.

Figure 3.2. shows patterns of missing data for clinical trial-related features. The missing values varied from 15 to 26% for clinical trial features of approved drugs and 25 to 32% for clinical trial features of unapproved drugs.

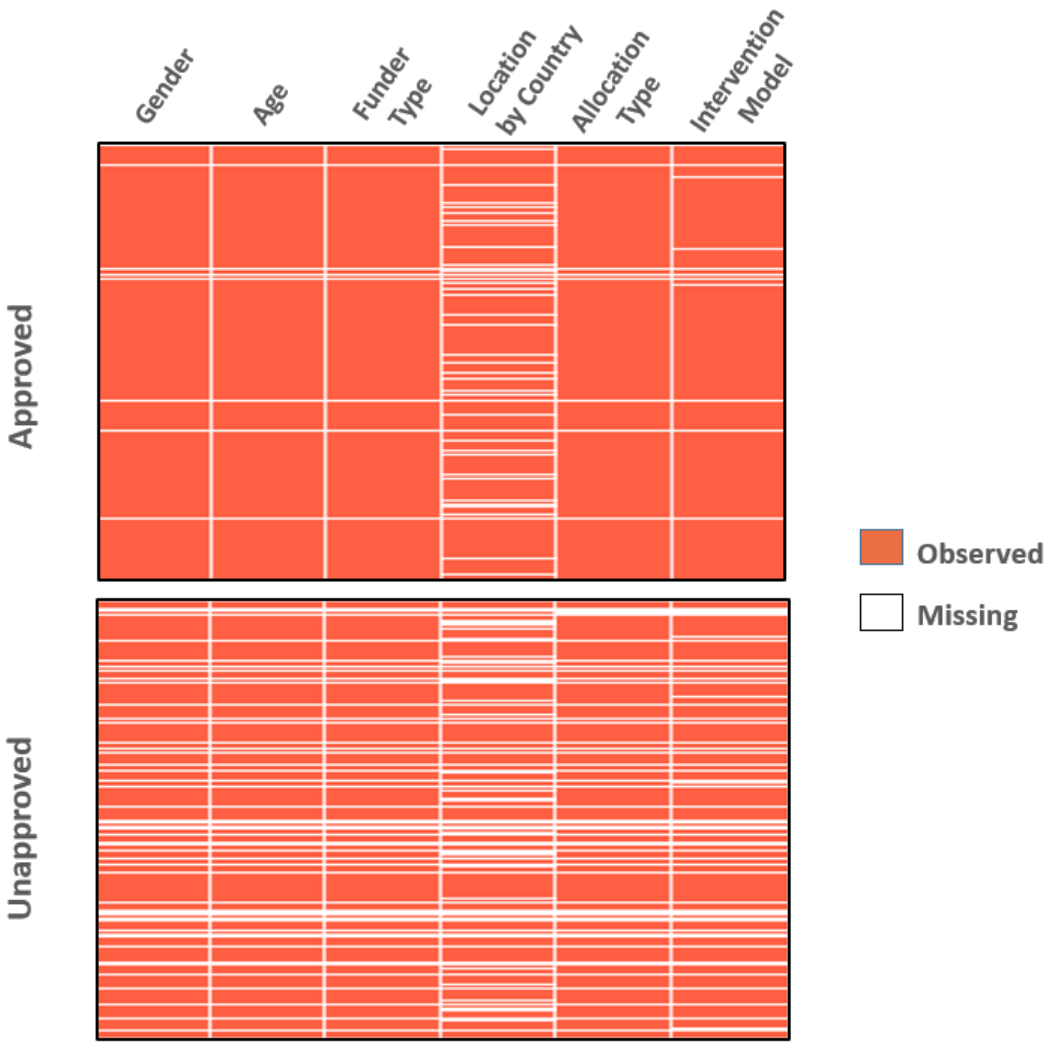


Figure 3.2. Patterns of missing data for clinical trial features.

Figure 3.3. shows patterns of missing data for patent-related features. For patent features, the percentages of missing values were 18 to 35% for approved drugs and 23 to 51 % for unapproved drugs.

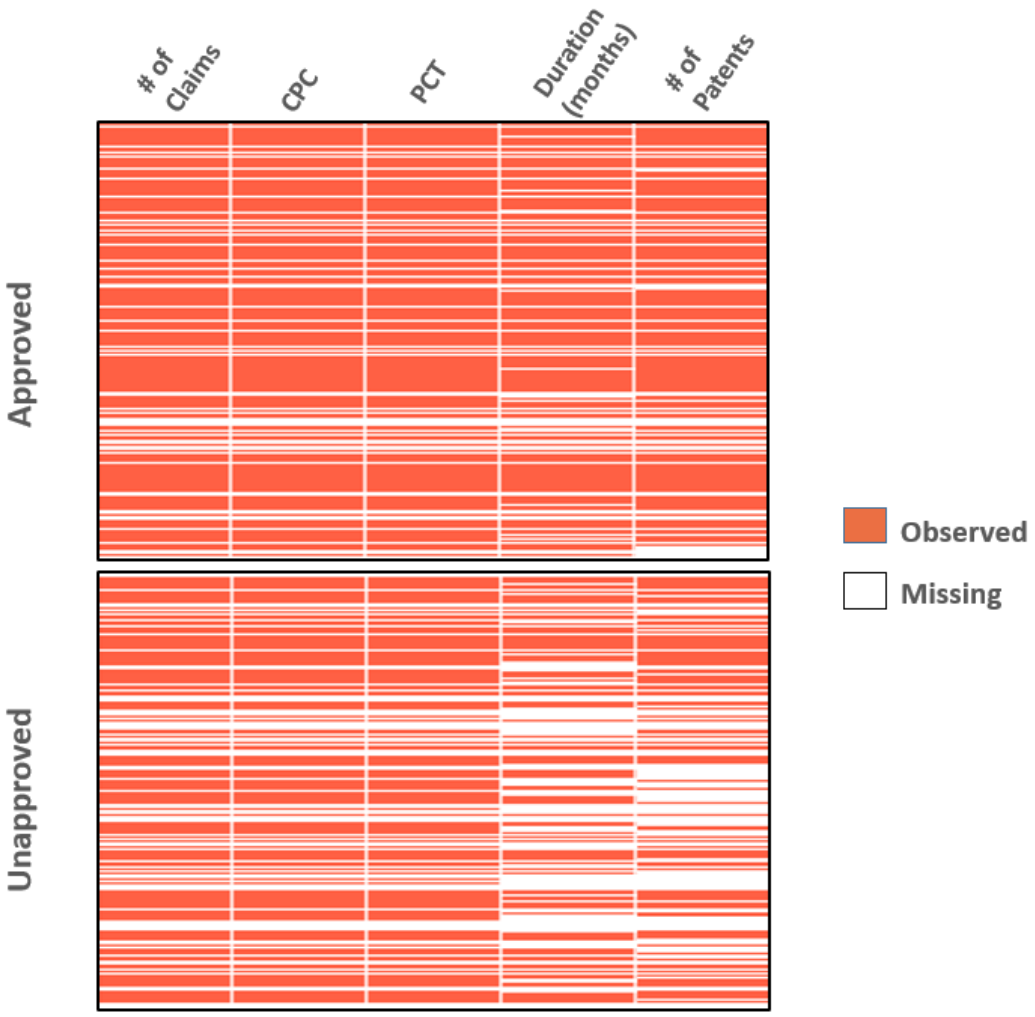


Figure 3.3. Patterns of missing data for patent features. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of international Patent Cooperation Treaty application, # of Patents: number of related patents.

Figure 3.4 shows the percentages of missing ECFP4, clinical trial, patent, and physicochemical features for each drug in the dataset. Each row corresponds to a specific drug, and all the column values of a row represent values that belong to the same drug.

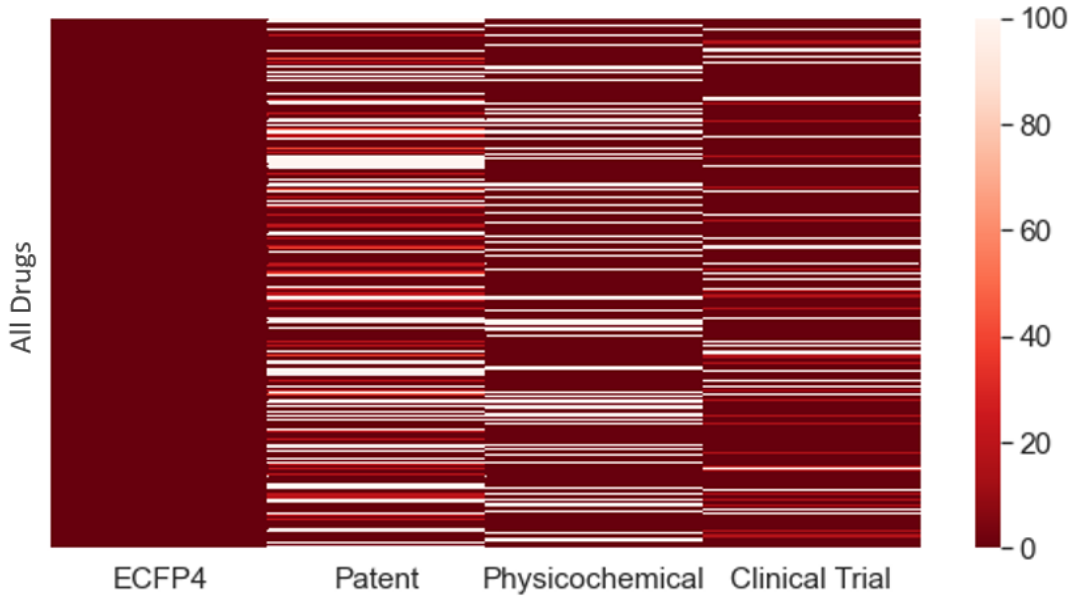


Figure 3.4. Patterns of missing data for all features. The bar on the right-hand side shows color codes representing the percentage of missing data.

We also conducted an empirical study in which we analyzed the performance of the imputation process by artificially filling varying degrees of missing features (i.e., 0%, 25%, 50%, and 75%) to a subset of the data at hand, which do not contain any missing features originally.

3.2.3. Modeling

Here, we modeled the task of predicting the approval (i.e., phase IV transition) of a drug-indication pair as a supervised binary classification problem (output classes: "approved" and "unapproved") using the random forest (RF) algorithm, given a set of input features. For each drug-indication pair (i.e., data point or sample), the corresponding molecular, physicochemical, clinical trial, and patent features are combined and presented to the prediction model as a fixed-sized feature vector (Figure 3.6).

Random forest (RF) is an ensemble learning algorithm that builds classification or regression models using a combination of decision trees. When constructing an individual tree, RF employs bagging and feature randomness. For classification tasks, the class that most trees choose constitutes the model's output. As the number of trees increases, the generalization error converges to a limit. The strength of each tree and the correlation in-between determine the generalization error of the forest (Breiman, 2001). Like most machine learning algorithms, RF involves hyperparameters to be set

by the user, and tuning these values can lead to improved model performance (Probst, 2019). Due to this reason, we focused on the following hyperparameters:

(i) The maximum number of splits until the final node (max_depth),
(ii) The minimum number of samples to split any given leaf node (min_samples_split),
(iii) The minimum number of samples to be in the leaf node (min_samples_leaf),
(iv) The number of trees/estimators in the forest (n_estimators),
(v) The maximum number of features is considered for the best split (max_features).

We tuned these hyper-parameters using grid search in a scheme of 5-fold cross-validation. The parameter values we used were; the number of trees/estimators: 100, 200, and 400, the maximum number of features: 30% of all features, square root of all features and all features, maximum depth: 4, 16, and no limitation, the minimum number of samples leaves: 1, 2 and 4, the minimum number of samples to split: 2, 4 and 8.

3.2.4. Temporal Analysis

As the processes of drug development, especially clinical trials, continuously change over time (Brunoni et al., 2010), it is important to observe that the success of modeling drug approval (using data from the past and predicting future events) varies over time. For this, we applied a temporal performance analysis. We split clinical trials into six different time frames (Figure 3.5). For instance, the first training time frame comprises clinical trial data until 2008 (starting from the oldest record), and the testing time frame includes the trial data between 2009 and 2020. For the second time frame, the training data is until 2010, whereas the testing data is from 2011-2020, etc. The model was trained and tested on each time slice independently. The RF classifier with the hyper-parameter values of; the number of trees: 200, the maximum feature number: square root of the total number of features; maximum depth: 16; minimum samples leaf: 2; minimum samples split: 2 were used. The models' performances were estimated using F1-score, accuracy, precision, recall, and MCC metrics.

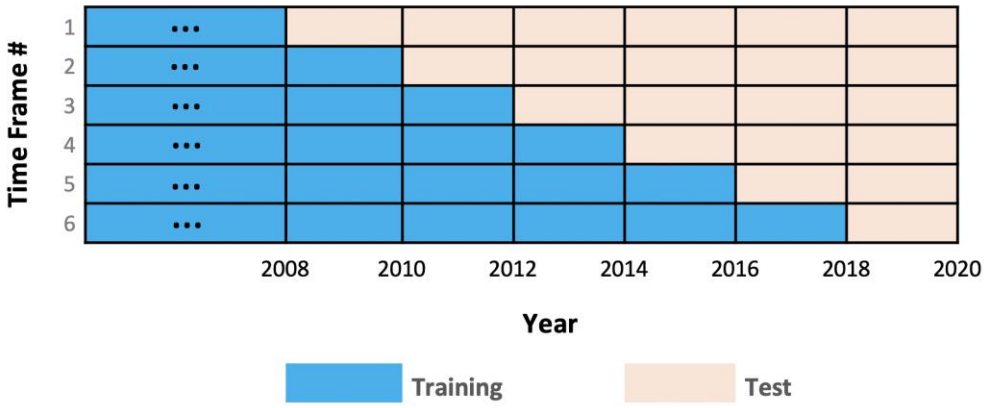


Figure 3.5. Time frames used for the temporal analysis. The first training time frame consists of clinical trial data until 2008; for the second time frame, the training data are selected up to 2010, etc. For each time frame, the test dataset comprises data left outside the training dataset.

3.2.5. Determination of Feature Importance

In this analysis, we extracted each model's top ten most informative features. Feature importance values were determined using the mean decrease impurity (MDI) and permutation feature importance methods. In mean decrease impurity, the decrease in node impurity for all nodes that split on that variable is determined, weighted by the probability of reaching that node, and averaged over all trees (Breiman et al., 1984). As MDI-based feature importance could lead to an inflation of the importance of numerical values, we also calculated permutation importance. Permutation feature importance takes into account the drop in the model's score when the values of the feature of interest are shuffled randomly. This method is based on the breakdown of the relationship between the feature and the target. Therefore, the decline in the model performance shows how much the model is dependent on the feature of interest (Breiman, 2001). We used the n_repeats parameter as 10, determining the number of times a specific feature is randomly shuffled.

3.2.6. Performance Evaluation Metrics

Evaluation metrics are used to observe the performances of prediction models on the testing set, which is important for the fair comparison of the success of different models. Here, we will refer to the approval of a drug as the positive class and the unapproval of a drug as the negative class for the prediction models regarding the regulatory approval of drugs. In classification, predictions of the model are compared to the true labels, which results in the following categories:

(i) TP: True positives, when the positive class is predicted correctly,

(ii) TN: True negatives, when the positive class is predicted correctly,

(iii) FP: False positives, when the positive class is predicted incorrectly,

(iv) FN: False negatives when the negative class is predicted incorrectly.

Among various classification metrics, accuracy represents the fraction of correctly predicted samples across all samples and is considered highly biased in the case of highly imbalanced datasets (Shamout et al., 2021). Precision (i.e., positive predictive value) corresponds to the fraction of correctly predicted positive class among all positive cases. On the other hand, recall refers to the proportion of correctly predicted samples across true positive and false negative cases. As inferred from the definitions, false negatives and false positives are not taken into account by precision and recall, respectively. Therefore, the evaluation of performances of models by only considering a single evaluation metric may lead to inaccurate results. In such a case, F1-score can be utilized, considering both false negatives and positives.

Furthermore, Matthews correlation coefficient (MCC) is another widely used metric and differs from F1-score by taking true negatives into account. The MCC value of zero indicates random prediction, while value 1 corresponds to perfect classification. Because of the equation (Eq. 5), using MCC usually results in lower performance values when compared to other evaluation metrics. Therefore, MCC around 0.5 is generally evaluated as a successful model. The selection of evaluation metrics may vary depending on different cases, and multiple evaluation metrics are suggested to reach more realistic conclusions.

In order to measure the performance of our predictive models, we used multiple evaluation metrics such as; accuracy, precision, recall, F1-score, and Matthews Correlation Coefficient (MCC). MCC is preferred over both accuracy and F1-score for highly imbalanced datasets due to its robustness (Chicco and Jurman, 2020). To further compensate for the imbalance between the sizes of approved and unapproved drug classes in our datasets, we incorporated metrics such as the balanced accuracy and weighted versions of precision, recall, and F1-score (Brodersen et al., 2010; Behera et al., 2019; Opitz and Burst, 2021). In weighted metrics, the performances are independently calculated for each label, and their weighted average is taken considering the number of positive cases for each label. The formulations of the basic versions of these metrics are given as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{FN + TP} \quad (3)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(FN + TP)(TN + FP)(TN + FN)}} \quad (5)$$

$$Balanced\ Accuracy = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \quad (6)$$

where TP, TN, FP, and FN correspond to true positive, true negative, false positive, and false negative predictions.

All work related to machine learning-based modeling has been done using Python scikit-learn package (Pedregosa et al., 2011).

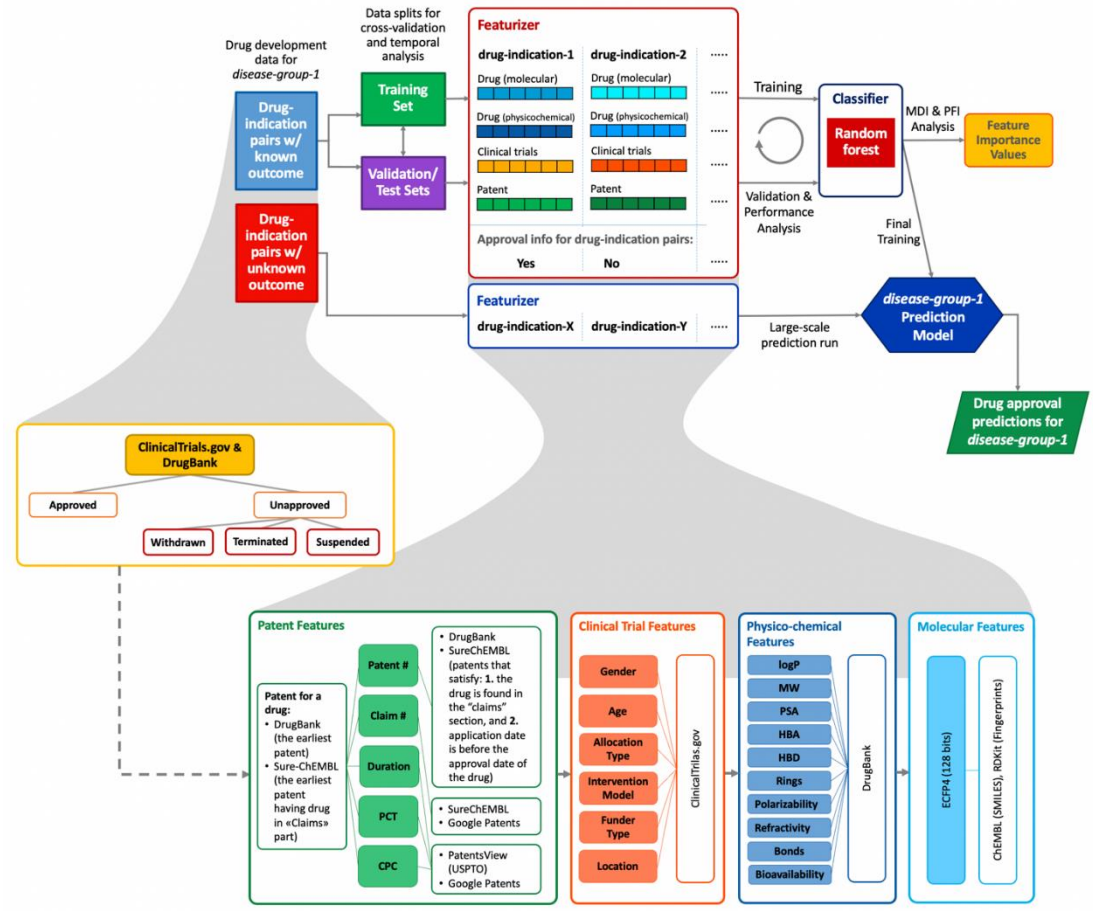


Figure 3.6. Schematic representation of the dataset preparation (left-side), featurization (middle- and bottom-side), and machine learning modeling (right-side) procedures applied in DrugApp, for training and testing a drug approval prediction model on an example disease group dataset (i.e., *disease-group-1*). In DrugApp, each data point is a drug-indication pair, and an independent prediction model is trained for each generic disease group, which are clusters of indications with commonalities (e.g., ovarian cancer belongs to the "cancers and other neoplasms" disease group). Approved and unapproved drug data have been obtained from DrugBank and ClinicalTrials.gov databases. Unapproved drug candidates are selected from clinical trial records with "suspended", "terminated," or "withdrawn" outcomes. Features are gathered from various databases (abbreviations; Claim #: number of claims, Patent #: number of related patents). Multiple training and validation experiments have been carried out to evaluate the performance of models and construct the finalized system.

CHAPTER 4

RESULTS AND DISCUSSION

4.1. Important Drug Substructures

Intending to determine the most important substructures that could give clues regarding drug approvals, we constructed RF-based approval prediction models using only drug substructures (represented by PubChem fingerprints) as input features. We predicted the outcome as drug approval (1) or unapproved (0) in a supervised binary classification scheme. This is followed by determining important substructures using RF built-in MDI-based feature importance function.

Table 4.1 summarizes the performance results of RF models for 14 disease groups. The performances of the models varied in the ranges of accuracy: 0.57-0.73, precision: 0.63-0.75, recall: 0.65-0.89, F1: 0.62-0.81 and MCC: 0.18-0.33. The highest performance results were observed in the "Sensory Diseases" group with the accuracy: 0.64-0.73, the precision: 0.71-0.75, recall: 0.73-0.89, F1: 0.70-0.81, and MCC: 0.33. The lowest performance results were observed in the "Blood Diseases" group, with the accuracy: 0.59-0.65, precision: 0.63-0.69, recall: 0.65-0.80, F1: 0.62-0.74, MCC: 0.19. Moreover, the performance of the combined datasets in "All Diseases" group was observed as accuracy: 0.62-0.67, precision: 0.66-0.71, recall: 0.67-0.82, F1: 0.66-0.76, MCC: 0.27. Most of the models belonging to specific disease groups, such as "Respiratory Diseases", performed better than the combined "all diseases" group as expected (since it contains drugs with more diverse substructures, which would be more difficult for the model to generalize), while others, such as models for "Rare Diseases" and "Neoplasms", performed more poorly. The reason behind this may be "Rare Diseases," and "Neoplasms" groups can also have diverse features due to heterogeneity in associated tissues and target proteins.

Additionally, the performance results of the "Blood Diseases" group are also lower than the combined "all disease" group, probably due to the fact that this group has a lower sample size leading to a weak representation of its space. The overall performances of the models shown in Table 4.1 were observed to be low (i.e., MCC scores are close to 0, near-random predictions). These results indicate that using drug substructures (i.e., PubChem fingerprints) alone as the input feature is insufficient to successfully model regulatory drug approval since many additional factors affect the drug development processes and their outcome.

Table 4.1. The performance of RF models for 14 disease groups. The highest scores for each evaluation metric are shown in bold. Abbreviations are; Acc: accuracy, Accb: balanced accuracy, Pre: precision, Prew: precision weighted, Rec: recall, Recw: recall weighted, F1: F1-score, F1w: F1 weighted, MCC: Matthews Correlation Coefficient, TN: true negative, FP: false positive, FN: false negative, TP: true positive.

Disease	**Acc**	**Accb**	**Pre**	**Prew**	**Rec**	**Recw**	**F1**	**F1w**	**MCC**	**TN**	**FP**	**FN**	**TP**
All	0.67	0.62	0.71	0.66	0.82	0.67	0.76	0.66	0.27	405	562	289	1,357
Rare	0.65	0.61	0.69	0.64	0.80	0.65	0.74	0.64	0.23	246	350	192	772
Nervous	0.68	0.57	0.70	0.64	**0.89**	0.68	0.78	0.64	0.18	102	304	91	721
Alimentary	0.68	0.57	0.71	0.65	0.88	0.68	0.78	0.64	0.19	104	291	92	698
Neoplasms	0.66	0.62	0.68	0.65	0.79	0.65	0.73	0.64	0.26	228	272	157	588
Dermato-logical	0.68	0.59	0.71	0.65	0.86	0.68	0.78	0.65	0.21	113	255	100	636
Urinary	0.69	0.60	0.72	0.67	**0.89**	0.69	0.79	0.66	0.24	104	238	76	608
Heart	0.68	0.57	0.70	0.64	**0.89**	0.68	0.79	0.63	0.18	65	205	58	482
Immuno-logical	0.70	0.60	0.72	0.68	0.89	0.70	0.79	0.67	0.24	105	230	75	595
Infective	0.68	0.57	0.71	0.64	**0.89**	0.68	0.78	0.64	0.18	68	200	60	476
Respiratory	0.71	0.62	0.73	0.70	0.88	0.71	0.80	0.68	0.28	113	196	77	541
Hormonal	0.69	0.59	0.72	0.67	0.88	0.69	0.79	0.66	0.23	74	169	57	429
Blood	0.65	0.59	0.69	0.63	0.80	0.65	0.74	0.62	0.19	99	164	94	371
Musculo-skeletal	0.69	0.59	0.72	0.66	0.87	0.69	0.79	0.66	0.22	45	98	36	250
Sensory	**0.73**	**0.64**	**0.75**	**0.71**	**0.89**	**0.73**	**0.81**	**0.70**	**0.33**	40	61	22	180

Table 4.2 and 4-13-4.26 (please see Appendix A) shows the top twenty most important substructures for 14 disease groups. We observed that drug substructures, for example, ">=16 H", "C(-O)(=O)", "C(~O)(~O)", ">=16 C", "O=C-N-C-C", "C(-N)(=O)", ">=4 any ring size 6", ">=3 any ring size 6", "N(~C)(~C)(~C)", ">=1 saturated or aromatic heteroatom-containing ring size 6", "[#1]-C-C-N-[#1]", "N(~C)(~C)(~H)",

"C-F" and ">=1 F" could be important for predicting drug approvals in "All Diseases" group. Moreover, "C-F" and ">=1 F" substructures were observed to be higher in number in the unapproved drug category (Table 4.2), showing that these substructures may have important associations with failure.

Table 4.2. The top twenty most important substructures for "All Diseases" group. Features correspond to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
>=1 Cl	0.008166	267	180
C-F	**0.006334**	**185**	**204**
C(~H)(~H)(~H)	0.005601	1157	706
>=1 F	**0.005578**	**192**	**204**
O-H	0.005200	918	505
>=16 H	0.005111	1144	754
>= 4 O	0.005076	807	477
C(-O)(=O)	0.004905	629	307
C(~O)(~O)	0.004889	683	338
>=16 C	0.004745	1023	712
O=C-N-C-C	0.004656	529	445
C(~C)(~C)(~C)(~H)	0.004633	505	290
>=1 S	0.004630	361	234
C(-N)(=O)	0.004556	614	514
>=4 any ring size 6	0.004415	225	218
>=3 any ring size 6	0.004388	577	449
N(~C)(~C)(~C)	0.004388	721	485
>=1 saturated or aromatic heteroatom-containing ring size 6	0.004367	570	462
[#1]-C-C-N-[#1]	0.004312	710	555
N(~C)(~C)(~H)	0.004198	656	516

Overall, different substructures were observed to be important in different disease groups. This is due to the fact that the associated tissues and target proteins for drugs in different diseases are dissimilar. The models performed poorly, showing that the prediction of drug approvals using only drug substructure features is not meaningful. The results of this preliminary study show that substructures of drug candidate molecules are not a clear and direct indicator for regulatory approval.

4.2. Prediction of Drug Approvals

4.2.1. Data Exploration

Intending to observe differences between approved and unapproved drugs in terms of the value distributions of a clinical trial, patent, and physicochemical features we employed in this study, we conducted a simple comparison-based exploratory analysis. In Figure 4.1, we provided feature value distribution histograms for the "all diseases" group (i.e., the whole dataset).

The "number of related patents" was calculated by considering the patents containing the drug of interest within the claims. These numbers were obtained from the SureChEMBL database, which implements extensive text and image mining from patent documents covering all types of compounds, including cosmetic products (Papadatos, 2016). Therefore, the "number of related patents" generally returned high values. Nevertheless, the predictive model evaluates these values in comparison to each other (i.e., compares the number of patents for each drug with each other). As a result, it is sufficient to provide consistent values (i.e., calculated using the same procedure). As observed in Figure 4.1, the mean number (dashed vertical lines on each histogram) of related patent applications is much higher among the approved drugs compared to the unapproved ones. This is expected as highly promising drug candidates are to be patented in many countries and tried for a high number of indications for patent protection. Similarly, the number of PCT applications (international patent applications), is observed to be higher in the approved drugs dataset than in the unapproved ones. Considering the "number of claims", the "duration" of the patent procedure, and "CPC", the data distributions for approved and unapproved drugs are not observed to be significantly different from each other.

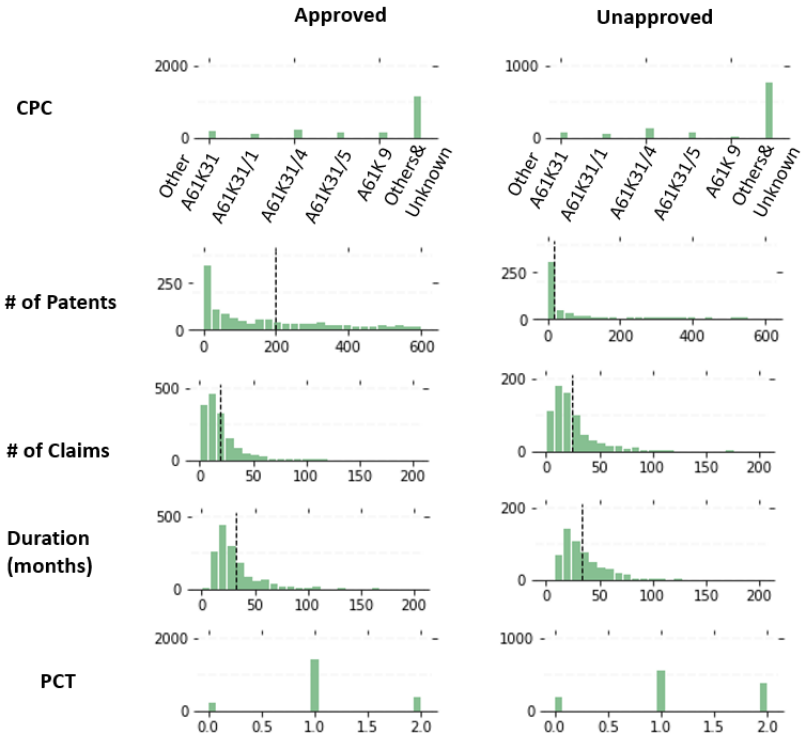


Figure 4.1. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "All Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): the presence of PCT application, PCT (#2): unknown.

For clinical trial features, "age" and "allocation type" slightly differ between the approved and unapproved cases (Figure 4.2). While "parallel assignment" constitutes most of the intervention model types in the approved drugs group, "other intervention models and unknown" and "single group assignment" dominate the unapproved drugs dataset along with the "parallel assignment". While in a parallel assignment, two or more participant groups undertake different interventions/treatments, all participants receive the same intervention/treatment in a single group assignment. The parallel assignment is the most commonly used intervention model and is suggested to be utilized in most trials (Nair, 2019). USA is where most clinical trials are conducted, considering both the approved and unapproved datasets. "Global Participation In Clinical Trials Report" (FDA, 2016) shows that the USA is the leading country in terms of participants in clinical trials (40.83 %), whereas the other countries have less than 10 % participation in clinical trials. Finally, while "industry" is the dominant funder type for unapproved drugs, "other funder types" constitute a more significant portion of the approved cases. This result could be indicative of the risk-taking nature

of the industry. Without risk, a company's path to commercialization would probably not be viable in the pharmaceutical field (Mollah, 2014).

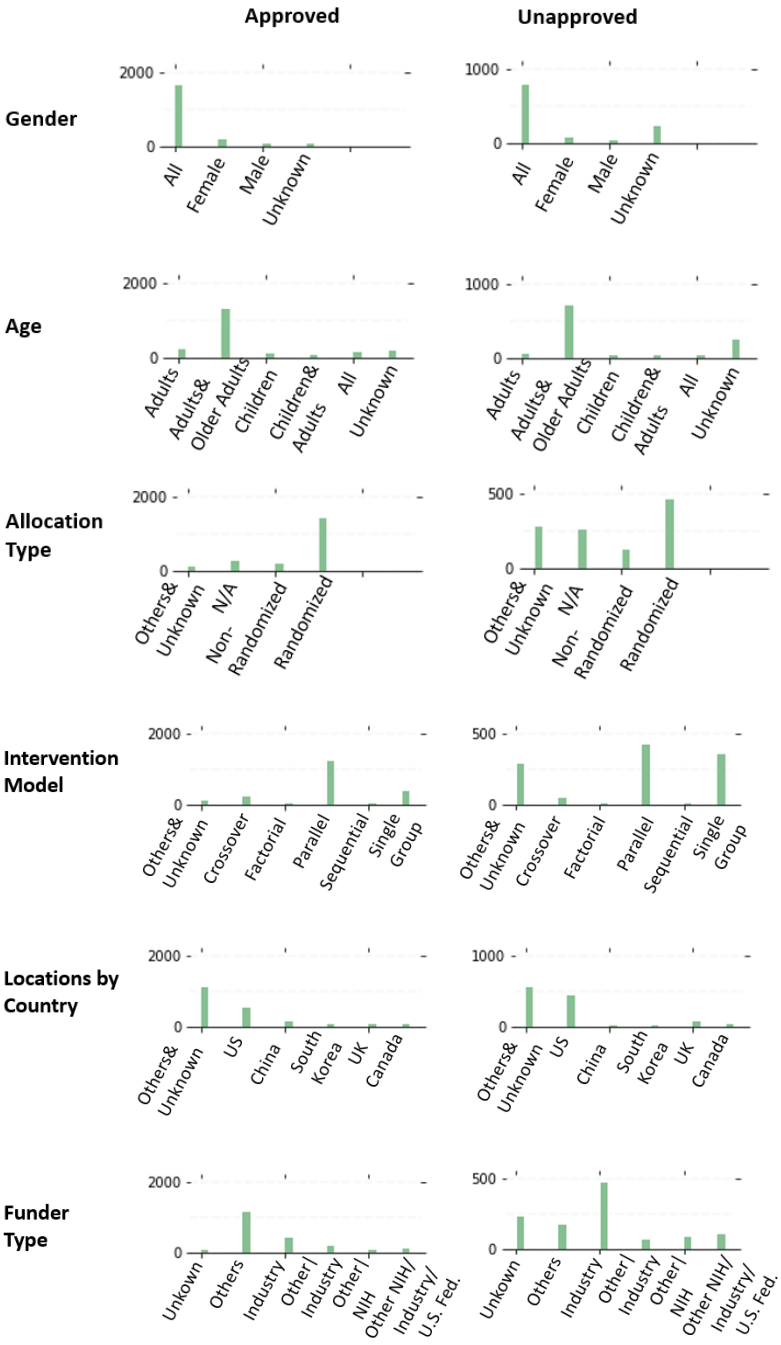


Figure 4.2. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the "All Diseases" class.

Physicochemical properties are critical for the concept of drug-likeness. Although the "rule of five" (Lipinski et al., 2001) has widely been accepted as an approach to reducing attrition in drug discovery, recent trends in newly discovered drugs indicate additional patterns. Our study revealed that the mean molecular weight – MW (417±9.95 Da) and mean the number of rings (2.90±0.05) values for the approved drugs are different from the mean MW (491±17.88 Da) and mean the number of rings (3.16±0.22) values for the unapproved drugs (Figure 4.3). Ghose rule (1999) states that MW should be between 160-480 Da and Ritchie et al. (2009) discloses that more than three aromatic rings in a molecule correlate with poorer drug developability. Considering other drug physicochemical properties, the mean ± error values for approved and unapproved drugs (respectively) are; logP: 1.88±0.09 and 2.46±0.09, polar surface area: 106±0.92 and 125±1.72, hydrogen bond donor: 5.6±0.20 and 6.56±0.34, refractivity: 108±2.41 and 129±4.38, polarizability: 42±0.92 and 50±1.72, and the number of rotatable bonds: 6.82±0.29 and 8.27±0.44. These results indicate apparent differences between the approved and unapproved drugs, which isverified by the student's t-test with p values lower than 0.05 for all listed features. These results are also consistent with the literature (Veber, 2002; Ghose, 1999; Vistoli et al., 2008). Finally, HBD and the bioavailability values of the approved drugs are observed to be similar to the unapproved ones. Please see Figures 4.15-56(Appendix C-E) for the data distributions of specific disease groups, which are largely consistent with the data distributions of the "All Diseases" group.

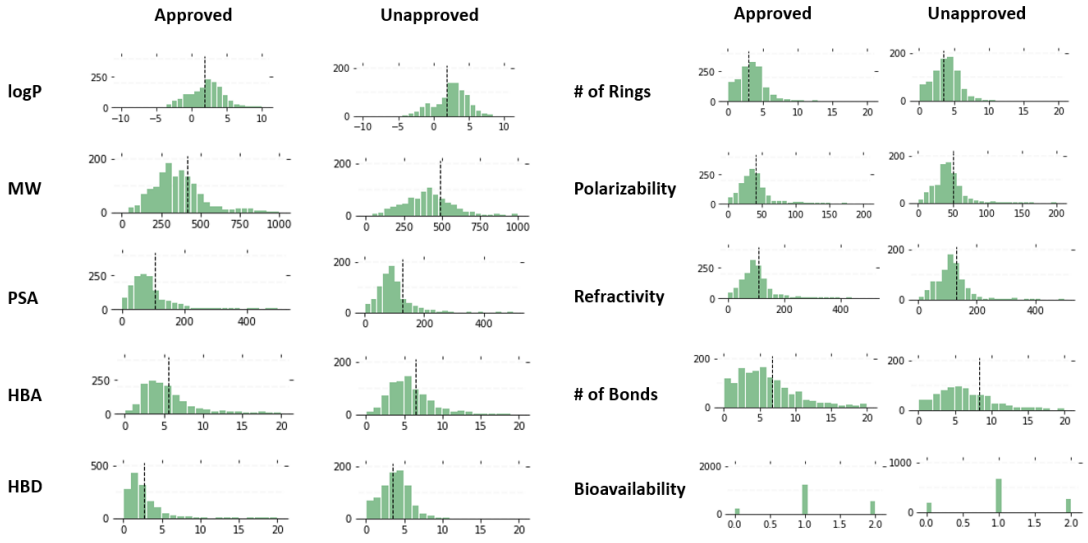


Figure 4.3. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "All Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted

bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. The dashed line corresponds to the mean value for numerical features.

As physicochemical properties are critical for the concept of drug-likeness, which is the fundamental part of the drug development process, we further inspected the physicochemical properties of drugs in detail. We compared the mean values of logP, MW, number of rings, PSA, HBA, HBD, refractivity, polarizability, and number of rotatable bonds across different disease groups (Figures 4.4-4.12).

Figure 4.4 shows the mean logP values of approved and unapproved drugs across 14 disease groups. The mean logP values significantly differ between approved and unapproved drugs in “Rare Diseases”, “Neoplasms”, “Dermatological Diseases”, “Immunological Diseases”, “Infective Diseases”, “Heart Diseases”, “Musculoskeletal Diseases”, “Blood Diseases”, Hormonal Diseases” groups (Student’s t test, p values <0.05).

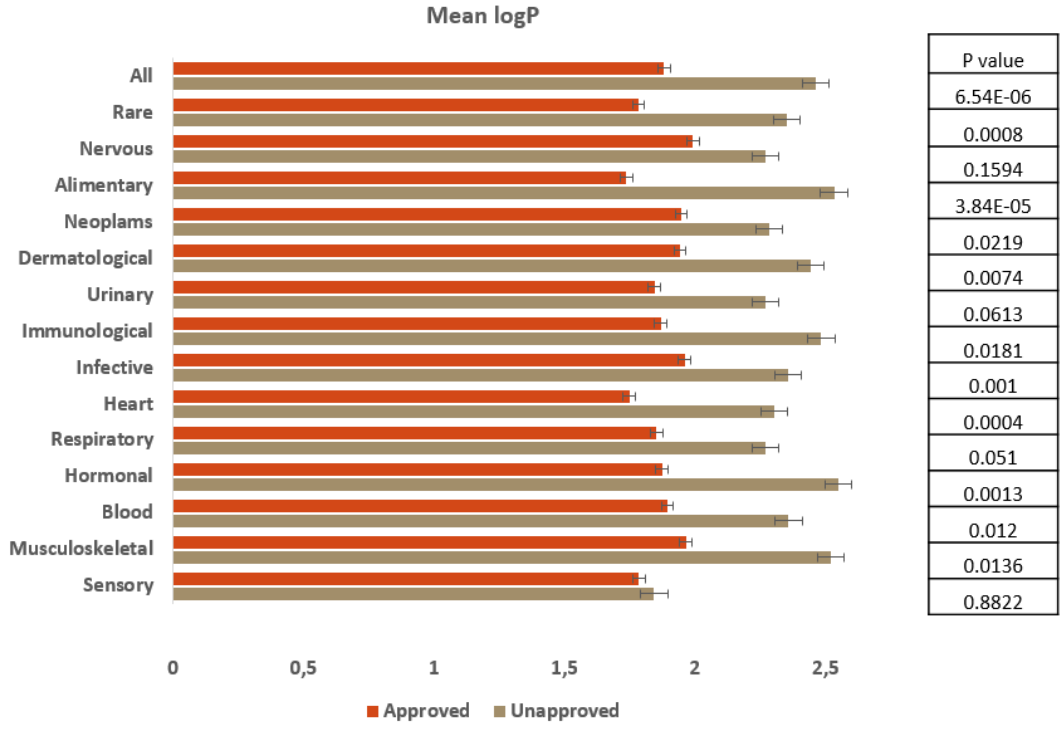


Figure 4.4. Mean logP values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student’s t-test, p values <0.05).

Figure 4.5 shows the mean MW values of approved and unapproved drugs across 14 disease groups. The mean MW values significantly differ between approved and

unapproved drugs in “Rare Diseases”, “Nervous Diseases”, “Neoplasms”, “Dermatological Diseases”, “Immunological Diseases”, “Infective Diseases”, “Blood Diseases”, “Blood Diseases”, Respiratory Diseases” groups (Student’s t test, p values <0.05).

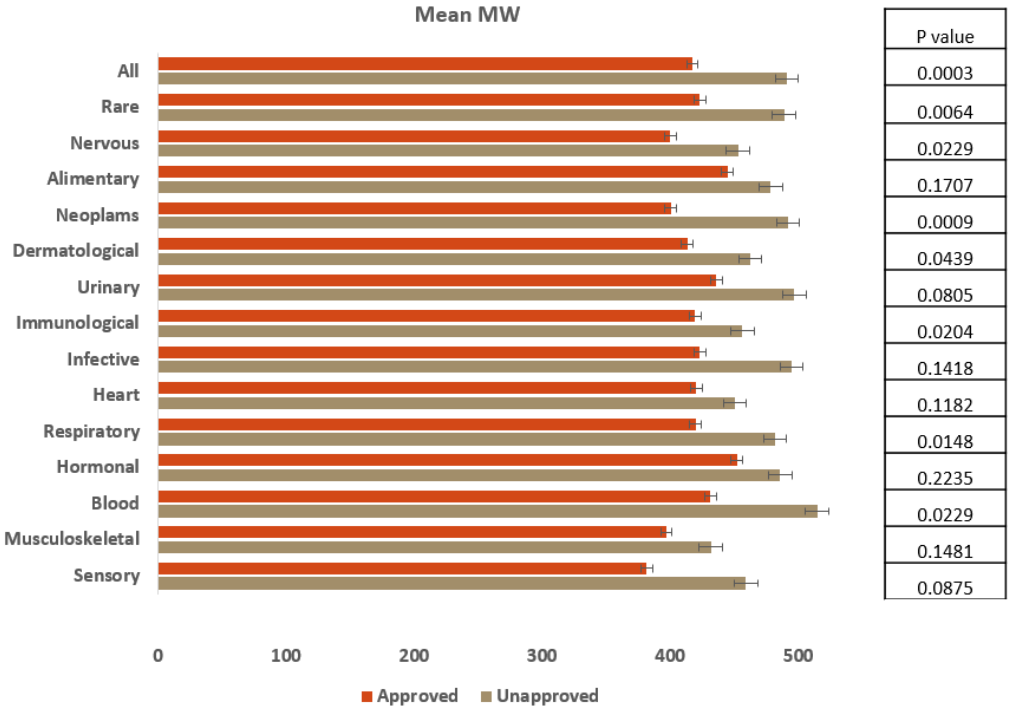


Figure 4.5. Mean MW values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student’s t-test, p values <0.05).

Figure 4.6 shows the mean PSA values of approved and unapproved drugs across 14 disease groups. The mean PSA values significantly differ between approved and unapproved drugs in “Neoplasms” group (Student’s t-test, p values <0.05).

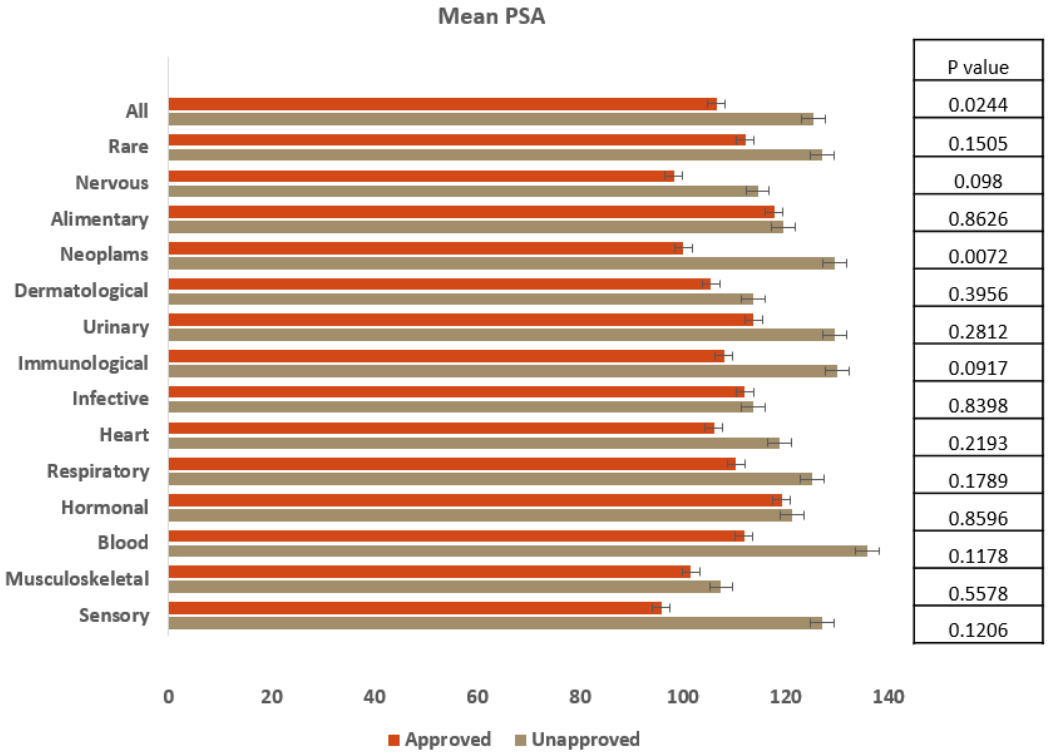


Figure 4.6. Mean PSA values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student's t-test, p values <0.05).

Figure 4.7 shows the mean HBA values of approved and unapproved drugs across 14 disease groups. The mean HBA values significantly differ between approved and unapproved drugs in "Neoplasms" group (Student's t-test, p values <0.05).

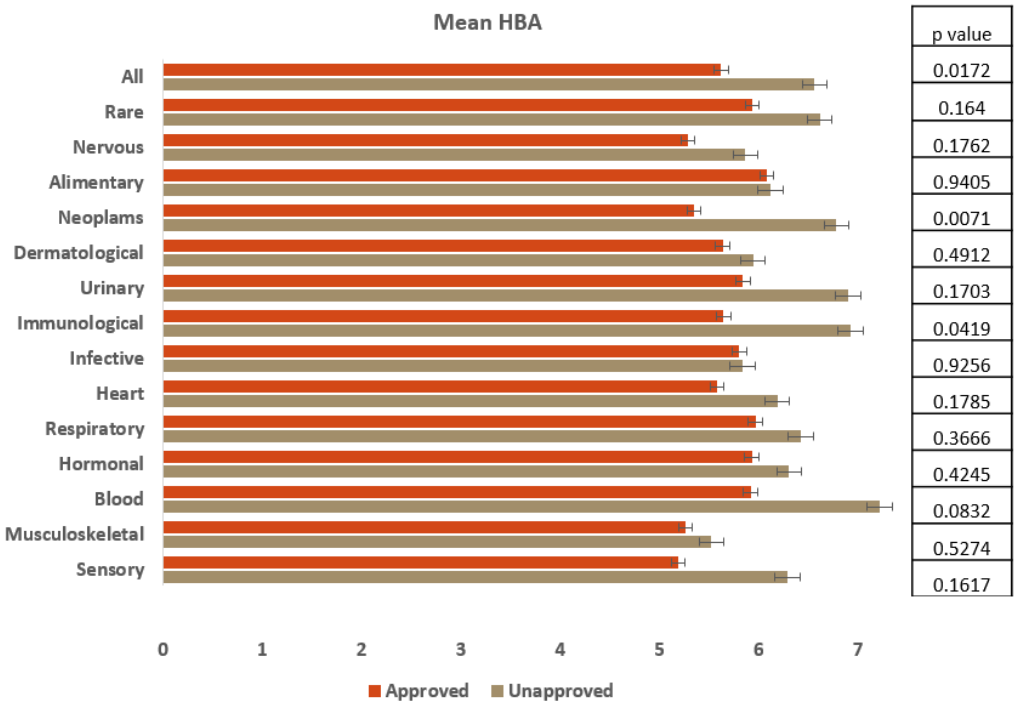


Figure 4.7. Mean HBA values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student's t-test, p values <0.05).

Figure 4.8 shows the mean HBD values of approved and unapproved drugs across 14 disease groups. The mean HBD values significantly differ between approved and unapproved drugs in "Neoplasms" group (Student's t-test, p values <0.05).

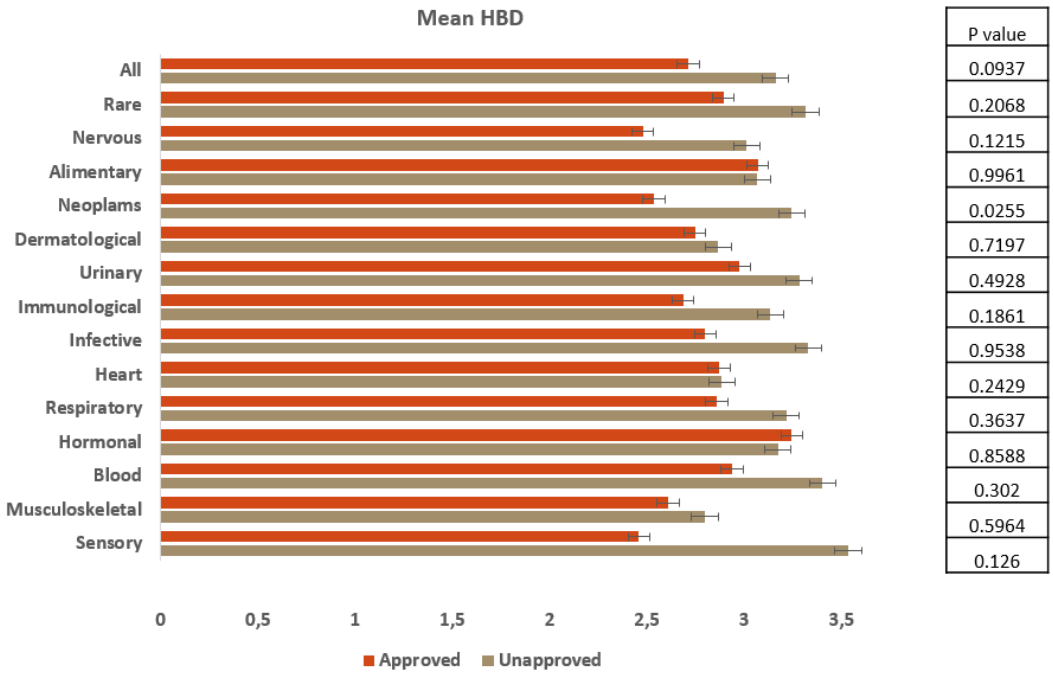


Figure 4.8. Mean HBD values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student's t-test, p values <0.05).

Figure 4.9 shows the mean number of ring values of approved and unapproved drugs across 14 disease groups. The mean number of rings values significantly differs between approved and unapproved drugs in "Rare Diseases", "Alimentary Diseases", "Neoplasms", "Dermatological Diseases", "Urinary Diseases", "Immunological diseases", "Infective Diseases", "Heart Diseases", "Respiratory Diseases", "Hormonal Diseases", "Blood Diseases" and "Musculoskeletal Diseases" groups (Student's t-test, p values <0.05).

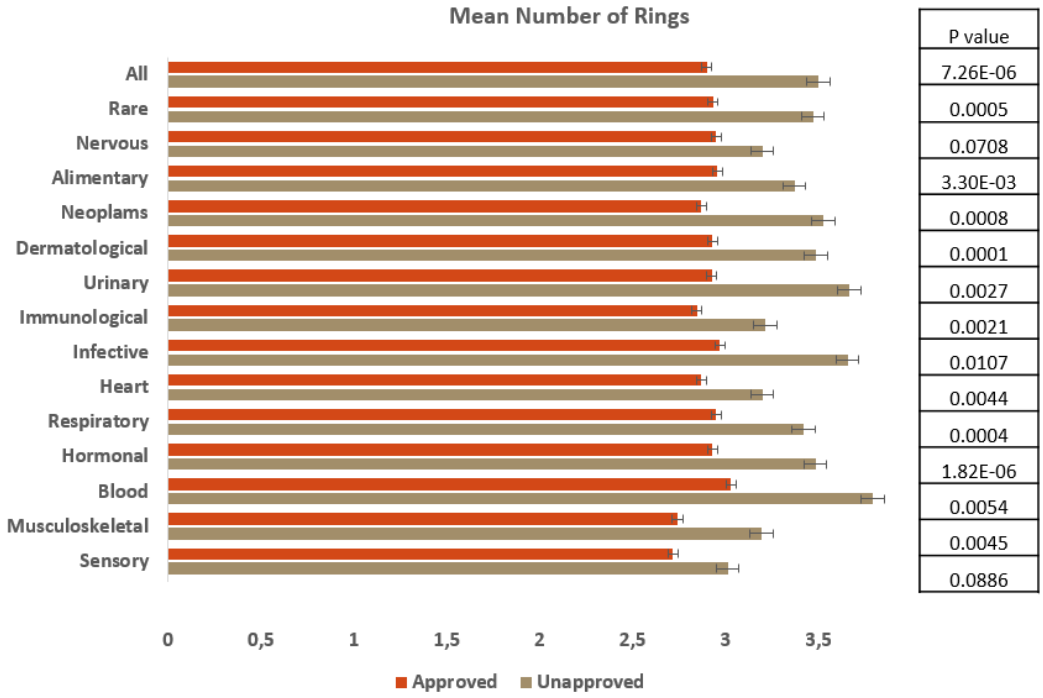


Figure 4.9. Mean number of rings values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student's t-test, p values <0.05).

Figure 4.10 shows the mean refractivity values of approved and unapproved drugs across 14 disease groups. The mean refractivity values significantly differ between approved and unapproved drugs in "Rare Diseases", "Neoplasms", "Dermatological Diseases", "Urinary Diseases", "Immunological diseases", "Heart Diseases", "Respiratory Diseases", and "Blood Diseases" groups (Student's t test, p values <0.05).

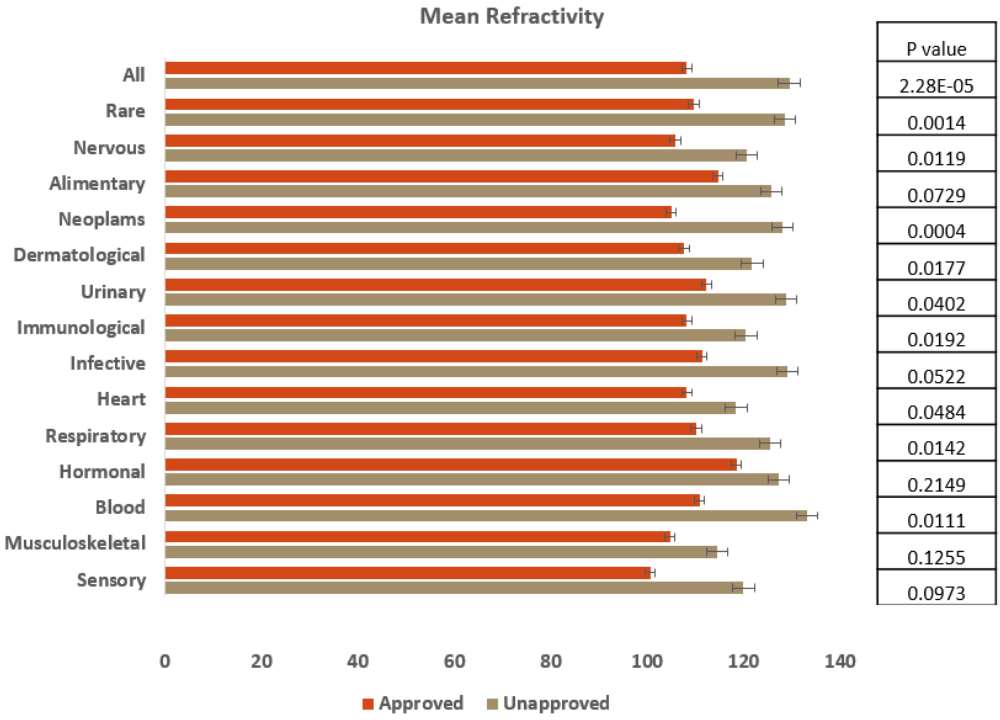


Figure 4.10. Mean refractivity values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student's t-test, p values <0.05).

Figure 4.11 shows the mean polarizability values of approved and unapproved drugs across 14 disease groups. The mean polarizability values significantly differ between approved and unapproved drugs in "Rare Diseases", "Neoplasms", "Dermatological Diseases", "Immunological diseases", "Respiratory Diseases", and "Blood Diseases" groups (Student's t test, p values <0.05).

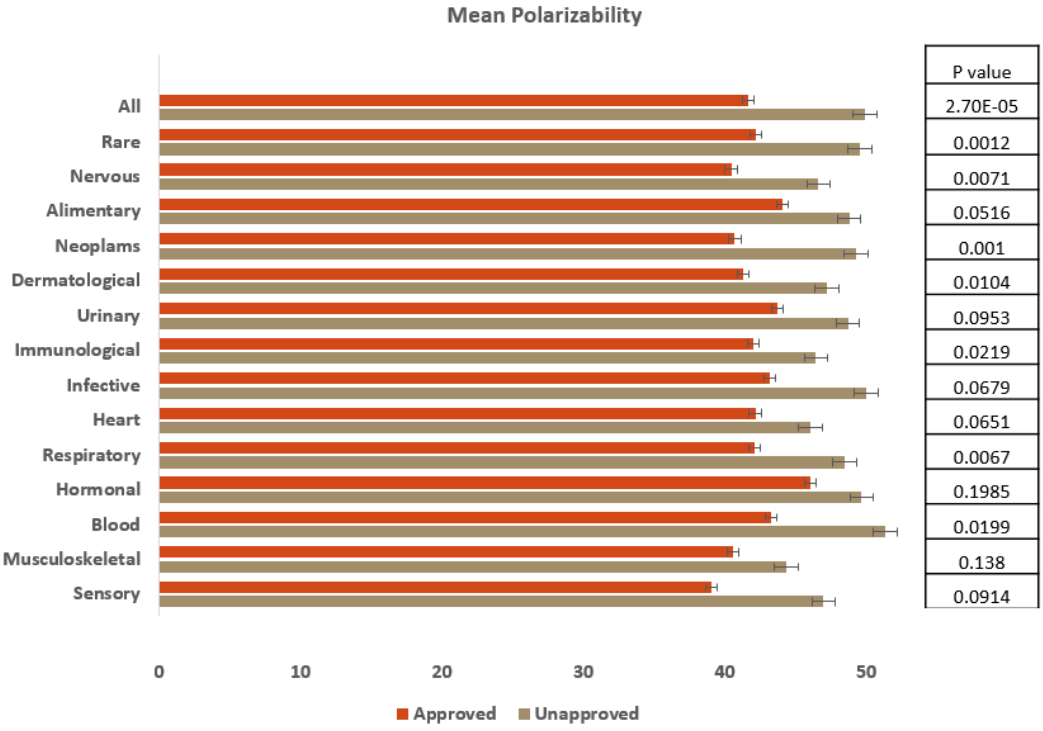


Figure 4.11. Mean polarizability values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student's t-test, p values <0.05).

Figure 4.12 shows the mean number of rotatable bond values of approved and unapproved drugs across 14 disease groups. The mean number of rotatable bond values significantly differ between approved and unapproved drugs in "Nervous Diseases", "Neoplasms" and "Blood Diseases" groups (Student's t-test, p values <0.05).

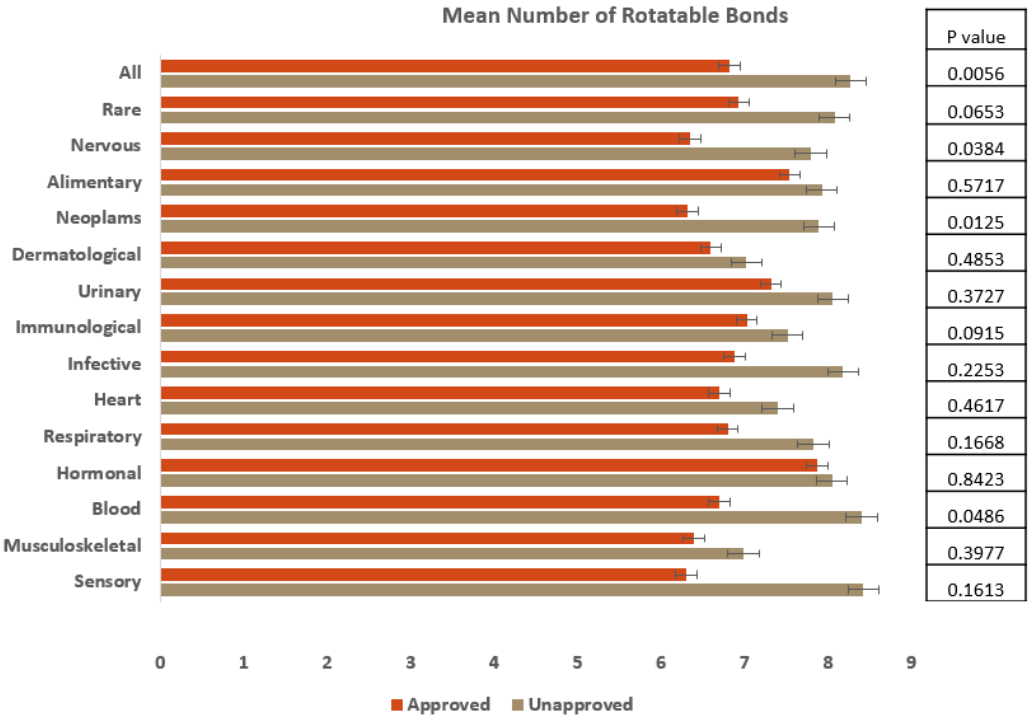


Figure 4.12. Mean number of rotatable bonds values for approved and unapproved drugs among 14 disease groups. Comparing approved and unapproved drugs shows statistically significant differences at the right part for certain disease groups (Student's t-test, p values <0.05).

4.2.2. Prediction Performance Analysis

Both as an ablation study and overall performance evaluation, we constructed models using each type of feature separately (i.e., molecular compound/drug features, physicochemical compound/drug features, clinical trial-related features, and patent-related features), and all of them at once (by concatenating all features under a single feature vector). We trained and validated independent RF models for each disease group via 5-fold cross-validation as described in Methods, section 2.3. At this point, a hyperparameter optimization test was performed on the "all diseases" group prediction model, and the best parameters were determined as; the number of trees: 200; the maximum number of features: the square root of the total number of features; maximum depth: 16; minimum samples leaf: 2 and minimum samples split: 2. The same set of hyperparameter values were used for the rest of the study since hyperparameter tuning process is computationally intensive. All the results reported below are obtained using this hyperparameter value set.

Table 4.3 summarizes the performance results of models constructed using drug-indication pair. As a drug may exist in multiple drug-indication data points in a disease group's dataset (because of being paired with different indications in the same disease group); molecular, physicochemical, and patent features are replicated for the drug of interest, and only clinical trial features remain different. Therefore, this situation led to overoptimistic performance results due to high similarity between training and test samples (i.e., a data leakage), as can be seen in Table 4.3. In order to avoid this situation, we randomly discarded duplicate data points of drugs under each disease group, thereby avoiding bias in terms of data leakage from training to test datasets (Table 4.4 & 4.5).

Table 4.3. The performance of models using drug indication pairs across 14 disease groups. The highest scores for each evaluation metric are shown in bold. Abbreviations are; Acc: accuracy, Accb: balanced accuracy, Pre: precision, Prew: precision weighted, Rec: recall, Recw: recall weighted, F1: F1-score, F1w: F1 weighted, MCC: Matthews Correlation Coefficient, TN: true negative, FP: false positive, FN: false negative, TP: true positive.

Disease	**Acc**	**Accb**	**Pre**	**Prew**	**Rec**	**Recw**	**F1**	**F1w**	**MCC**	**TN**	**FP**	**FN**	**TP**
All	**0.97**	0.88	**0.97**	**0.97**	**0.99**	**0.97**	**0.98**	**0.97**	**0.83**	1,321	1,159	1,774	16,653
Rare	0.92	0.89	0.92	0.92	0.97	0.92	0.94	0.92	0.81	1,125	643	618	3,268
Nervous	0.94	0.76	0.94	0.94	**0.99**	0.94	0.97	0.94	0.67	219	589	499	5,898
Alimentary	0.93	0.79	0.93	0.93	**0.99**	0.93	0.96	0.92	0.71	298	563	302	4,530
Neoplasms	0.91	**0.90**	0.91	0.91	0.96	0.91	0.93	0.91	0.82	1,070	460	543	1,907
Dermato-logical	0.91	0.81	0.92	0.92	**0.99**	0.92	0.95	0.91	0.73	314	488	198	3,048
Urinary	0.92	0.82	0.92	0.92	**0.99**	0.92	0.95	0.92	0.74	371	447	130	3,085
Heart	0.94	0.73	0.94	0.94	**0.99**	0.94	0.97	0.93	0.64	105	411	111	4,261
Immuno-logical	0.92	0.85	0.91	0.91	**0.99**	0.92	0.95	0.91	0.78	356	492	249	2,245
Infective	0.93	0.76	0.93	0.93	**0.99**	0.93	0.96	0.92	0.68	162	420	108	3,658
Respiratory	0.91	0.84	0.91	0.91	0.97	0.91	0.94	0.91	0.75	358	330	211	1,936
Hormonal	0.92	0.86	0.92	0.92	0.98	0.92	0.95	0.92	0.78	364	284	89	1,961
Blood	0.90	0.89	0.89	0.89	0.93	0.90	0.91	0.90	0.79	568	207	160	878
Musculo-skeletal	0.91	0.68	0.91	0.91	**0.99**	0.91	0.95	0.89	0.54	54	186	26	1,457
Sensory	0.91	0.79	0.91	0.91	**0.99**	0.91	0.95	0.91	0.68	121	185	117	1,261

Table 4.4 summarizes our models' drug approval prediction performances using separate and combined feature sets for all datasets (i.e., specific disease groups and a dataset that includes all data points from all disease groups). Upon comparing the models that utilize individual feature types, it is observed that the performance of the clinical trial features is

the highest (with the highest accuracy: 0.76, precision: 0.92, recall: 0.85, F1: 0.80, and MCC: 0.54), followed by the patent features (with the highest accuracy: 0.71, precision: 0.81, recall: 0.77, F1: 0.79, and MCC: 0.37). Molecular and physicochemical features of drugs displayed lower performances. These results indicate that the most important signals for predicting drug approvals come from the clinical trial and patent features. Especially for nervous, alimentary, and dermatological disease groups, patent features perform better than clinical trial features in certain evaluation metrics.

On the other hand, models' performances using physicochemical features are observed to be the lowest among all, though still better than a random prediction, indicating their moderate value for predicting drug approvals. The reason behind this could be that drug development time drug candidates are modified to follow generally accepted rules regarding desired physicochemical feature values such as Lipinski's rule of five (Lipinski et al., 2001). Therefore, most unapproved and approved drugs are more or less consistent with RO5.

Finally, the performances of the models that utilize the combination of all features are observed to be the highest for all of the disease groups, considering most of the evaluation metrics (with the highest accuracy: 0.81, precision: 0.82, recall: 0.96, F1: 0.88, and MCC: 0.62). The precision values of clinical trial features and the recall values of molecular drug features are higher than that of combined features; however, weighted versions of these evaluation metrics for these individual-feature models are still lower than those that utilize combined features. These results highlight the advantage of combining different features to maximize the prediction performance by exploiting their complementarity.

Table 4.4. The performance of our models for disease groups using four different feature sets and their combination (the top-most block corresponds to a model in which data points from all disease groups are utilized together). The highest scores for each evaluation metric and disease group are shown in bold. Abbreviations are; Acc: accuracy, Accb: balanced accuracy, Pre: precision, Prew: precision weighted, Rec: recall, Recw: recall weighted, F1: F1-score, F1w: F1 weighted, MCC: Matthews Correlation Coefficient, TN: true negative, FP: false positive, FN: false negative, TP: true positive.

Disease / Utilized features	**Acc**	**Accb**	**Pre**	**Prew**	**Rec**	**Recw**	**F1**	**F1w**	**MCC**	**TN**	**FP**	**FN**	**TP**
All diseases													
ECFP4	0.71	0.63	0.72	0.70	**0.90**	0.71	0.80	0.68	0.33	417	710	203	1,792
Clinical	0.72	0.74	**0.87**	0.76	0.65	0.72	0.75	0.72	0.46	914	213	692	1,303
Patent	0.69	0.69	0.80	0.71	0.69	0.69	0.74	0.70	0.37	780	347	656	1,339
Physico-chemical	0.65	0.58	0.69	0.63	0.84	0.65	0.76	0.63	0.19	353	774	335	1,660
All	**0.80**	**0.77**	0.82	**0.80**	0.89	**0.80**	**0.85**	**0.80**	**0.56**	734	393	233	1,762
Rare diseases													
ECFP4	0.66	0.64	0.67	0.66	0.80	0.66	0.73	0.65	0.29	395	440	224	909
Clinical	0.70	0.71	**0.82**	0.73	0.61	0.70	0.70	0.70	0.42	675	160	435	698
Patent	0.67	0.66	0.72	0.67	0.68	0.67	0.70	0.67	0.32	541	294	400	733
Physico-chemical	0.58	0.58	0.65	0.59	0.58	0.58	0.61	0.58	0.16	236	259	303	534
All	**0.77**	**0.76**	0.78	**0.77**	**0.84**	**0.77**	**0.81**	**0.77**	**0.52**	578	257	303	830
Nervous diseases													
ECFP4	0.70	0.58	0.71	0.68	**0.93**	0.70	0.81	0.65	0.23	128	427	90	1,042
Clinical	0.71	**0.73**	**0.86**	0.76	0.67	0.71	0.75	0.72	0.43	432	123	379	753
Patent	0.71	0.68	0.80	0.72	0.77	0.71	0.78	0.71	0.36	319	236	298	834

Table 4.4 (cont.)

Physico-chemical	0.59	0.53	0.69	0.59	0.72	0.59	0.70	0.58	0.06	125	430	159	973
All	**0.79**	0.71	0.79	**0.78**	**0.93**	**0.79**	**0.85**	**0.77**	**0.49**	263	292	129	1,003
Alimentary diseases													
ECFP4	0.70	0.59	0.70	0.69	**0.94**	0.70	0.80	0.65	0.26	142	413	82	993
Clinical	0.69	0.72	**0.86**	0.75	0.63	0.69	0.73	0.69	0.41	446	109	391	684
Patent	0.69	0.67	0.80	0.71	0.71	0.69	0.74	0.69	0.34	351	204	351	724
Physico-chemical	0.60	0.56	0.70	0.60	0.69	0.60	0.69	0.60	0.11	243	312	378	697
All	**0.80**	**0.75**	0.81	**0.80**	0.92	**0.80**	**0.86**	**0.79**	**0.54**	298	257	130	945
Neoplasms													
ECFP4	0.64	0.63	0.65	0.64	0.76	0.64	0.70	0.63	0.27	351	353	217	652
Clinical	0.72	0.71	0.72	0.72	0.81	0.72	0.76	0.71	0.43	427	277	167	702
Patent	0.66	0.66	0.70	0.66	0.66	0.66	0.68	0.66	0.31	467	237	312	557
Physico-chemical	0.59	0.59	0.64	0.59	0.59	0.59	0.61	0.59	0.18	394	310	381	488
All	**0.78**	**0.77**	**0.78**	**0.78**	**0.83**	**0.78**	**0.80**	**0.78**	**0.55**	494	210	217	652
Dermatological diseases													
ECFP4	0.69	0.61	0.70	0.67	0.88	0.69	0.78	0.65	0.26	162	359	93	837
Clinical	0.67	0.70	**0.84**	0.73	0.61	0.67	0.70	0.68	0.39	404	117	366	564
Patent	0.69	0.66	0.77	0.69	0.72	0.68	0.74	0.68	0.32	308	213	296	634
Physico-chemical	0.61	0.59	0.70	0.62	0.66	0.61	0.68	0.61	0.17	245	276	328	602
All	**0.78**	**0.73**	0.79	**0.78**	**0.90**	**0.78**	**0.84**	**0.77**	**0.51**	277	244	136	794

Table 4.4 (cont.)

Urinary tract diseases													
ECFP4	0.69	0.61	0.71	0.67	0.88	0.69	0.78	0.66	0.26	151	355	82	832
Clinical	0.72	0.75	**0.88**	0.77	0.66	0.72	0.75	0.73	0.48	432	74	323	591
Patent	0.68	0.66	0.76	0.69	0.73	0.68	0.73	0.68	0.31	280	226	273	641
Physico-chemical	0.61	0.56	0.68	0.60	0.72	0.61	0.70	0.60	0.12	139	367	167	747
All	**0.80**	**0.76**	0.80	**0.80**	**0.91**	**0.80**	**0.85**	**0.79**	**0.55**	291	215	115	799
Heart diseases													
ECFP4	0.73	0.58	0.74	0.71	**0.96**	0.73	0.84	0.68	0.24	64	301	48	835
Clinical	0.69	0.74	**0.92**	0.79	0.61	0.69	0.73	0.70	0.44	315	50	340	543
Patent	0.71	0.67	0.81	0.72	0.77	0.71	0.79	0.71	0.33	211	154	223	660
Physico-chemical	0.61	0.53	0.73	0.62	0.71	0.61	0.72	0.61	0.07	135	230	264	619
All	**0.81**	**0.73**	0.82	**0.82**	0.95	**0.82**	**0.88**	**0.80**	**0.53**	168	197	83	800
Immunological diseases													
ECFP4	0.69	0.62	0.70	0.68	0.89	0.69	0.78	0.65	0.29	159	336	93	744
Clinical	0.68	0.70	**0.84**	0.73	0.60	0.68	0.70	0.68	0.40	317	178	279	558
Patent	0.69	0.67	0.76	0.69	0.73	0.69	0.75	0.69	0.34	294	201	249	588
Physico-chemical	0.61	0.59	0.70	0.62	0.67	0.61	0.69	0.62	0.19	236	259	303	534
All	**0.79**	**0.74**	0.78	**0.79**	**0.92**	**0.79**	**0.84**	**0.77**	**0.53**	268	227	128	709
Infective diseases													
ECFP4	0.72	0.57	0.73	0.70	0.95	0.72	0.83	0.66	0.23	88	309	46	860
Clinical	0.68	0.71	**0.88**	0.76	0.62	0.68	0.73	0.69	0.40	315	82	330	576

Table 4.4 (cont.)

Patent	0.68	0.66	0.80	0.70	0.72	0.68	0.76	0.69	0.30	229	168	273	633
Physico-chemical	0.60	0.53	0.71	0.60	0.71	0.60	0.71	0.60	0.06	157	240	312	594
All	**0.79**	**0.70**	0.80	**0.79**	**0.96**	**0.79**	**0.86**	**0.77**	**0.48**	182	215	88	818
Respiratory diseases													
ECFP4	0.71	0.67	0.71	0.71	**0.87**	0.71	0.78	0.70	0.38	225	245	106	593
Clinical	0.66	0.67	0.78	0.70	0.64	0.66	0.69	0.66	0.35	260	210	171	528
Patent	0.66	0.66	0.73	0.66	0.69	0.66	0.71	0.66	0.31	283	187	227	472
Physico-chemical	0.61	0.60	0.69	0.62	0.64	0.61	0.66	0.61	0.20	261	209	259	440
All	**0.77**	**0.75**	**0.78**	**0.77**	0.86	**0.77**	**0.82**	**0.77**	**0.53**	288	182	127	572
Hormonal diseases													
ECFP4	0.71	0.65	0.71	0.72	0.91	0.71	0.80	0.69	0.35	152	222	68	563
Clinical	0.65	0.69	**0.86**	0.73	0.53	0.65	0.65	0.65	0.37	250	124	205	426
Patent	0.68	0.66	0.76	0.69	0.70	0.68	0.73	0.68	0.33	225	149	209	422
Physico-chemical	0.61	0.59	0.70	0.62	0.66	0.61	0.67	0.61	0.18	110	264	116	515
All	**0.81**	**0.78**	0.81	**0.81**	**0.91**	**0.81**	**0.86**	**0.81**	**0.59**	220	154	66	565
Blood diseases													
ECFP4	0.65	0.64	0.65	0.65	0.72	0.65	0.69	0.64	0.29	224	179	132	331
Clinical	0.77	0.76	0.75	0.77	**0.85**	0.77	0.80	0.76	0.54	267	136	66	397
Patent	0.63	0.63	0.66	0.63	0.63	0.63	0.64	0.63	0.26	252	151	178	285
Physico-chemical	0.61	0.60	0.63	0.61	0.63	0.61	0.63	0.60	0.21	231	172	185	278

Table 4.4 (cont.)

All	**0.81**	**0.81**	**0.81**	**0.81**	0.84	**0.81**	**0.83**	**0.81**	**0.62**	277	126	77	386
Musculoskeletal diseases													
ECFP4	0.74	0.58	0.75	0.72	0.96	0.74	0.84	0.69	0.25	29	173	23	493
Clinical	0.65	0.68	**0.87**	0.75	0.60	0.65	0.71	0.67	0.33	148	54	212	304
Patent	0.68	0.63	0.80	0.70	0.75	0.68	0.77	0.69	0.26	103	99	138	378
Physico-chemical	0.66	0.52	0.73	0.83	0.66	0.61	0.77	0.63	0.05	38	164	60	456
All	**0.80**	**0.67**	0.80	**0.80**	**0.96**	**0.80**	**0.87**	**0.77**	**0.45**	75	127	26	490
Sensory diseases													
ECFP4	0.75	0.65	0.75	0.75	0.94	0.75	0.84	0.72	0.39	84	134	34	417
Clinical	0.72	0.67	0.78	0.72	0.81	0.72	0.79	0.72	0.35	111	107	122	329
Patent	0.65	0.61	0.74	0.66	0.73	0.65	0.74	0.65	0.22	102	116	114	337
Physico-chemical	0.64	0.60	0.74	0.65	0.73	0.64	0.73	0.64	0.20	105	113	120	331
All	**0.79**	**0.71**	**0.79**	**0.80**	**0.95**	**0.79**	**0.86**	**0.77**	**0.50**	107	111	64	387

Since utilizing all features together yields the best performance in each disease group model, we selected these models (i.e., all /combined features) with 149 dimensions to continue our analyses with the total size. Table 4.5 summarizes all/combined features-based performance results for 14 disease groups. The performances of the models for the majority of disease groups are observed to be satisfactory and also similar to each other; varying in the ranges of accuracy: 0.67-0.81, precision: 0.77-0.82, recall: 0.77-0.96, F1-score: 0.77-0.88 and MCC: 0.45-0.62, indicating the effectiveness of the drug approval prediction approach proposed in this study. As mentioned in Section 3.2.5., MCC around 0.5 is generally evaluated as a successful model because of the equation for MCC (Eq. 5).

Table 4.5. The performance of finalized models for 14 disease groups. The highest scores for each evaluation metric are shown in bold. Abbreviations are; Acc: accuracy, Accb: balanced accuracy, Pre: precision, Prew: precision weighted, Rec: recall, Recw: recall weighted, F1: F1-score, F1w: F1 weighted, MCC: Matthews Correlation Coefficient, TN: true negative, FP: false positive, FN: false negative, TP: true positive.

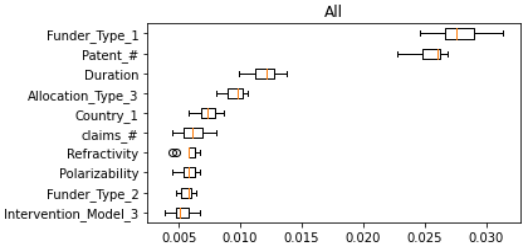
Disease	Acc	Accb	Pre	Prew	Rec	Recw	F1	F1w	MCC	TN	FP	FN	TP
All	0.80	0.77	**0.82**	0.80	0.89	0.80	0.85	0.80	0.56	734	393	233	1,762
Rare	0.77	0.76	0.78	0.77	0.84	0.77	0.81	0.77	0.52	578	257	303	830
Nervous	0.79	0.71	0.79	0.78	0.93	0.79	0.85	0.77	0.49	263	292	129	1,003
Alimentary	0.80	0.75	0.81	0.80	0.92	0.80	0.86	0.79	0.54	298	257	130	945
Neoplasms	0.78	0.77	0.78	0.78	0.83	0.78	0.80	0.78	0.55	494	210	217	652
Dermato-logical	0.78	0.73	0.79	0.78	0.90	0.78	0.84	0.77	0.51	277	244	136	794
Urinary	0.80	0.76	0.80	0.80	0.91	0.80	0.85	0.79	0.55	291	215	115	799
Heart	**0.81**	0.73	**0.82**	**0.82**	0.95	**0.82**	**0.88**	0.80	0.53	168	197	83	800
Immuno-logical	0.79	0.74	0.78	0.79	0.92	0.79	0.84	0.77	0.53	268	227	128	709
Infective	0.79	0.70	0.80	0.79	0.96	0.79	0.86	0.77	0.48	182	215	88	818
Respiratory	0.77	0.75	0.78	0.77	0.86	0.77	0.82	0.77	0.53	288	182	127	572
Hormonal	**0.81**	0.78	0.81	0.81	0.91	0.81	0.86	**0.81**	0.59	220	154	66	565
Blood	**0.81**	**0.81**	0.81	0.81	0.84	0.81	0.83	**0.81**	**0.62**	277	126	77	386
Musculo-skeletal	0.80	0.67	0.80	0.80	**0.96**	0.80	0.87	0.77	0.45	75	127	26	490
Sensory	0.79	0.71	0.79	0.80	0.95	0.79	0.86	0.77	0.50	107	111	64	387

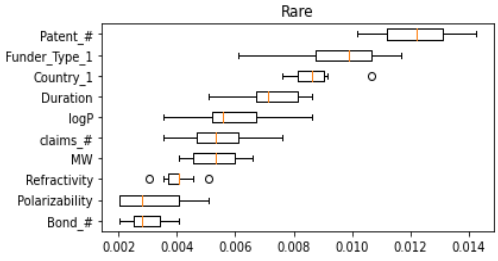
We observed that some of the specific disease group models performed slightly better than the combined “all diseases” model (e.g., “blood and lymph diseases”); however, others performed relatively poorly (e.g., “sensory diseases”). Target proteins and the

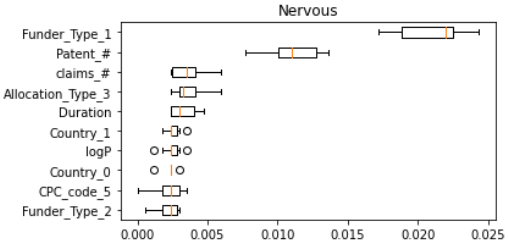
characteristics of affected tissues greatly vary in different disease groups, which causes a high amount of heterogeneity in the data. We believe this is the main reason behind observing the performance differences among disease groups. For example, "blood diseases" can be considered to have lower heterogeneity than "rare diseases", which may explain their performance difference. Another possible reason behind variable performances may be the size of the datasets, together with the state of balance between the sizes of tasks/classes (i.e., approved and unapproved drugs), where larger datasets with better coverage over the sample space generally result in better performances, provided that all other factors remain constant. Relatively lower performances of "sensory", "musculoskeletal," and "infective" diseases can be given as examples to this point.

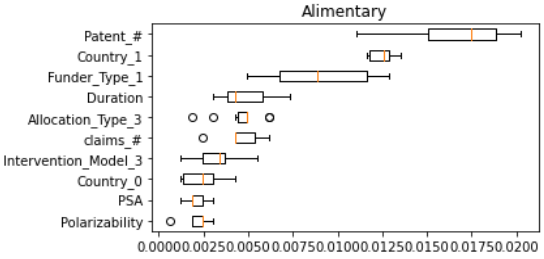
To observe the most informative features for accurately predicting drug approvals, we calculated both the permutation importance and mean decrease impurity (MDI) values for each feature. Figure 4.13 shows the top ten important features for all disease groups, determined using permutation importance. Here, we observe that the number of related patents and the funder type of clinical trials is among the most important features for all disease groups (Figure 4.13). Additionally, number of claims, allocation and location (by country) of clinical trials, duration of patent process, lipophilicity (logP), molecular weight, PSA, polarizability and refractivity of the drugs are frequently ranked among the top features. The fact that three patent-related features are among the most important ones demonstrates that patents have critical roles in the approval of drugs. About the most important patent feature (i.e., "number of related patents"), when a drug candidate is highly promising, numerous patent applications have been made before the approval date of that drug. However, this feature cannot be evaluated in a causality relationship because, most probably, getting high number of patents beforehand do not affect the approval decision. However, when the developers see high chance of approval, they tend to protect the candidate with more patents.

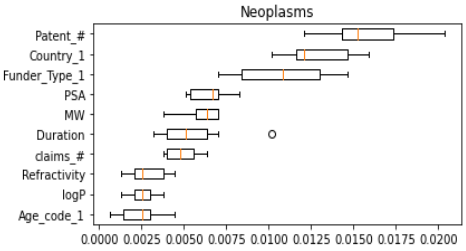
Additionally, multiple clinical trial-related features such as the funder type, allocation type, location by country, and intervention model are among the most important features, indicating the importance of clinical trial specifications in drug approvals. As a physicochemical feature, the lipophilicity of a drug is frequently observed as important, which is understandable because lipophilicity is considered one of the most important drug-like physical properties (Leeson et al., 2007). For the MDI-based feature importance results, please see Table 4.27(Appendix B). In general, the most important features obtained by the MDI method seem to be largely consistent with the results of the permutation importance analysis.

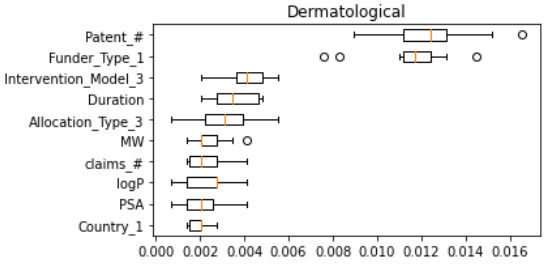

All
Funder_Type_1
Patent_#
Duration
Allocation_Type_3
Country_1
claims_#
Refractivity
Polarizability
Funder_Type_2
Intervention_Model_3
0.005
0.010
0.015
0.020
0.025
0.030

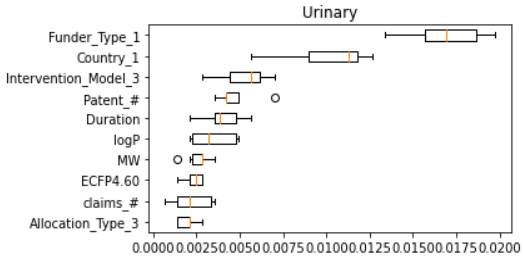

Rare
Patent_#
Funder_Type_1
Country_1
Duration
logP
claims_#
MW
Refractivity
Polarizability
Bond_#
0.002
0.004
0.006
0.008
0.010
0.012
0.014

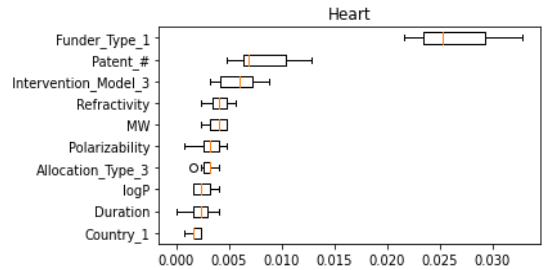

Nervous
Funder_Type_1
Patent_#
claims_#
Allocation_Type_3
Duration
Country_1
logP
Country_0
CPC_code_5
Funder_Type_2
0.000
0.005
0.010
0.015
0.020
0.025

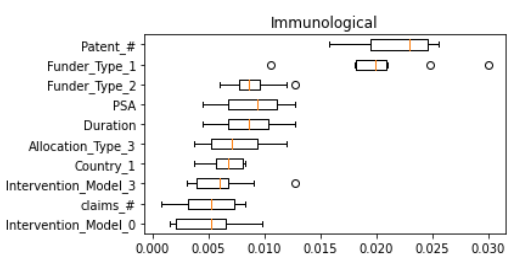

Alimentary
Patent_#
Country_1
Funder_Type_1
Duration
Allocation_Type_3
claims_#
Intervention_Model_3
Country_0
PSA
Polarizability
0.00000.00250.00500.00750.01000.01250.01500.01750.0200

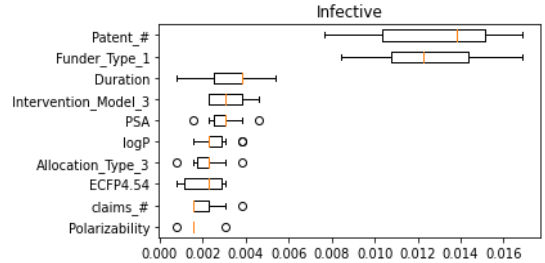

Neoplasms
Patent_#
Country_1
Funder_Type_1
PSA
MW
Duration
claims_#
Refractivity
logP
Age_code_1
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200


Dermatological
Patent_#
Funder_Type_1
Intervention_Model_3
Duration
Allocation_Type_3
MW
claims_#
logP
PSA
Country_1
0.000 0.002 0.004 0.006 0.008 0.010 0.012 0.014 0.016


Urinary
Funder_Type_1
Country_1
Intervention_Model_3
Patent_#
Duration
logP
MW
ECFP4.60
claims_#
Allocation_Type_3
0.0000 0.0025 0.0050 0.0075 0.0100 0.0125 0.0150 0.0175 0.0200


Heart
Funder_Type_1
Patent_#
Intervention_Model_3
Refractivity
MW
Polarizability
Allocation_Type_3
logP
Duration
Country_1
0.000 0.005 0.010 0.015 0.020 0.025 0.030


Immunological
Patent_#
Funder_Type_1
Funder_Type_2
PSA
Duration
Allocation_Type_3
Country_1
Intervention_Model_3
claims_#
Intervention_Model_0
0.000 0.005 0.010 0.015 0.020 0.025 0.030


Infective
Patent_#
Funder_Type_1
Duration
Intervention_Model_3
PSA
logP
Allocation_Type_3
ECFP4.54
claims_#
Polarizability
0.000 0.002 0.004 0.006 0.008 0.010 0.012 0.014 0.016

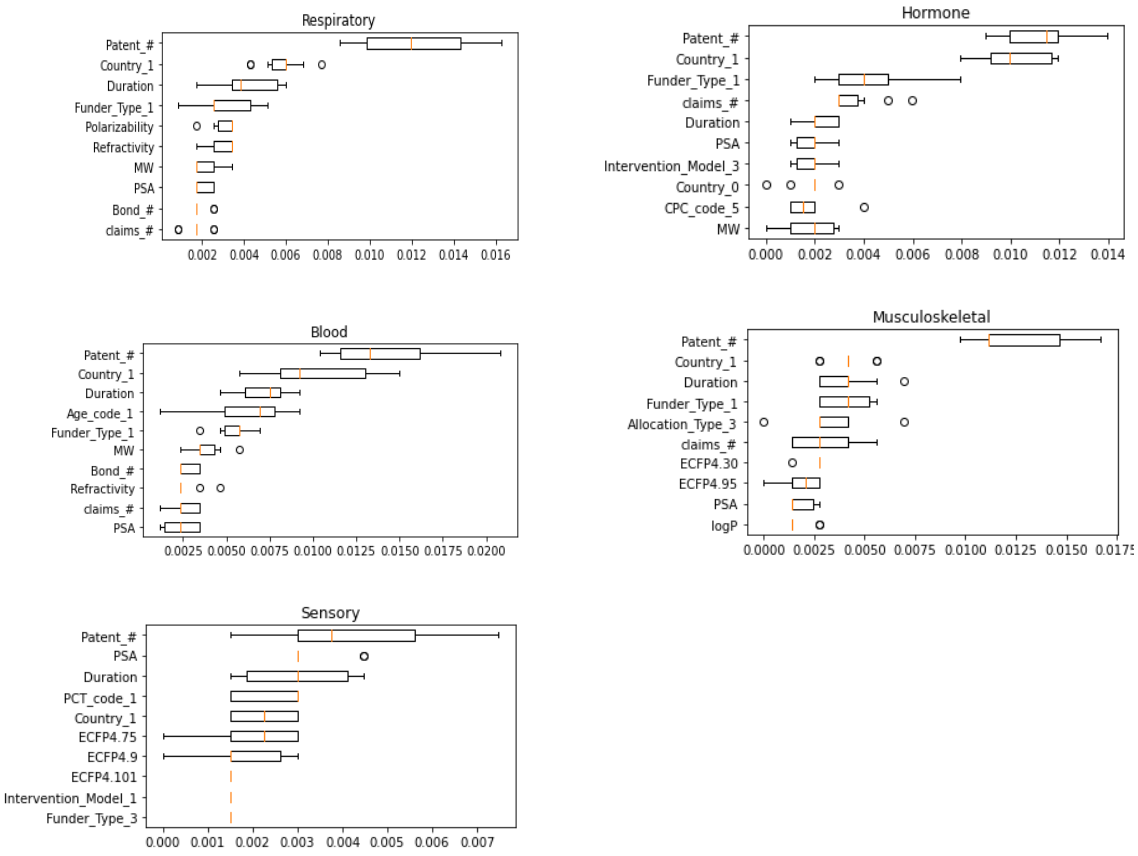


Figure 4.13. The top ten important features for accurately predicting drug approvals in selected disease groups, determined using the permutation feature importance (PFI) method. Abbreviations; Claim #: number of claims, Patent #: number of related patents, Bond #: number of rotatable bonds, PSA: polar surface area, logP: lipophilicity of the molecule. Each categorical feature is represented by a numerical code after one hot encoding; Funder_Type_1: Others (in clinicaltirals.gov), Funder_Type_2: Industry, Country_1: US, Country_0: Others, Intervention_Model_3: Parallel Assignment, Allocation_Type_3: Randomized, CPC_code_5: Others, Age_code_1: Adults & Older Adults.

4.2.3. Imputation Performance Analysis

To investigate the high degrees of missing features in our dataset (i.e., 15% to 51%), we calculated the performance change caused by the imputation process by artificially adding varying degrees of missing features into the dataset (i.e., 0%, 25%, 50%, and 75%). In this analysis, we used a subset of our original dataset, composed of 1486 drug-indication pairs without any missing features. We separated 10% of this dataset (i.e., 149 data points) upfront as our test set. We used the remaining 1337 as our training dataset. We trained four models, each differing in terms of the amount of artificially created missing features (i.e., 0% - meaning the complete dataset is directly used- 25%, 50%, and 75% missing/imputed features). For example, we randomly selected 25% of all features among all drug-indication pairs in the training dataset for our second model and changed their original value with imputation.

Following the training and the calculation of performance values, we observed that performances are decreasing with the increasing degrees of missing features, which was expected (Table 4.6). The performance decrease of the model that use data with 25% missing features, compared to the one that utilizes the complete dataset (i.e., 0% missing), were 11% and 45% according to weighted F1-score and MCC, respectively. After that point, the performance decrease of the models (that use 50% and 75% missing data) is slow, mainly due to getting closer to random prediction. These results indicate that DrugApp can benefit from data filling procedures that are more sophisticated than mean/median imputing. It is also important to note that, in this analysis, we tested marginal cases (e.g., 75% of all features in the whole dataset is missing) to observe the performance difference. However, in our original dataset, the minimum and maximum amounts of missing data for a single feature are 15% and 51%, respectively.

Table 4.6. The performance results of the empirical analysis on missing value imputation. Performances were analyzed by adding varying degrees of missing features into the dataset. Abbreviations are; Acc: accuracy, Accb: balanced accuracy, Pre: precision, Prew: precision weighted, Rec: recall, Recw: recall weighted, F1: F1-score, F1w: F1 weighted, MCC: Matthews Correlation Coefficient, TN: true negative, FP: false positive, FN: false negative, TP: true positive.

	Acc	**Accb**	**Pre**	**Prew**	**Rec**	**Recw**	**F1**	**F1w**	**MCC**	**TN**	**FP**	**FN**	**TP**
Complete data (0% missing)	0.79	0.69	0.80	0.78	0.93	0.79	0.86	0.77	0.46	20	24	7	98
25% missing/ imputed	0.73	0.58	0.74	0.71	0.95	0.73	0.83	0.68	0.25	9	35	5	100
50% missing/ imputed	0.71	0.54	0.72	0.70	0.98	0.72	0.83	0.63	0.17	4	40	2	103
75% missing/ imputed	0.71	0.52	0.71	0.68	0.98	0.71	0.83	0.62	0.12	3	41	2	103

4.2.4. Temporal Analysis

In order to examine the performance of our method in prospectively predicting the outcome of drug development processes with continuing clinical trial procedures, we applied a temporal analysis based on the approval date of the drugs. For this purpose, we split our dataset into six-time intervals, as shown in Table 4.7, which also displayed the statistics of each temporal dataset. The test datasets comprise data left outside of the training data time frames. In order to prevent using future trial information for predicting the results of earlier trials, testing was selected only from future time frames. In this analysis, we did not split data into specific disease groups and used all data points under a single model (i.e., the "all diseases" group) due to the reduced size of disease group-specific temporal datasets.

Table 4.7. Statistics of the temporal analysis datasets. The sample sizes of regulatorily approved and unapproved drug-indication pairs in six different time frames. The time frame for the training set comprises the years before the threshold year, and the testing set consists of the years after the threshold year.

Threshold year (to separate train-test sets)	# of approved (training set)	# of unapproved (training set)	# of approved (test set)	# of unapproved (test set)
2008	752	198	1,781	722
2010	952	307	1,711	635
2012	1,127	397	1,636	565
2014	1,274	497	1,532	467
2016	1,441	610	1,388	362
2018	1,649	726	1,184	221

Table 4.8 summarizes the performance results of the temporal analysis. We observe a trend of increasing performance from the threshold year 2008 (accuracy: 0.57-0.75, precision: 0.74-0.78, recall: 0.75-0.99, F1: 0.69-0.85, and MCC: 0.32) to 2018 (accuracy: 0.81-0.91, precision: 0.90-0.94, recall: 0.91-0.95, F1: 0.90-0.94, and MCC: 0.64). This result is expected since the size of the training datasets is getting higher with increasing threshold years, and larger datasets usually yield a better representation of sample spaces. The performances of temporal models (especially the last two of them) are observed to be not only satisfactory but also higher compared to the performance of the previous "all disease groups" model on which the performance was measured via random splitting-based cross-validation (Accuracy: 0.79-0.80, Precision: 0.80-0.82, Recall: 0.80-0.89, F1: 0.80-0.85, MCC: 0.56). This indicates that our approach is promising in a real-world setting where future drug approvals are to be predicted.

Table 4.8. The temporal analysis performance results. Performance scores of models from six different time frames for the "all diseases" group. The highest scores for each evaluation metric are shown in bold. Abbreviations are; Acc: accuracy, Accb: balanced accuracy, Pre: precision, Prew: precision weighted, Rec: recall, Recw: recall weighted, F1: F1-score, F1w: F1 weighted, MCC: Matthews Correlation Coefficient, TN: true negative, FP: false positive, FN: false negative, TP: true positive.

Training data time frames	Acc	Accb	Pre	Prew	Rec	Recw	F1	F1w	MCC	TN	FP	FN	TP
≤ 2008	0.75	0.57	0.74	0.78	0.99	0.75	0.85	0.69	0.32	131	591	20	1,761
≤ 2010	0.81	0.67	0.80	0.81	0.97	0.81	0.88	0.78	0.46	229	406	46	1,665
≤ 2012	0.84	0.72	0.84	0.83	0.96	0.84	0.90	0.82	0.53	268	297	65	1,571
≤ 2014	0.86	0.76	0.88	0.86	0.95	0.86	0.91	0.85	0.59	263	204	71	1,461
≤ 2016	0.88	0.78	0.90	0.88	0.95	0.88	0.93	0.88	0.61	219	143	65	1,323
≤ 2018	**0.91**	**0.81**	**0.94**	**0.90**	**0.95**	**0.91**	**0.94**	**0.90**	**0.64**	150	71	60	1,124

For the important features of temporal analysis, please see Figure 4.14 (permutation importance) and Table 4.28 (MDI-based feature importance) in Appendix B. The results obtained by MDI-based and PFI methods seem primarily consistent. In time, it was observed that while ECFP4, age (adults& older adults), and CPC disappeared from the top ten most important features, number of claims, other countries in terms of location of trials and industry as funder type appeared among the most important features for predicting drug approvals. However, despite not being present in the top ten most important features determined by PFI method, claim numbers were among the most important features for the 0-2008 time interval, revealed by the MDI-based method. For ECFP4 and CPC features, the fact that they appeared in the top most important features for the 0-2008 time interval might not be meaningful because the 0-2008 model performed poorly due to the low sample size.

Additionally, we observed that the USA was among the most important features for all time frames; however, in recent years, other countries emerged as important for predicting drug approvals. WHO report (2022) reveals that the USA had the highest number of trials; however, after 2010, Europe, and after 2016, the Western Pacific region became the region with the highest number of trials being registered, thereby being consistent with our results. For the age category, adults & older adults disappeared from being among the most important features in future time frames. FDA (2020) report showed that the difference between the number of participants younger than 65 and the number of

participants older than (or equal to) 65 did not seem to be significantly different between the years 2015 and 2018, consistent with our results. Finally, for the funder type, the industry emerged as an important feature in future time frames. GlobalData (2017) revealed that the relative contribution of industry sponsors, when compared to non-industry sponsors, increased from 63% to 72% (and the non-industry sponsors decreased from 37% to 28%), starting from the year 2012 in Australia, which is also consistent with our results.

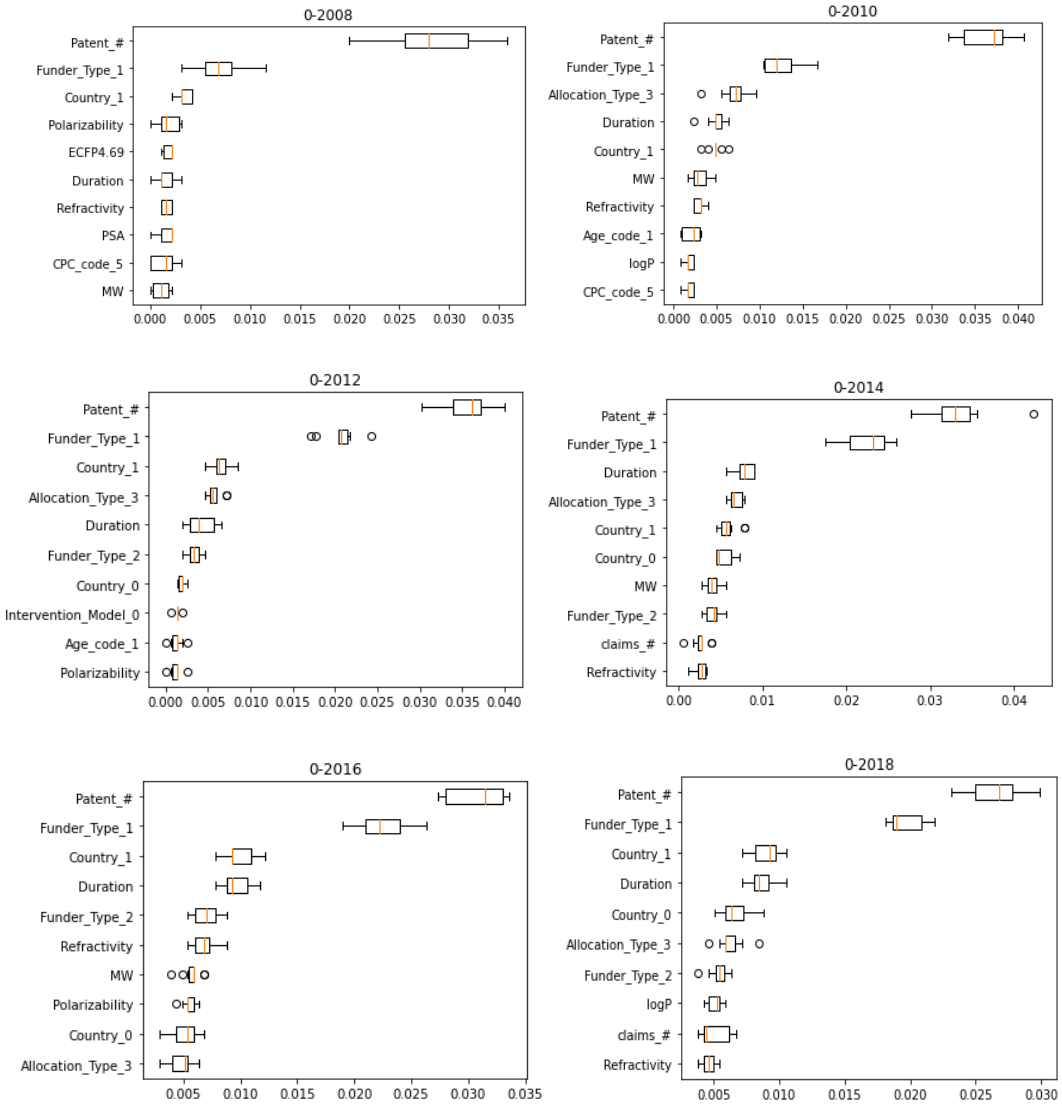


Figure 4.14. The top ten important features for predicting drug approvals in the temporal analysis, calculated using permutation importance. Abbreviations; Claim #: number of claims, Patent #: number of related patents, Bond #: number of rotatable bonds, PSA: polar surface area, logP: lipophilicity of the molecule. Each categorical feature is

represented by a numerical code after one hot encoding; Funder_Type_1: Others (in clinicaltirals.gov), Funder_Type_2: Industry, Country_1: US, Country_0: Others, Intervention_Model_0: Others, Allocation_Type_3: Randomized, CPC_code_5: Others, Age_code_1: Adults & Older Adults.

4.2.5. Performance Comparison with Other Methods

We performed a performance comparison with a baseline method to further investigate DrugApp and justify the algorithm choice of random forest. For this, we employed logistic regression (LR) since it is a widely known and used algorithm in drug discovery and development (Tolles & Meurer, 2016). We used the same datasets for all disease groups and calculated the scores based on a 10-fold cross-validation analysis to make results comparable to the ones provided in Table 4.5. The performances of the LR models varied in the ranges of accuracy: 0.63-0.72, precision: 0.67-0.83, recall: 0.65-0.75, F1-score: 0.65-0.78 and MCC: 0.25-0.41 (Table 4.9), in comparison to random forest-based DrugApp models within the ranges of accuracy: 0.67-0.81, precision: 0.77-0.82, recall: 0.77-0.96, F1-score: 0.77-0.88 and MCC: 0.45-0.62. A one-to-one comparison for each disease group between RF and LR models revealed that the RF models display better performances for all disease groups (without any intersections in scores at one standard deviation of the mean), justifying the choice of using the random forest algorithm over the baseline.

Table 4.9. The performance of finalized the baseline prediction models using logistic regression for 14 disease groups. The highest scores for each evaluation metric are shown in bold. Abbreviations are; Acc: accuracy, Accb: balanced accuracy, Pre: precision, Prew: precision weighted, Rec: recall, Recw: recall weighted, F1: F1-score, F1w: F1 weighted, MCC: Matthews Correlation Coefficient, TN: true negative, FP: false positive, FN: false negative, TP: true positive.

Disease group	**Acc**	**Accb**	**Pre**	**Prew**	**Rec**	**Recw**	**F1**	**F1w**	**MCC**	**TN**	**FP**	**FN**	**TP**
All diseases	0.70	**0.70**	0.80	0.72	0.71	0.70	0.75	0.71	**0.38**	779	348	596	1,399
Rare	0.68	0.68	0.74	0.69	0.69	0.68	0.71	0.68	0.35	540	295	378	755
Nervous	0.66	0.65	0.79	0.69	0.68	0.66	0.73	0.67	0.29	340	215	366	766
Alimentary	0.66	0.66	0.79	0.69	0.66	0.66	0.72	0.67	0.29	390	165	402	390
Neoplasms	0.68	0.68	0.72	0.68	0.68	0.68	0.70	0.68	0.36	480	224	305	564
Dermatological	0.67	0.67	0.79	0.70	0.67	0.67	0.72	0.68	0.34	352	169	346	584
Urinary	0.66	0.65	0.77	0.68	0.68	0.66	0.72	0.67	0.30	313	193	278	636

Table 4.9 (cont.)													
Heart	0.68	0.67	**0.83**	0.72	0.69	0.68	0.75	0.69	0.31	231	134	283	600
Immunological	0.65	0.65	0.76	0.67	0.65	0.65	0.70	0.65	0.29	317	178	312	525
Infective	0.66	0.65	0.81	0.70	0.67	0.66	0.73	0.67	0.29	313	193	278	636
Respiratory	0.66	0.65	0.74	0.67	0.67	0.66	0.70	0.66	0.29	279	191	227	472
Hormonal	0.68	0.67	0.77	0.69	0.68	0.68	0.72	0.68	0.34	247	127	185	446
Blood	0.70	**0.70**	0.72	0.71	0.72	0.70	0.72	0.70	0.41	273	130	142	321
Musculoskeletal	0.65	0.63	0.81	0.70	0.68	0.65	0.74	0.67	0.25	114	88	178	338
Sensory	**0.72**	**0.70**	0.81	**0.73**	**0.75**	**0.72**	**0.78**	**0.72**	**0.38**	130	88	144	307

Moreover, in order to provide a comparison against the state-of-the-art, we analyzed the performance of our model on the dataset of a recently proposed method, HINT, which was trained for clinical trial outcome prediction under three different predictions tasks, namely the prediction of meeting primary endpoints for Phase I, II and III trials (Fu et al., 2022). For this analysis, we employed the dataset provided in the HINT study. Data points from the HINT training and test datasets were intersected with our dataset using SMILES notations of drugs/compounds, and the intersection set is used as our train and test datasets (statistics are shown in Table 4.10). Data points that contain drug combinations were ignored. Outcome labels from the HINT study were directly used, initially obtained by manual curation in the respective study (1 and 0 mean the primary endpoints were met or not, respectively) (Fu et al., 2022). An example of primary endpoints is the percentage of hypertension patients with controlled blood pressure (e.g., cytosolic blood pressure lower than 140 mm Hg) for an antihypertensive drug trial.

As reported in the respective study, the HINT models achieved 0.66, 0.62, and 0.85 F1 scores for predicting the meeting of endpoints for phase I, II, and III clinical trials, respectively; whereas, DrugApp displayed 0.73, 0.71, and 0.82 F1-scores for the same tasks (Table 4.10). The results show that DrugApp and HINT have roughly comparable performances. It is important to note that we could not use the same datasets provided in the HINT study due to differences in dataset preparation and featurization steps between the two studies (e.g., we manually curated patent features for the data points in our original dataset. Also it was not possible to redo this step for the HINT dataset, due to limited sources). Therefore, DrugApp could utilize 74% and 70% of the original HINT training and test datasets, respectively (Table 4.10). Additionally, HINT is centered around clinical trials, whereas DrugApp is based on drug candidate compounds (i.e., the aim of DrugApp is directly predicting the approval of the drug candidate for the given indication). Consequently, it was not possible to obtain an intersection set over shared clinical trials (as we formed our approved drug-indication pairs dataset by using clinical trials that are

transitioned in phase IV, whereas in HINT, authors utilized trials at phases I, II, and III). Similar limitations exist in the quantitative comparison of computational methods with not-perfectly overlapping aims/modeling approaches/datasets.

Performance comparison results indicate that both the artificial learning and featurization approaches utilized in DrugApp effectively model the drug development process. Furthermore, DrugApp can generalize information, as it performs quite well when trained and tested on a prediction task that it was not originally designed for (i.e., predicting the meeting of primary endpoints in clinical trials).

Table 4.10. The performance comparison between DrugApp and HINT for three different prediction tasks, namely the prediction of meeting primary endpoints for Phase I, II, and III trials. The best performance for each analysis is shown in bold font.

	HINT training dataset size (# of unique SMILES)	**HINT test dataset size (# of unique SMILES)**	**DrugApp training dataset size (# of unique SMILES)**	**DrugApp test dataset size (# of unique SMILES)**	**DrugApp F1-score**	**HINT F1-score**
Phase I	493	174	354	130	**0.73**	0.66
Phase II	1114	233	764	154	**0.71**	0.62
Phase III	811	185	675	133	0.82	**0.85**

4.2.6. A Case Study

Some drugs are withdrawn from the market for health and safety reasons, such as severe drug side effects or problems in regulatory and manufacturing processes (Siramshetty et al., 2015). Approximately 4 % of the drugs are withdrawn from the market due to health and safety-related issues caused by reasons such as limited sample sizes or short duration of phase I–III trials (Zhang et al., 2016). To observe whether our method has the potential to address this issue, we employed our model on the drugs withdrawn from the market after being approved to see whether it would be possible to foresee their withdrawal before they even enter phase IV (by predicting them as "unapproved" for the intended disease).

The ChEMBL database provides extensive data about drugs, including withdrawal information at https://www.ebi.ac.uk/chembl/g/#browse/drugs. We downloaded the dataset of withdrawn drugs in ChEMBL v30. Among the total of 65 withdrawn drugs in this list, 12 were found in the "Nervous System Diseases" group as their intended indication. We decided to focus on these drugs in our use-case study. These drugs are presented in Table 4.11, with the reasons for withdrawal after their regulatory approval,

such as cardiotoxicity, hepatotoxicity, and neurotoxicity. Upon testing the drugs provided in Table 4.11 (using the data from phase III trials dating before their approval in the first place) in our "Nervous System Diseases" prediction model, using the leave-one-out approach, we found that 8 out of 12 drugs (66%) are predicted as unapproved, correctly. The prediction scores (which can also be read as the probability for approval) produced by the model for these drugs vary between 0.33 and 0.48, given the threshold of approval used in our study: 0.5. These scores are produced using the "predict_proba" function in the Python Scikit-learn package. The results indicate that our method is promising in revealing drugs to be withdrawn from the market sometime after their phase III approval and has the potential to prevent large-scale losses in terms of both health and the economy by providing an early warning at the end of phase III trials.

Table 4.11. The list of drugs withdrawn from the market during phase 4 trials for the indications in the "Nervous System Diseases" group, included in our use-case study. The table also includes the reason(s) for withdrawal and our prediction results. Our model produces scores that are binarized afterward (using a threshold of 0.5) to obtain classification results. Correct predictions (i.e., unapproved) are shown in bold font.

Drug Name	Reason(s) of withdrawal	Prediction	Score
Dextropropoxyphene	Cardiotoxicity	Approved	0.75
Levacetylmethadol	**Cardiotoxicity**	**Unapproved**	**0.36**
Nomifensine	**Hematological toxicity; Hepatotoxicity**	**Unapproved**	**0.37**
Pemoline	**Hepatotoxicity**	**Unapproved**	**0.48**
Pentobarbital	Misuse	Approved	0.75
Pergolide	**Cardiotoxicity**	**Unapproved**	**0.44**
Remoxipride	**Hematological toxicity**	**Unapproved**	**0.48**
Secobarbital	**Misuse**	**Unapproved**	**0.42**
Thioridazine	Cardiotoxicity	Approved	0.78
Tolcapone	Hepatotoxicity	Approved	0.65
Triazolam	**Neurotoxicity; Psychiatric toxicity**	**Unapproved**	**0.45**
Veralipride	**Neurotoxicity; Dermatological and Psychiatric toxicity**	**Unapproved**	**0.33**

4.2.7. Prospective Approval Prediction for Drug Candidates in Continuing Trials

Finally, we employed our model to predict the approval of drug candidates still in different phases of the clinical trial process (excluding phase IV) at the study's date. Here, we only considered clinical trials for single drug treatments, not drug combinations. Selected approved drug predictions (for each disease group) are listed in Table 4.12, and their prediction scores (higher scores indicate a higher chance of approval). We randomly selected four drugs from each disease group whose probability of approval is higher than our threshold score (0.5). The complete list of approved drug predictions is provided in our GitHub repository.

Table 4.12. Approval predictions for drug candidates currently in phase I, II, or III of clinical trials as of May 2022. Drug candidates are given below (randomly selected four drug-indication pairs for each disease group) and are predicted as "approved" by DrugApp models. Scores can be interpreted as the predicted probability of approval.

Disease group	Drug	Indication	Phase	Score
Rare diseases	Niclosamide	Familial Adenomatous Polyposis	2	0.80
	Bromocriptine	Peripartum Cardiomyopathy	3	0.81
	Salsalate	Preeclampsia	1	0.77
	Pazopanib	Glioblastoma Multiforme	1|2	0.61
Nervous system diseases	Atorvastatin	Obstructive Sleep Apnea	1|2	0.98
	Guanfacine	Alzheimer Disease	3	0.96
	Mifepristone	Alcohol Use Disorders	1|2	0.91
	Cariprazine	Major Depressive Disorder	3	0.56
Alimentary diseases	Nitazoxanide	Gastroenteritis	2|3	0.87
	Ozanimod	Ulcerative Colitis	3	0.60
	Benzocaine	Obesity	1	0.81
	Niacin	Type 2 Diabetes	1	0.88
Neoplasms	Enzalutamide	Ovarian Cancer	2	0.87
	Icotinib	Lung Cancer	3	0.95
	Pazopanib	Glioblastoma Multiforme	1|2	0.65
	Rocuronium	Prostate Cancer	2	0.86
Dermatological diseases	Upadacitinib	Atopic Dermatitis	3	0.57
	Cholecalciferol	Atopic Dermatitis	2	0.88
	Nifedipin	Preterm Premature Rupture of Membrane	2	0.54
	Dronabinol	Osteoarthritis	3	0.92
Urinary diseases	Flutamide	Polycystic Ovary Syndrome	1	0.82
	Darulotamide	Prostate Cancer	3	0.73
	Carvedilol	Prostate Cancer	2	0.65
	Tamsulosin	Urinary Retention	1|2	0.70
Heart diseases	Sodium Selenite	Heart Disease	3	0.61
	Milrinone	Cardiomyopathy	1	0.60
	Edoxaban	Aortic Valve Stenosis	3	0.92
	Exenatide	Acute Ischemic Stroke	2	0.92

Table 4.12 (cont.)

Immunological diseases	Enzastaurin	Diffuse Large B-Cell Lymphoma	3	0.59
	Ixazomib	Mantle Cell Lymphoma	2	0.64
	Dexamethasone	Graft vs Host Disease	2	0.97
	Tofacitinib	Juvenile Idiopathic Arthritis	3	0.83
Infective diseases	Nitazoxanide	Chronic Hepatitis B	2	0.86
	Bedaquiline	Tuberculosis	1|2	0.71
	Clofazimine	Mycobacterium Avium Complex	2	0.80
	Clevudine	COVID-19	2	0.89
Respiratory diseases	Amikacin	Pneumonia	3	0.91
	Icotinib	Lung Cancer	3	0.94
	Losartan	COVID-19	1	0.63
	Sildenafil	Bronchopulmonary Dysplasia	2	0.90
Hormonal diseases	Naloxone	Type 1 Diabetes	2	0.78
	Flutamide	Polycystic Ovary Syndrome	1	0.57
	Moxifloxacin	Diabetes	1	0.56
	Enzalutamide	Ovarian Cancer	2	0.59
Blood diseases	Cytarabine	Myelodysplastic Syndrome	2	0.96
	Ixazomib	Mantle Cell Lymphoma	2	0.54
	Pentoxifylline	Acute Lymphoblastic Leukemia	2|3	0.70
	Riociguat	Sickle Cell Disease	2	0.67
Musculoskeletal diseases	Etoricoxib	Musculoskeletal Pain	2	0.98
	Baricitinib	Juvenile Idiopathic Arthritis	3	0.77
	Montelukast	Pain	3	0.79
	Tofacitinib	Psoriatic Arthritis	3	0.89
Sensory diseases	Carbidopa	Age-Related Macular Degeneration	2	0.56
	Furosemide	Nasal Polyps	2	0.79
	Latanoprost	Open Angle Glaucoma	1	0.94
	Dexlansoprazole	Gastro Esophageal Reflux	1	0.90

CHAPTER 5

CONCLUSION

In this study, we proposed a new method to predict prospective approval of drug candidates (i.e., phase IV transition) using heterogeneous datasets and machine learning. For this, we constructed feature vectors composed of molecular and physicochemical compound/drug features and a clinical trial- and patent-related features. We employed molecular fingerprints (i.e., ECFP4) as our molecular compound/drug feature. Since molecular fingerprints represent each different sub-structure in a different, the total size of these vectors is usually high (i.e., 512, 1024, or 2048). However, this high dimensionality might dominate overall feature vectors since the remaining features were relatively low dimensional. To mitigate this risk, we decided to use the 128-bit version of ECFP4. For missing data, mean and median imputation techniques were utilized. We employed one-hot encoding for the categorical features to distinguish them from real values features. We employed the random forest (RF) algorithm to build our model. The reason behind using RF is that it has long been used in drug discovery and development studies and shown to be highly successful on many different tasks (Rifaioglu et al., 2020). We performed our predictive performance analyses using various classification metrics.

Overall, our prediction models (each comprising a different disease group) achieved considerably high performances, indicating the potential of our approach for successfully predicting drug approval using heterogeneous data. We also performed temporal analysis by dividing our dataset into six different time frames and splitting each frame into training and test folds over a threshold point in time. Again, high-performance scores indicated that our method could be utilized to generate future drug approval predictions. Furthermore, we showed that our method is also promising in terms of revealing drugs to be withdrawn from the market, even before their approval in the first place, with an acceptable sensitivity (recall) score of 0.66.

This study, for the first time in literature, explored the extensive use of patent-related features and revealed that these features have critical importance for predicting regulatory approval. The most important features for successfully predicting approval were found as the “number of related patents”, “number of claims”, “funder type”, “allocation and the location of clinical trials”, “duration of patent process”, drug’s lipophilicity (logP), molecular weight and refractivity. In future studies, only the approved patent numbers can be considered for the number of related patents. Additionally, we observed that a high number of related patents is important for predicting drug approvals, showing that when

the developers see a high chance of approval, they tend to protect the candidate with more patents. However, in the future, patent numbers may increase for all types of drug candidates regardless of whether they are promising. For future studies, it is suggested to retrain the models using current data and to use these current models for the prediction.

Limitations of this study mainly resided in the dataset preparation part. For example, the size of the unapproved dataset was smaller than the approved dataset for most of the disease groups mainly due to the lack of proper reporting of negative results, which caused dataset imbalance. In order to overcome this problem, we used class weights during the training of our prediction models. Another problem was the missing data, especially considering clinical trial and patent features, for which we used simple data imputation methods such as mean/median imputation. However, mean and median imputation can introduce bias on the mean, reduce the variance and not consider the relationship with other variables (Zhang, 2016). In case there exists a low correlation between the variables and only less than 10% of data is missing, the limitations of the mean imputation almost disappear (Raymond, 1986; Tsikriktsis, 2005). Alternative imputation methods can be utilized, such as regression-based imputation by predicting the missing values from non-missing ones or the maximum likelihood method, which produces maximum likelihood estimates of missing values (Lodder, 2013).

We had varying degrees of missingness in our dataset, with the highest degree of 51%, which was only observed in the unapproved drug dataset for the patent duration feature. Additionally, we did not have any missing values for the ECFP4 feature (128-bit). Therefore, our data contained approximately 3% missing features in total. As an alternative solution, mean imputation can be performed by taking the mean of the most similar samples to the missing sample (similarity can be based on the non-missing features). Another limitation is that the source databases for clinical trials and patent features were semi-structured, meaning that many of the data fields are filled with free text, making it extremely difficult to use fully automated pipelines. We had to curate the data manually at different study steps to overcome this issue. One further limitation was the scarcity of patent data in the literature, mainly due to accessibility-related issues. Upon utilizing multiple databases and integrating data from each one, we constructed a patent dataset with a considerable size and coverage; however, this was still a limiting factor.

The models' performances that utilize the combination of all features are observed to be the highest for all disease groups, indicating the advantage of combining different types of features to maximize the prediction performance. The most important signals for predicting drug approvals come from the clinical trial and patent features, showing the importance of using these feature types for predicting drug approvals. Drug physicochemical properties are observed to be less powerful than other features. The reason behind this could be that during the drug development process, unapproved drugs are also adjusted to follow the rules regarding drug physicochemical properties, such as

rule of five (RO5). Therefore, most unapproved and approved drugs seem consistent with the existing rules. However, although both approved and unapproved drug physicochemical properties essentially fit the existing rules, we still observed significant differences between the physicochemical properties of approved and unapproved drugs for some disease groups. This indicates that there is more to the existing rules. Therefore, our results can contribute to the literature regarding narrowing the optimized physicochemical property ranges depending on various disease classes.

This study also revealed different substructures that might be important in different disease groups. However, poor performances of the models show that substructures of drug candidate molecules are not a clear and directly detectable indicator for regulatory approval. Additionally, interpreting these substructures is still difficult because, although specific, one disease group may comprise drugs with various structural properties. If the model performances were high, we would expect the following behind the basic structural factors: the groups increasing the drug solubility would certainly have a better chance of approval for the indication groups where the route of administration requires soluble molecules. On the other hand, if the drug should pass a lipophilic barrier like the blood-brain barrier, lipophilic subgroups would be necessary for the success of the drug. In future studies, if the drugs can be categorized more precisely according to the associated tissues and target proteins, more meaningful results can be obtained regarding significant drug substructures for predicting drug approvals.

To summarize the main contributions of our study:

1. Data: To prepare enriched datasets in terms of size and scope, numerous data resources have been incorporated, and their data were integrated by both automated and manual curation processes. We utilized these datasets while developing our prediction system, but they will also be valuable for the research community in terms of being reliable benchmarks for training and evaluating other future drug development-related prediction models. During this process, it was challenging to incorporate patent data due to its primarily unstructured representation of information. Here, we combined data from multiple patent databases to obtain specific patent-related features for each drug. To our knowledge, these patent features are used here for the first time.

2. Prediction method: We trained an independent drug approval prediction model for each of the 14 disease groups, using our heterogeneous datasets that contain features from clinical trials, patents, and drugs' physicochemical and molecular properties, together with the random forest classification algorithm, which allowed us to optimize each model according to the aspects of the corresponding disease group. We rigorously tested our prediction system's potential via random split-based cross validation and temporal split-based test experiments. In the end, we observed that the proposed method has a high and stable predictive performance, indicating its reliability.

3. Investigating drugs withdrawn from the market after approval: We showed that our method has the potential to reveal drugs that were withdrawn from the market sometime after their regulatory approval, within a certain level of confidence, which indicates that it may be possible to act upon similar cases in the first place, to prevent critical losses regarding human health.

4. Sharing datasets, codes, and results: We strongly believe it is crucial to fully and adequately share in silico methods and results with the research community. For this purpose, we constructed a data repository for our study and shared all datasets with source codes and results at https://github.com/HUBioDataLab/DrugApp. We considered alignment with FAIR data principles to contribute to the movement towards fully open access, standard, and efficiently (re)usable biomedical data.

In the future, patent features can be examined more in detail. In this context, new patent-related features can be selected. Moreover, the performance of models including all non-patent features can also be compared with models using all features of this study in order to assess the effect of patent-related features in more detail. Additionally, new types of data can be incorporated into drug approval prediction systems to provide models with the ability to generalize information and produce more sensible output, especially in complicated cases (e.g., out-of-distribution samples). In this context, bioactivity profiles (obtained from target protein and cell-based assays) and previously documented side effect associations of drug candidate compounds, as well as ADMETox features, can be incorporated into the system together with text-based features of indicated diseases (e.g., curated descriptions/definitions from disease centric databases and biomedical text from the literature).

REFERENCES

Artemov, A. V., Putin, E., Vanhaelen, Q., Aliper, A., Ozerov, I., V. & Zhavoronkov, A. (2016). Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes. *BioRxiv*, 095653.

Behera, B., Kumaravelan, G., & Kumar, B. P. (2019). Performance evaluation of deep learning algorithms in biomedical document classifcation. *11th international conference on advanced computing (ICoAC)*, 220–224.

Beinse, G., Tellier, V., Charvert, V., Deutch, E., Borget, I., Massard, C., Hollebeque, a. & Verlingue, L. (2019). Prediction of Drug Approval After Phase I Clinical Trials in Oncology: RESOLVED2. *Clinical Cancer Informatics*, 3, 1-10.

Bethesda, MD, National Institutes of Health. 2009. Retrieved on 01.07.2022 from https://ftp.ncbi.nlm.nih.gov/pubchem/specifications/pubchem_fingerprints.pdf.

Breiman, L., Friedman, J., Olshen, R. & Stone, C. (1984) Classification and Regression Trees. *Chapman and Hall*, Wadsworth, New York.

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Brodersen, K.H, Ong, C.N., Stephan, K.E. & Buhmann, J.M. (2010). The balanced acuracy and its posterior distribution. *20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 3121-3124.

Brunoni A. R., Tadini L., & Fregni F. (2010). Changes in clinical trials methodology over time: a systematic review of six decades of research in psychopharmacology. *PLoS One*, 5(3):e9479.

Carracedo-Reboredo, P., Liñares-Blanco,J., Rodríguez-Fernández, N., Cedrón, F., Novoa, F. J., Carballal, A., Maojo, V., Pazos, A. & Fernandez-Lozano, C. (2021). A review on machine learning approaches and trends in drug discovery. *Comput Struct Biotechnol J*.; 19: 4538–4558.

Chicco, D. & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21, 1–13.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y. & Claster, B. V. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. J. *Clinical Epidemiol*, 110; 12–22.

Cox, D. R. (1958). The Regression Analysis of Binary Sequences. *Journal of Royal Statistical Society*, XX:2.

Deore, A. B., Jayprabha, R. D., Hrushikesh, V. W., & Rushikesh, B. S. (2019). The Stages of Drug Discovery and Development Process. *Asian Journal of Pharmaceutical Research and Development.* 2019; 7(6): 62-67.

DiMasi, J. A., Hansen, R. W. & Grabowski, H. G. (2003). The price of innovation: new estimates of drug development costs. *Journal of Health Economics*, 151-185.

DiMasi, J. A., Hermann, J. C., Twyman, K., Kondru, R. K., Stergiopoulos, S., Getz, K. A., & Rackoff, W. (2015). A tool for predicting regulatory approval after Phase II testing of new oncology compounds. *Clinical Pharmacology & Therapeutics*, 98(5), 506-513.

Doğan, T. (2018). HPO2GO: Prediction of Human Phenotype Ontology Term Associations For Proteins Using Cross Ontology Annotation Co-occurrences. *PeerJ* 6:e5298.

Doğan, T., Güzelcan, E. A., Baumann, M., Koyas, A., Atas, H., Baxendale, I., Martin, M., & Cetin-Atalay, R. (2021). Protein Domain-Based Prediction of Compound–Target Interactions and Experimental Validation on LIM Kinases. *PLOS Computational Biology*, 17(11), e1009171.

Doğan, T., Atas, H., Joshi, V., Atakan, A., Rifaioglu, A.S., Nalbat, E., Nightingale, A., Saidi, R., Volynkin, V., Zellner, H., Cetin-Atalay, R., Martin, M. J. & Atalay, V. (2021). CROssBAR: Comprehensive Resource of Biomedical Relations with Knowledge Graph Representations. *Nucleic Acids Research*, 49(16), e96-e96.

Duan, J., Dixon, S. L., Lowrie, J. F. & Sherman, W.(2010). Analysis and comparison of 2D fingerprints: insights into database screening performance using eight fingerprint methods. *J Mol Graph Model*; 29:157–70.

EPO (2022). Chapter II – Content of a European patent application (other than claims). Retrieved on 17.07.2022 from https://www.epo.org/law-practice/legal-texts/html/guidelines/e/f_ii.htm.

FDA (2016). Global Participation In Clinical Trials Report. Retrieved on 17.07.2022 from https://www.fda.gov/files/drugs/published/2015---2016-Global-Clinical-Trials-Report.pdf.

FDA (2020). 2015-2019 Drug Trials Snapshots Summary Report. Retrieved on 30.07.2022 from https://www.fda.gov/media/143592/download.

FDA (2022). Exclusivity and Generic Drugs: What Does It Mean? Retrieved on 17.07.2022 from https://www.fda.gov/files/drugs/published/Exclusivity-and-Generic-Drugs--What-Does-It-Mean-.pdf.

Feijoo, F., Palopoli, M., Bernstein, J., Siddiqui, S. & Albright, T. E.. (2020) Key indicators of phase transition for clinical trials through machine learning. *Drug Discovery Today, 25*: 414–421.

Fu, T., Huang, K., Xiao, C., Glass, L. M., & Sun, J. (2022). HINT: Hierarchical interaction network for clinical-trial-outcome predictions. *Patterns,* 3(4):1-12.

Gardner, S. & Vinter, A. (2010). Beyond Markush – protecting activity not chemical structure. Cresset-BMD Ltd. Retrieved on 08.05.2019, from https://www.cresset-group.com/.

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. & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic Acids Res*., 45(D1) D945-D954.

Gayvert, K. M., Madhukar, N. S., & Elemento, O. (2016). A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol*, 23: 1294–1301.

Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. (1999). A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. A qualitative and quantitative characterization of known drug databases. *J Comb Chem*, 1 (1): 55–68.

GlobalData (2017). Industry trials were more frequent than non-industry trials from 2012 – 2016. Retrieved on 30.07.2022 from https://www.clinicaltrialsarena.com/marketdata/industry-trials-frequent-non-industry-trials-2012-2016/.

Hou, T., Wang, J., Zhang, W. & Xu, X. (2007). ADME evaluation in drug discovery. 7. Prediction of oral absorption by correlation and classification. *J. Chem. Inf. Model.* 47, 208–218.

Kerns, E. H. & Di, L. (2008). Drug-like properties: concepts, structure design and methods; from ADME to toxicity optimization. Amsterdam: *Academic Press*; 978-0-12-369520-8.

Lal, R. (2015). Patents and Exclusivity. Retrieved on 17.07.2022 from https://www.fda.gov/media/92548/download.

Landrum, G. (2006). RDKit: Open-source cheminformatics, Retrieved from http://www.rdkit.org.

Leeson, P.D. & Springthorpe, B. (2007) The influence of drug-like concepts on decision-making in medicinal chemistry. *Nat Rev Drug Discov*., 26(11):881-90.

Linnankoski, J., Maekelae, J. M., Ranta, V.-P., Urtti, A. & Yliperttula, M. (2006) Computational prediction of oral drug absorption based on absorption rate constants in humans. *J. Med. Chem*. 49, 3674–3681.

Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. (2001). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv*. Rev., 46 (1–3): 3–26.

Lo, A. W., Siah, K. W., & Wong, C. H. (2018). Machine Learning with Statistical Imputation for Predicting Drug Approvals. Retrieved on 17.07.2022 from http://dx.doi.org/10.2139/ssrn.2973611.

Lodder, P. (2013). To Impute or not Impute: That's the Question. *Paper Methodological Advice*, Retrieved on 17.07.2022 from https://www.paultwin.com/wp-content/uploads/Lodder_1140873_Paper_Imputation.pdf.

Lu, J. J., Crimin, K., Goodwin, J. T., Crivori, P., Orrenius, L. X., Tandler, P. J., Vidmar, T. J., Amore, B. M., Wilson, A. G. E., Stouten, P. F. W. & Burton, P. S. (2004). Influence of molecular flexibility and polar surface area metrics on oral bioavailability in the rat. *J. Med. Chem*. 47, 6104–6107 (2004).

Mollah, A. H., Baseman, H. S., Long, M. & Rathore, A. S. (2014). A practical discussion of risk management for manufacturing of pharmaceutical products. *PDA J Pharm Sci Technol.*, 68(3):271-80.

Opitz, J., & Burst, S. (2021). Macro f1 and macro f1. *arXiv*:1911.03347.

Papadatos, G., Davies, M., Chambers, J., Gaulton, A., Siddle, J., Koks, R., Pettersson, J., Goncharoff, N., Hersey, A. & Overington, J. P.(2016). SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res*., 44(D1):D1220-8.

Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Duboutg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perror, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*. 12(Oct), 2825-2830.

Probst, P., Wright, M. N., & Boulesteix, A. (2019). Hyperparameters and tuning strategies for random forest. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9.

Raymond, M. R. (1986). Missing data in evaluation research. *Evaluation & the health professions*, 9 (4), 395–420.

Tsikriktsis, N. (2005). A review of techniques for treating missing data in om survey research. *Journal of Operations Management*, 24 (1), 53–62.

Ritchie, T. J. & Macdonald, J. F. (2009). The impact of aromatic ring count on compound developability – are too many aromatic rings a liability in drug design? *Drug Discovery Today*, Volume 14, Issues 21–22, 1011-1020.

Rifaioglu, A. S., Atas, H., Martin, M., Cetin Atalay, R., Atalay, V. & Doğan, T (2018). Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. *Briefings in Bioinformatics*, 1–36.

Rifaioglu, A. S., Cetin Atalay, R., Cansen Kahraman, D., Doğan, T., Martin, M., & Atalay, V. (2020). MDeePred: Novel Multi-Channel protein featurization for deep learning based binding affinity prediction in drug discovery. *Bioinformatics*, 37(5), 693-704.

Rifaioglu, A.S., Nalbat, E., Atalay, M.V., Martin, M.J., Cetin-Atalay, R. & Doğan, T. (2020). DEEPScreen: High Performance Drug-Target Interaction Prediction with Convolutional Neural Networks Using 2-D Structural Compound Representations. *Chemical Science*, 11(9), 2531-2557.

Rifaioglu, A.S. (2020). Deep Learning for Prediction of Drug-Target Interaction Space and Protein Functions. *A Thesis Submitted* To *The Graduate School Of Natural And Applied Sciences, METU*, June 2020.

Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5), 742-754.

USPTO (2019). PatentsView database. Retrieved on 22.12.2019 from https://www.patentsview.org/download/.

Sastry, M., Lowrie, J. F., Dixon, S. L. & Sherman, W. (2010). Large-scale systematic analysis of 2D fingerprint methods and parameters to improve virtual screening enrichments. *J Chem Inf Model*. 2010;50:771–784.

Shamout, F., Zhu, T. & Clifton, D. A. (2021). Machine Learning for Clinical Outcome Prediction. *IEEE Reviews In Biomedical Engineering*, 14.

Siramshetty, V. B., Nickel, J., Omieczynski, C., Gohlke, B. O., Drwal, M. N. & Preissner, R. (2016). WITHDRAWN—a resource for withdrawn and discontinued drugs. *Nucleic Acids Research*, Volume 44, Issue D1, D1080–D1086.

Suvarna, V. (2010). Phase IV of Drug Development. *PICR*, 1(2).

Todeschini, R. & Consonni, V. 2008. Handbook of Molecular Descriptors. Weinheim, Germany: *Wiley-VCH*, ISBN 3-52-29913-0.

Todeschini, R. & Consonni, V. 2009. Molecular Descriptors for Chemoinformatics, Weinheim, Germany: *Wiley-VCH*, ISBN 978-3-527-31852-0.

Tolles, J. & Meurer, W. J. (2016). Logistic regression relating patient characteristics to outcomes. *JAMA*, 316(5): 533–4.

Valance, E. H. (1961). Understanding the Markush claim in chemical patents. *J Chem Doc*. 1:87–92.

Veber, D. F., Johnson, S. R., Cheng, H. Y., Smith, B. R., Ward, K.W. & Kopple, K. D. (2002). Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem*., 45 (12): 2615–23.

Vidal, D., Thormann, M. & Pons, M. (2005). LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J Chem Inf Model*; 45(2):386–93

Vistoli, G., Pedretti, A. & Testa, B. (2008). Assessing drug-likeness--what are we missing? *Drug Discov* Today ;13(7-8):285-94.

WHO (2022). Number of clinical trials by year, country, WHO region and income group (1999-2021). Retrieved on 30.07.2022 from https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/number-of-clinical-trials-by-year-country-who-region-and-income-group.

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., Chang, Z. & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic Acids Res*, 1;34: D668-72. 16381955.

WIPO (2022). Patents. Retrieved on 17.07.2022 from https://www.wipo.int/patents/en/.

WIPO (2022). Time Limits for Entering National/Regional Phase under PCT Chapters I and II. Retrieved on 17.07.2022 from https://www.wipo.int/pct/en/texts/time_limits.html.

Wood, D. J., Vlieg, J., De Wagener, M. & Ritschel, T. (2012). Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *J Chem Inf Model*;52(8):2031–43.

Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. & Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8), 3370-3388.

Zarin, D. A., Tse, T., Williams, R. J., Califf, R. M., M.D., & Ide, N. C. (2011). The ClinicalTrials.gov results database—update and key issues. *N Engl J Med*., 364(9):852-60.

Zhang, X., Zhang, Y., Ye, X., Guo, X., Zhang, T. & He, J. (2016). Overview of phase IV clinical trials for postmarket drug safety surveillance: a status report from the ClinicalTrials.gov registry. *BMJ Open*, 6:e010643.

Zhang, Z. (2016). Missing data imputation: focusing on single imputation. *Ann Transl Med*, 4(1): 9.

Zhavoronkov, A., Kudrin, R., Tutubalina, E., Kuzmina, A., Korbut, D., Kadurin, A., Polykovskiy, D. & Aliper, A. (2020). Multimodal AI Engine for Clinical Trials Outcome Prediction: Prospective Case Study Summer 2020. Retrieved on 27 May 2022 from https://www.researchgate.net/publication/342354346_Multimodal_AI_Engine_for_Clinical_Trials_Outcome_Prediction_Prospective_Case_Study_Summer_2020.

APPENDICES

APPENDIX A

IMPORTANT DRUG SUBSTRUCTURES OF REGULATORILY APPROVED AND UNAPPROVED DRUGS

Table 4.13 shows the top twenty most important substructures for "Rare Diseases" group. We observed that ">=16 C", "N(~C)(~H)", "O-H", "C(~O)(~O)", "[#1]-C-C-N-[#1]", "C(-N)(=O)", "N(~C)(~C)(~C)", ">=2 aromatic rings", "C(-O)(=O)", ">=2 any ring size 6", ">= 1 any ring size 5", "C(-C)(-N)(=O)", "C-F", ">=1 F", "C(~F)(:C)" and "N-O" could be important for predicting drug approvals in "Rare Diseases" group. Moreover, "C-F", ">=1 F", "C(~F)(:C)" and "N-O" substructures were observed to be higher in number in the unapproved drug category (Table 4.3) , showing that these substructures may have important associations with failure.

Table 4.13. The top twenty most important substructures for "Rare Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
C-F	**0.006282**	**115**	**135**
>=1 F	**0.005667**	**118**	**135**
>=1 Cl	0.005612	147	114
C(~H)(~H)(~H)	0.005152	680	431
>=16 C	0.005141	597	442
N(~C)(~H)	0.005128	532	425
O-H	0.004912	555	305
C(~O)(~O)	0.004882	402	209
[#1]-C-C-N-[#1]	0.004741	435	364
C(-N)(=O)	0.004706	371	326
C(~F)(:C)	**0.004703**	**56**	**87**
>= 4 O	0.004659	494	283
C(~C)(~C)(~C)(~H)	0.004654	302	161
N(~C)(~C)(~C)	0.004501	415	289
>=2 aromatic rings	0.004490	424	376
C(-O)(=O)	0.004441	367	188
>=2 any ring size 6	0.004424	561	407
N-O	**0.004348**	**45**	**56**
>= 1 any ring size 5	0.004295	465	322
C(-C)(-N)(=O)	0.004268	328	285

Table 4.14 shows the top twenty most important substructures for “Nervous Diseases” group. We observed that drug substructures, for example, “[#1]-C-C-N-[#1]”, “N-H”, “C(~O)(~O)”, “N(~C)(~H)”, “N-C-C-C-N”, “C(-C)(-N)(=O)”, “C-C-O-C-C”, “C-F”, “>=1 F” and “C(~F)(:C)” could be important for predicting drug approvals in “Nervous Diseases” group. Moreover, “C-F”, “>=1 F” and “C(~F)(:C)” substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.14. The top twenty most important substructures for “Nervous Diseases” group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
C(~H)(~H)(~H)	0.007666	574	283
>=1 F	**0.007394**	**100**	**102**
C-F	**0.007341**	**98**	**101**
>=1 Cl	0.006187	133	75
C(~F)(:C)	**0.006172**	**48**	**64**
[#1]-C-C-N-[#1]	0.005397	329	224
>= 4 O	0.005344	361	180
N-H	0.004875	414	267
C(~O)(~O)	0.004819	316	132
C-C(C)-C(C)-C	0.004764	302	159
>=32 H	0.004690	172	100
O-H	0.004659	435	207
[#1]-N-C-[#1]	0.004574	257	166
C(~C)(~H)(~O)	0.004351	388	191
N(~C)(~H)	0.004351	404	266
N-C-C-C-N	0.004301	167	141
O=C-C-C-C-C-C	0.004300	334	176
>=2 saturated or aromatic carbon-only ring size 6	0.004294	337	181
C(-C)(-N)(=O)	0.004276	262	181
C-C-O-C-C	0.004275	280	136

Table 4.15 shows the top twenty most important substructures for “Alimentary Diseases” group. We observed that drug substructures, for example, “[#1]-C-C-N-[#1]”, “>=1 Cl”, “>=3 any ring size 6”, “N(~C)(~C)(~H)”, “N(~C)(~H)”, “O=C-C=C-[#1]”, “>=2 N”, “C-F”, “>=1 F” and “>=2 F” could be important for predicting drug approvals in “Alimentary Diseases” group. Moreover, “C-F”, “>=1 F” and “>=2 F” substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.15. The top twenty most important substructures for "Alimentary Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
>=1 F	**0.008819**	**93**	**94**
>=1 Cl	0.007603	112	79
C-F	**0.006746**	**88**	**93**
[#1]-C-C-N-[#1]	0.006682	331	233
>=3 any ring size 6	0.006562	274	184
C(~H)(~H)(~H)	0.005885	556	285
N(~C)(~C)(~H)	0.005480	313	216
>= 4 O	0.005180	411	193
N(~C)(~H)	0.005075	405	267
O=C-C=C-[#1]	0.004783	168	122
C(~O)(~O)	0.004538	336	157
O-H	0.004506	456	215
C(-O)(=O)	0.004397	305	143
O(~C)(~H)	0.004377	421	200
>=2 N	0.004305	438	284
>=2 F	**0.004297**	**42**	**49**
O=C-C-C-C-C-O	0.004225	120	36
N(~C)(~C)(~C)	0.004191	356	188
>=1 saturated or aromatic nitrogen-containing ring size 6	0.004147	219	163
C(-C)(-O)(=O)	0.004069	275	125

Table 4.16 shows the top twenty most important substructures for "Dermatological Diseases" group. We observed that drug substructures, for example, ">=1 Cl", ">=4 N", "C-C-N-C-C", ">=3 any ring size 6", ">=2 N", ">=2 O", "O=C-N-C-[#1]", ">=1 S", "C-F", ">=1 F", ">=2 F" and "C(~F)(:C)" could be important for predicting drug approvals in "Dermatological Diseases" group. Moreover, "C-F", ">=1 F", ">=2 F" and "C(~F)(:C)" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.16. The top twenty most important substructures for "Dermatological Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
C-F	**0.009622**	**86**	**95**
>=1 F	**0.006922**	**88**	**96**
>=1 Cl	0.006283	125	70
>=4 N	0.006252	162	157
>= 4 O	0.005659	346	192
>= 1 any ring size 5	0.005067	333	207
C-C-N-C-C	0.005059	380	275
>=3 any ring size 6	0.005055	265	184
C(~H)(~H)(~H)	0.004659	509	258
C(~F)(:C)	**0.004572**	**43**	**57**
>=2 N	0.004554	392	279
>=16 C	0.004540	460	269
O-H	0.004479	429	211
>=2 F	**0.004469**	**44**	**52**
O(~C)(~H)	0.004345	402	187
>=1 saturated or aromatic heteroatom-containing ring size 5	0.004332	246	177
>=2 O	0.004319	572	312
O=C-N-C-[#1]	0.004259	172	141
C(-O)(=O)	0.004167	277	118
>=1 S	0.004139	150	97

Table 4.17 shows the top twenty most important substructures for "Infective Diseases" group. We observed that drug substructures, for example, "O=C-C=C-[#1]", ">=16 C", ">=1 Cl", ">=4 N", "C-C-N-C-C", "N-O", "C-F", ">=1 F", and ">=2 F" could be important for predicting drug approvals in "Infective Diseases" group. Moreover, "C-F", ">=1 F", ">=2 F" and ">=1 saturated or aromatic nitrogen-containing ring size 9" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.17. The top twenty most important substructures for "Infective Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
>=1 F	**0.007422**	**54**	**66**
C-F	**0.007367**	**53**	**65**
>=2 F	**0.006902**	**27**	**41**
O=C-C=C-[#1]	0.006539	110	86
>=16 C	0.005950	328	206
>=1 Cl	0.005832	84	58
>=4 N	0.005578	163	119
>=1 any ring size 9	0.005252	117	83
C-C-N-C-C	0.005247	302	197
N-O	0.005119	29	27
>= 4 O	0.004761	285	145
>=1 saturated or aromatic nitrogen-containing ring size 9	**0.004731**	**54**	**60**
C(~O)(~O)	0.004703	240	101
O=C-N-C-C	0.004629	189	126
>=32 H	0.004616	140	86
>=2 any ring size 5	0.004609	73	60
C(~C)(~C)(~H)(~N)	0.004566	213	94
>=1 saturated or aromatic heteroatom-containing ring size 9	0.004527	73	66
C(~H)(~H)(~H)	0.004400	392	191
[#1]-C-O-[#1]	0.004394	181	82

Table 4.18 shows the top twenty most important substructures for "Urinary Diseases" group. We observed that drug substructures, for example, ">=1 Cl", ">=3 any ring size 6", ">=2 N", ">= 4 O", "O=C-N-C-[#1]", ">=4 N", "O=C-C=C-C", "O-C-C-C-C-C-N-C", "C(~O)(~O)", ">=32 H", "N-C-C-C-C", ">=1 saturated or aromatic heteroatom-containing ring size 6", "C-F", ">=1 F", "C(~F)(:C)" and ">= 1 B" could be important for predicting drug approvals in "Urinary Diseases" group. Moreover, "C-F", ">=1 F", "C(~F)(:C)" and ">= 1 B" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.18. The top twenty most important substructures for "Urinary Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
>=1 Cl	0.009429	101	66
>=1 F	**0.007968**	**68**	**87**
C-F	**0.007395**	**66**	**87**
>=3 any ring size 6	0.006620	244	170
>=2 N	0.005093	390	267
O-C-C-C-C-C-N-C	0.005044	102	89
>= 4 O	0.004943	336	182
O=C-N-C-[#1]	0.004885	184	136
>=4 N	0.004880	173	150
O-H	0.004817	389	197
C(~H)(~H)(~H)	0.004776	494	244
O=C-C=C-C	0.004612	192	131
[#1]-C-C-N-[#1]	0.004584	305	203
O=C-C=C-[#1]	0.004517	128	104
>=1 saturated or aromatic heteroatom-containing ring size 6	0.004218	243	162
C(~O)(~O)	0.004186	278	117
>=32 H	0.004186	160	103
C(~F)(:C)	**0.004163**	**39**	**55**
N-C-C-C-C	0.004131	444	284
>= 1 B	**0.003945**	**2**	**5**

Table 4.19 shows the top twenty most important substructures for "Heart Diseases" group. We observed that drug substructures, for example, ">= 1 any ring size 5", ">=3 any ring size 6", "N(~C)(~C)(~H)", ">=8 O", "O=C-C=C-[#1]", "C-F" and ">=1 F" could be important for predicting drug approvals in "Heart Diseases" group. Moreover, "C-F" and ">=1 F" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.19. The top twenty most important substructures for "Heart Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
C-F	**0.007749**	**56**	**57**
>= 4 O	0.006090	264	136
>=32 H	0.005883	110	71
[#1]-N-C-[#1]	0.005748	196	113
N(~C)(~C)(~C)	0.005580	234	117
[#1]-C-C-N-[#1]	0.005276	244	141
>= 1 any ring size 5	0.005247	241	144
>=1 F	**0.005119**	**57**	**57**
>=1 S	0.004981	114	68
>=3 any ring size 6	0.004930	182	129
>=16 H	0.004816	384	208
C(~C)(~H)(~O)	0.004786	264	135
N(~C)(~C)(~H)	0.004718	228	141
>=8 O	0.004614	88	29
>=1 saturated or aromatic heteroatom-containing ring size 5	0.004535	182	122
>=2 O	0.004523	418	223
C(-O)(=O)	0.004514	206	98
C-C-C-C-C-C(C)-C	0.004449	237	129
O=C-C=C-[#1]	0.004361	116	79
C(~H)(~H)(~H)	0.004356	382	184

Table 4.20 shows the top twenty most important substructures for "Neoplasms" group. We observed that drug substructures, for example, ">=1 Cl", ">=32 H", ">=16 C", ">=2 N", "N-H", "N(~C)(~C)(~C)", "C-C-N-C-C", "C(-C)(-N)(=O)", "N(~C)(~H)", ">=4 N", ">=1 F", and "N-C-C-C-N" could be important for predicting drug approvals in "Neoplasms" group. Moreover, ">=4 N", ">=1 F", "N-C=C-[#1]" and "N-C-C-C-N" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.20. The top twenty most important substructures for "Neoplasms" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
>=1 Cl	0.006397	103	91
>=4 N	**0.006345**	**197**	**236**
>=32 H	0.006220	158	138
>=16 C	0.005611	460	376
>=2 N	0.005413	410	385
N-H	0.005397	408	364
N(~C)(~C)(~C)	0.005352	314	257
[#1]-C-C-N-[#1]	0.005346	332	303
C(~H)(~H)(~H)	0.005123	523	367
>= 4 O	0.004956	349	239
N-C=C-[#1]	**0.004938**	**252**	**277**
>=1 F	**0.004917**	**100**	**111**
C-C-N-C-C	0.004812	399	375
N-C-C-C-N	**0.004696**	**171**	**198**
C(-C)(-N)(=O)	0.004685	240	238
N(~C)(~H)	0.004589	400	361
O-H	0.004454	406	253
C(~C)(~H)(~O)	0.004348	337	228
O(~C)(~H)	0.004314	378	228
>=2 saturated or aromatic heteroatom-containing ring size 6	0.004198	85	107

Table 4.21 shows the top twenty most important substructures for "Immunological Diseases" group. We observed that drug substructures, for example, ">=4 N", "O=C-C=C-C", ">=16 C", "O=C-N-C-[#1]", ">= 1 any ring size 5", "O(~C)(~C)", ">=1 Cl", "C-F", ">=1 F" and ">=2 F" could be important for predicting drug approvals in "Immunological Diseases" group. Moreover, "C-F", ">=1 F" and ">=2 F" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.21. The top twenty most important substructures for "Immunological Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
C-F	**0.006957**	**71**	**83**
>=1 F	**0.006890**	**75**	**84**
>=4 N	0.006207	172	161
>=2 F	**0.005555**	**34**	**47**
O=C-C=C-C	0.005533	192	141
>=4 any ring size 6	0.005125	101	91
C(~C)(~C)(~H)(~O)	0.005076	225	92
>=16 C	0.004604	419	262
>=1 saturated or aromatic heteroatom-containing ring size 5	0.004597	233	161
C(~C)(~C)(~C)(~H)	0.004588	224	96
O-C-C-N	0.004479	240	124
>= 4 O	0.004469	324	171
[#1]-C-C-N-[#1]	0.004354	279	203
O=C-N-C-[#1]	0.004340	167	133
>= 1 any ring size 5	0.004311	303	185
O(~C)(~C)	0.004307	292	176
O=C-C=C-[#1]	0.004295	150	115
>=1 Cl	0.004283	114	74
O-C-C-N-C	0.004272	212	115
C(~O)(~O)	0.004223	266	111

Table 4.22 shows the top twenty most important substructures for "Respiratory Diseases" group. We observed that drug substructures, for example, "O=C-N-C-[#1]", ">=4 N", "N-C-C-N-C", ">=2 N", ">=2 O", ">=2 aromatic rings", ">=2 any ring size 5", "C-F", ">=1 F" and ">=2 F" could be important for predicting drug approvals in "Respiratory Diseases" group. Moreover, "C-F", ">=1 F", ">=2 F" and >=2 saturated or aromatic heteroatom-containing ring size 6" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.22. The top twenty most important substructures for "Respiratory Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
O=C-N-C-[#1]	0.007241	156	129
>=1 F	**0.006384**	**73**	**76**
>=4 N	0.006318	149	148
>=2 F	**0.005914**	**31**	**46**
>=4 any ring size 6	0.005849	85	76
C-F	**0.005523**	**71**	**75**
N-C-C-N-C	0.005482	169	128
>=1 Cl	0.005157	96	55
C(-O)(=O)	0.004918	229	100
O-H	0.004909	348	177
O=C-C-C-C-N	0.004858	67	55
>=2 N	0.004831	350	246
>=2 saturated or aromatic heteroatom-containing ring size 6	**0.004815**	**69**	**71**
C(~O)(~O)	0.004815	252	107
C(~H)(~H)(~H)	0.004532	447	215
[#1]-C-C-N-[#1]	0.004454	275	179
>=2 O	0.004370	483	263
>=3 any ring size 6	0.004262	220	150
>=2 aromatic rings	0.004211	278	205
>=2 any ring size 5	0.004162	75	65

Table 4.23 shows the top twenty most important substructures for "Hormonal Diseases" group. We observed that drug substructures, for example, ">=32 H", "O=C-C=C", ">=1 F", "C-F", "O=C-N-C-[#1]", "N-C-C:C-C", ">=2 O", "[#1]-C-C-N-[#1]", ">=16 C", ">= 1 B", "S(-O)(=O)" and "S(~C)(~O)" could be important for predicting drug approvals in "Hormonal Diseases" group. Moreover, ">= 1 B", "S(-O)(=O)", S(=O)(=O) and "S(~C)(~O)" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.23. The top twenty most important substructures for "Hormonal Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
>= 1 B	**0.006614**	**0**	**4**
>= 4 O	0.006552	235	135
S(-O)(=O)	**0.006193**	**26**	**35**
>=32 H	0.005656	105	78
S(=O)(=O)	**0.005450**	**26**	**35**
O=C-C=C	0.005263	157	102
>=1 Cl	0.005098	81	53
>=1 F	0.005045	60	58
C-F	0.005013	58	57
O=C-N-C-[#1]	0.004969	127	94
N-C-C:C-C	0.004903	197	150
S(~C)(~O)	**0.004867**	**25**	**30**
C(~C)(~H)(~O)	0.004817	234	127
C(-O)(=O)	0.004688	172	88
>=2 O	0.004688	377	204
O-C-C-C-N	0.004688	133	67
C(~H)(~H)(~H)	0.004677	348	177
>=3 any ring size 6	0.004634	176	124
[#1]-C-C-N-[#1]	0.004582	212	138
>=16 C	0.004503	322	192

Table 4.24 shows the top twenty most important substructures for "Blood Diseases" group. We observed that drug substructures, for example, ">=2 N", "O=C-N-C-[#1]", "O-H", "O=C-C=C-C", ">=2 aromatic rings", "N(~C)(~C)(~H)", "N-H", "N-C-C=C", "[#1]-C-O-[#1]", "O=C-C=C-[#1]", "C-F", ">=1 F", "N=C-C-C", "N-O" and "N-C-C-O-C" could be important for predicting drug approvals in "Blood Diseases" group. Moreover, "C-F", ">=1 F", "N=C-C-C", "N-O" and "N-C-C-O-C" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.24. The top twenty most important substructures for "Blood Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
>= 4 O	0.006429	227	131
C-F	**0.005752**	**48**	**64**
>=2 N	0.005746	256	208
O=C-N-C-[#1]	0.005662	121	101
>=1 F	**0.005424**	**49**	**64**
>=3 any ring size 6	0.005257	176	131
N=C-C-C	**0.005212**	**137**	**137**
O-H	0.005125	287	127
O=C-C=C-C	0.005089	135	113
C(~C)(~C)(~H)(~N)	0.004935	159	89
>=2 aromatic rings	0.004852	201	175
N(~C)(~C)(~H)	0.004839	174	161
N-H	0.004770	246	198
N-O	**0.004763**	**18**	**20**
N-C-C=C	0.004684	225	192
C-S	0.004617	99	53
N-C-C-O-C	**0.004494**	**58**	**62**
[#1]-C-O-[#1]	0.004393	152	57
O=C-C=C-[#1]	0.004386	104	93
C(~H)(~H)(~H)	0.004385	319	190

Table 4.25 shows the top twenty most important substructures for "Musculoskeletal Diseases" group. We observed that drug substructures, for example, ">=16 H", "C-C-N-C-C", "C(-C)(-N)(=O)", "O=C-N-C-C", "C-F", ">=1 F", ">=2 F", ">=1 Sm", ">= 1 B", ">=3 aromatic rings", and "C(~F)(:C)" could be important for predicting drug approvals in "Musculoskeletal Diseases" group. Moreover; "C-F", ">=1 F", ">=2 F", ">=1 Sm", ">= 1 B", ">=3 aromatic rings", >=4 any ring size 6 and "C(~F)(:C)" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.25. The top twenty most important substructures for "Musculoskeletal Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
C-F	**0.009453**	**30**	**43**
>=2 F	**0.007813**	**12**	**26**
>=1 F	**0.007636**	**31**	**43**
>=1 Sm	**0.007583**	**0**	**1**
>= 1 B	**0.007534**	**0**	**2**
>=3 aromatic rings	**0.006248**	**54**	**56**
>=16 H	0.005973	188	109
C-C-N-C-C	0.005906	143	108
C(-C)(-N)(=O)	0.005783	88	70
O-H	0.005438	156	80
O=C-N-C-C	0.005319	83	63
>=3 any ring size 6	0.005270	77	72
>= 4 O	0.005092	131	69
>=2 aromatic rings	0.005066	118	93
>=32 H	0.005037	49	34
[#1]-N-C-[#1]	0.004970	90	56
>=4 any ring size 6	**0.004866**	**21**	**32**
C(~F)(:C)	**0.004670**	**16**	**24**
Cc1ccc(O)cc1	0.004663	41	33
C-C-C-O-[#1]	0.004631	130	64

Table 4.26 shows the top twenty most important substructures for "Sensory Diseases" group. We observed that drug substructures, for example, "N-C-C:C-C", "C-C-N-C-C", "N=C-C=C", "N=O" and "C-C-C-C-C-C-C", "C-F", ">=1 F", ">=2 F", "N-O", "N:C-C:C", "C(~F)(:C)", ">=4 aromatic rings", "C-C:N:C", "O-N-C-C", "C(~F)(~F)" and "C(~C)(:N)" could be important for predicting drug approvals in "Sensory Diseases" group. Moreover; "C-F", ">=1 F", ">=2 F", "N-O", "N:C-C:C", "C(~F)(:C)", ">=4 aromatic rings", "C-C:N:C", "O-N-C-C", "C(~F)(~F)" and "C(~C)(:N)" substructures were observed to be higher in number in the unapproved drug category, showing that these substructures may have important associations with failure.

Table 4.26. The top twenty most important substructures for "Sensory Diseases" group. Features corresponds to substructures from PubChem Fingerprints representing 881 substructures. Bold: substructures higher in number in the unapproved drugs.

Feature	Importance (MDI)	# of Approved	# of Unapproved
C-F	**0.012166**	**16**	**26**
>=1 F	**0.011330**	**16**	**26**
>=2 F	**0.009745**	**5**	**14**
N-O	**0.008224**	**4**	**10**
N-C-C:C-C	0.007563	68	62
>=3 any ring size 6	0.007330	61	52
N:C-C:C	**0.007253**	**29**	**46**
C(~F)(:C)	**0.007205**	**10**	**18**
>=4 aromatic rings	**0.006949**	**17**	**33**
N:C:C-C	**0.006854**	**38**	**49**
N=C-C-[#1]	**0.006392**	**43**	**51**
C-C:N:C	**0.006042**	**37**	**47**
O-N-C-C	**0.005843**	**2**	**6**
C-C-N-C-C	0.005771	105	78
C(~F)(~F)	**0.005701**	**2**	**9**
C(~C)(:N)	**0.005478**	**40**	**50**
N=C-C=C	0.005298	37	50
N=C-C:C-[#1]	0.005265	29	40
N=O	0.005247	2	5
C-C-C-C-C-C-C	0.005217	132	81

APPENDIX B

IMPORTANT FEATURES FOR PREDICTING DRUG APPROVALS

Table 4.27. The top ten important features for predicting drug approvals for each disease group, shown using mean decrease impurity (MDI).

Disease groups	Importance (MDI)
All diseases	
Number of related patents	0.065
Funder type of clinical trials	0.063
Duration of patent process	0.040
Number of claims	0.025
Polarizability of the drug	0.024
Refractivity of the drug	0.024
Lipophilicity (logP) of the drug	0.023
Molecular weight of the drug	0.022
Allocation of clinical trials	0.020
Polar surface area of the drug	0.020
Rare diseases	
Number of related patents	0.051
Funder type of clinical trials	0.038
Duration of patent process	0.034
Number of claims	0.028
Lipophilicity (logP) of the drug	0.027
Polarizability of the drug	0.025
Molecular weight of the drug	0.024
Refractivity of the drug	0.024
Polar surface area of the drug	0.022
Location of clinical trials	0.022
Nervous diseases	
Number of related patents	0.057
Funder type of clinical trials	0.050
Duration of patent process	0.033
Number of claims	0.027
Molecular weight of the drug	0.025
Polar surface area of the drug	0.023
Polarizability of the drug	0.023
Lipophilicity (logP) of the drug	0.023
Refractivity of the drug	0.021
Allocation of clinical trials	0.020
Alimentary diseases	
Number of related patents	0.059
Duration of patent process in months	0.033
Funder type of clinical trials	0.032
Number of claims	0.028
Lipophilicity (logP) of the drug	0.028

Table 4.27 (cont.)	
Location of clinical trials	0.026
Polar surface area of the drug	0.025
Refractivity of the drug	0.024
Polarizability of the drug	0.023
Allocation of clinical trials	0.020
Neoplasms	
Number of related patents	0.058
Funder type of clinical trials	0.036
Location of clinical trials	0.035
Duration of patent process in months	0.033
Polar surface area of the drug	0.027
Molecular weight of the drug	0.026
Number of claims	0.025
Refractivity of the drug	0.023
Polarizability of the drug	0.021
Lipophilicity (logP) of the drug	0.020
Dermatological diseases	
Number of related patents	0.057
Funder type of clinical trials	0.036
Duration of patent process	0.035
Refractivity of the drug	0.026
Polar surface area of the drug	0.025
Molecular weight of the drug	0.025
Lipophilicity (logP) of the drug	0.024
Polarizability of the drug	0.024
Number of claims	0.016
Allocation of clinical trials	0.014
Urinary tract diseases	
Funder type of clinical trials	0.049
Number of related patents	0.039
Duration of patent process	0.028
Polar surface area of the drug	0.027
Location of clinical trials	0.026
Number of claims	0.026
Lipophilicity (logP) of the drug	0.026
Molecular weight of the drug	0.026
Intervention model of clinical trials	0.026
Polarizability of the drug	0.023
Heart diseases	
Funder type of clinical trials	0.070
Number of related patents	0.044
Duration of patent process in months	0.033
Molecular weight of the drug	0.027
Polarizability of the drug	0.026
Refractivity of the drug	0.025
Lipophilicity (logP) of the drug	0.025
Number of claims	0.025
Polar surface area of the drug	0.024
Allocation of clinical trials	0.021

Table 4.27 (cont.)	
Immunological diseases	
Number of related patents	0.052
Duration of patent process	0.035
Molecular weight of the drug	0.030
Polar surface area of the drug	0.029
Number of claims	0.027
Polarizability of the drug	0.027
Refractivity of the drug	0.026
Funder type of clinical trials	0.026
Lipophilicity (logP) of the drug	0.025
Intervention model of clinical trials	0.020
Infective diseases	
Number of related patents	0.051
Funder type of clinical trials	0.045
Duration of patent process	0.030
Lipophilicity (logP) of the drug	0.027
Molecular weight of the drug	0.025
Polar surface area of the drug	0.025
Polarizability of the drug	0.024
Refractivity of the drug	0.024
Number of claims	0.023
Number of rotatable bonds	0.016
Respiratory diseases	
Number of related patents	0.053
Duration of patent process	0.036
Funder type of clinical trials	0.028
Polar surface area of the drug	0.027
Molecular weight of the drug	0.026
Number of claims	0.025
Lipophilicity (logP) of the drug	0.025
Polarizability of the drug	0.025
Refractivity of the drug	0.024
Location of clinical trials	0.022
Hormonal diseases	
Number of related patents	0.053
Location of clinical trials	0.036
Duration of patent process	0.032
Allocation of clinical trials	0.030
Funder type of clinical trials	0.030
Molecular weight of the drug	0.028
Number of claims	0.026
Polarizability of the drug	0.025
Polar surface area of the drug	0.025
Lipophilicity (logP) of the drug	0.022
Blood diseases	
Number of related patents	0.053
Location of clinical trials	0.048
Duration of patent process	0.031
Age of participants	0.029

Table 4.27 (cont.)	
Funder type of clinical trials	0.029
Intervention model of clinical trials	0.027
Molecular weight of the drug	0.027
Refractivity of the drug	0.027
Polar surface area of the drug	0.026
Polarizability of the drug	0.024
Musculoskeletal diseases	
Number of related patents	0.052
Duration of patent process	0.035
Allocation of clinical trials	0.030
Refractivity of the drug	0.026
Polar surface area of the drug	0.026
Polarizability of the drug	0.025
Number of claims	0.025
Molecular weight of the drug	0.024
Lipophilicity (logP) of the drug	0.024
Number of related patents	0.052
Sensory diseases	
Duration of patent process	0.032
Polarizability of the drug	0.031
Number of related patents	0.031
Molecular weight of the drug	0.029
Polar surface area of the drug	0.028
Refractivity of the drug	0.028
Number of claims	0.023
Location of clinical trials	0.022
Lipophilicity (logP) of the drug	0.021
Allocation of clinical trials	0.019

Table 4.28. The top ten most important features for predicting drug approvals in temporal analysis, shown using mean decrease impurity (MDI).

Training time frames	Importance (MDI)
0-2008	
Number of related patents	0.086
Duration of patent process	0.030
Molecular weight of the drug	0.029
Funder type of clinical trials	0.028
Polar surface area of the drug	0.027
Location of clinical trial	0.023
Refractivity of the drug	0.022
Polarizability of the drug	0.023
Lipophilicity (logP) of the drug	0.022
Number of claims	0.018
0-2010	
Number of related patents	0.092
Funder type of clinical trials	0.043
Duration of patent process	0.035
Allocation of clinical trials	0.024
Molecular weight of the drug	0.023
Number of claims	0.022
Location of clinical trial	0.021
Refractivity of the drug	0.020
Polarizability of the drug	0.020
Lipophilicity (logP) of the drug	0.019
0-2012	
Number of related patents	0.089
Funder type of clinical trials	0.055
Duration of patent process	0.038
Allocation of clinical trials	0.024
Number of claims	0.024
Location of clinical trial	0.022
Molecular weight of the drug	0.022
Polarizability of the drug	0.020
Refractivity of the drug	0.020
Polar surface area of the drug	0.019
0-2014	
Number of related patents	0.081
Funder type of clinical trials	0.055
Duration of patent process	0.041
Allocation of clinical trials	0.024
Number of claims	0.023
Molecular weight of the drug	0.021
Polar surface area of the drug	0.021
Polarizability of the drug	0.020
Lipophilicity (logP) of the drug	0.019
Refractivity of the drug	0.019
0-2016	

Table 4.28 (cont.)	
Number of related patents	0.077
Funder type of clinical trials	0.054
Duration of patent process	0.039
Number of claims	0.024
Refractivity of the drug	0.022
Polar surface area of the drug	0.022
Molecular weight of the drug	0.021
Polarizability of the drug	0.021
Lipophilicity (logP) of the drug	0.021
Location of clinical trial	0.020
0-2018	
Number of related patents	0.066
Funder type of clinical trials	0.059
Duration of patent process	0.036
Number of claims	0.025
Allocation of clinical trials	0.023
Lipophilicity (logP) of the drug	0.022
Refractivity of the drug	0.021
Polar surface area of the drug	0.021
Location of clinical trial	0.020
Polarizability of the drug	0.020

APPENDIX C

PATENT-RELATED FEATURE DISTRIBUTIONS OF REGULATORILY APPROVED AND UNAPPROVED DRUGS

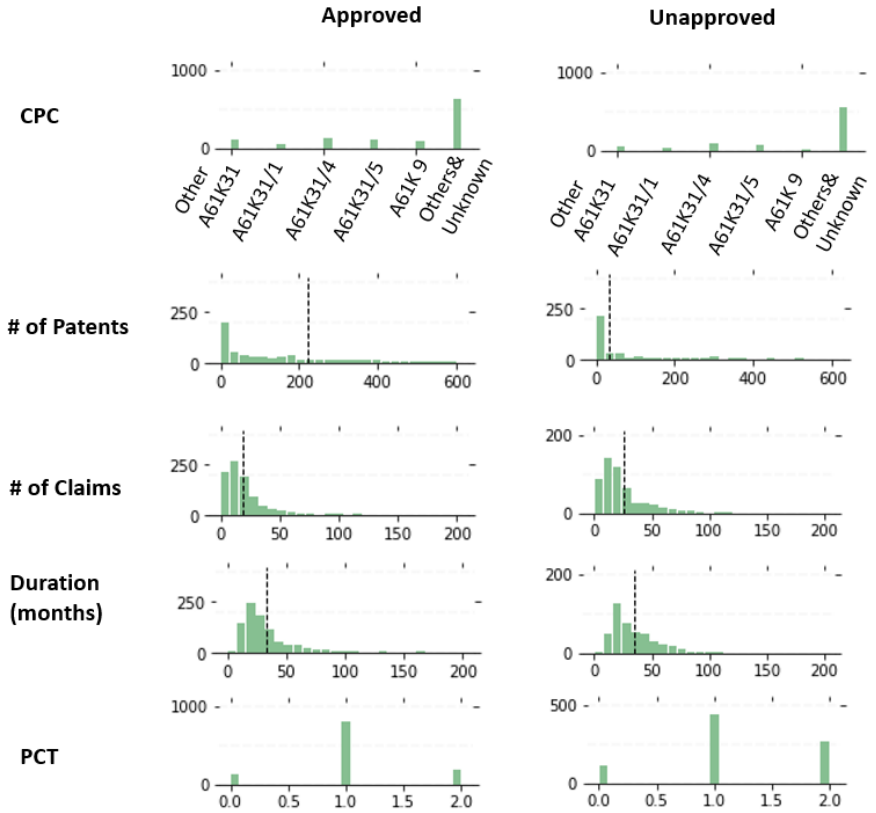


Figure 4.15. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Rare Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

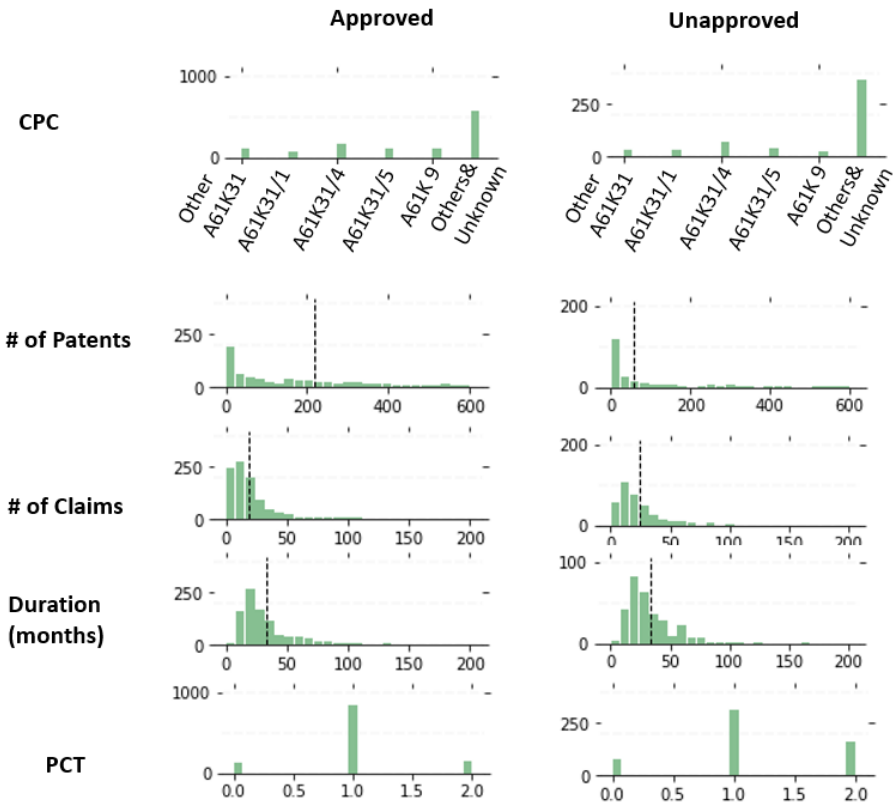


Figure 4.16. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Nervous Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

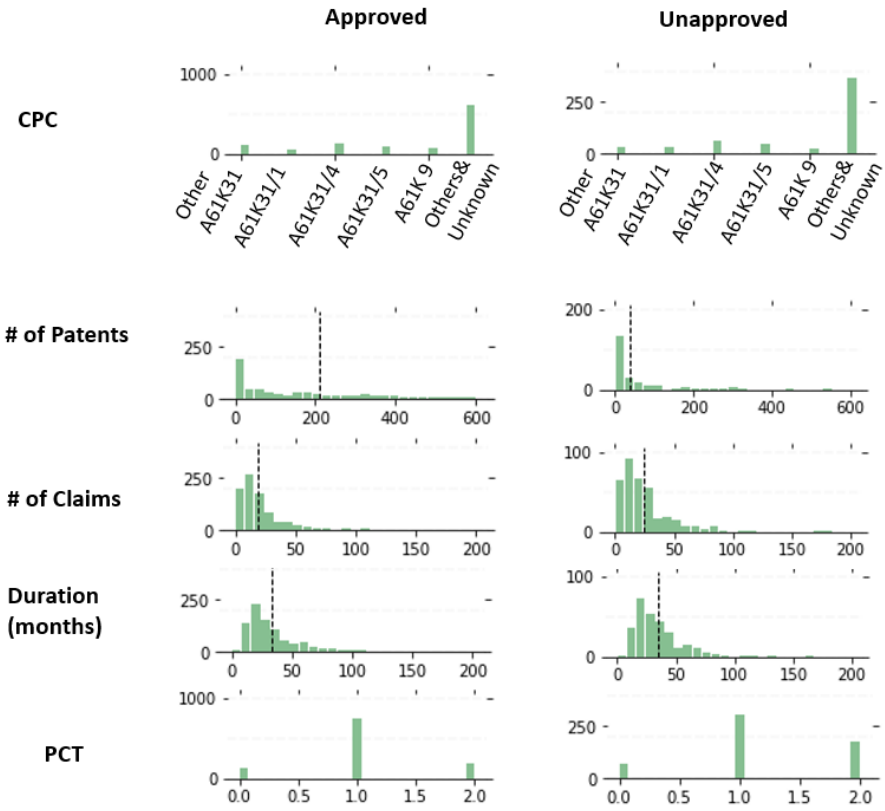


Figure 4.17. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Alimentary Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

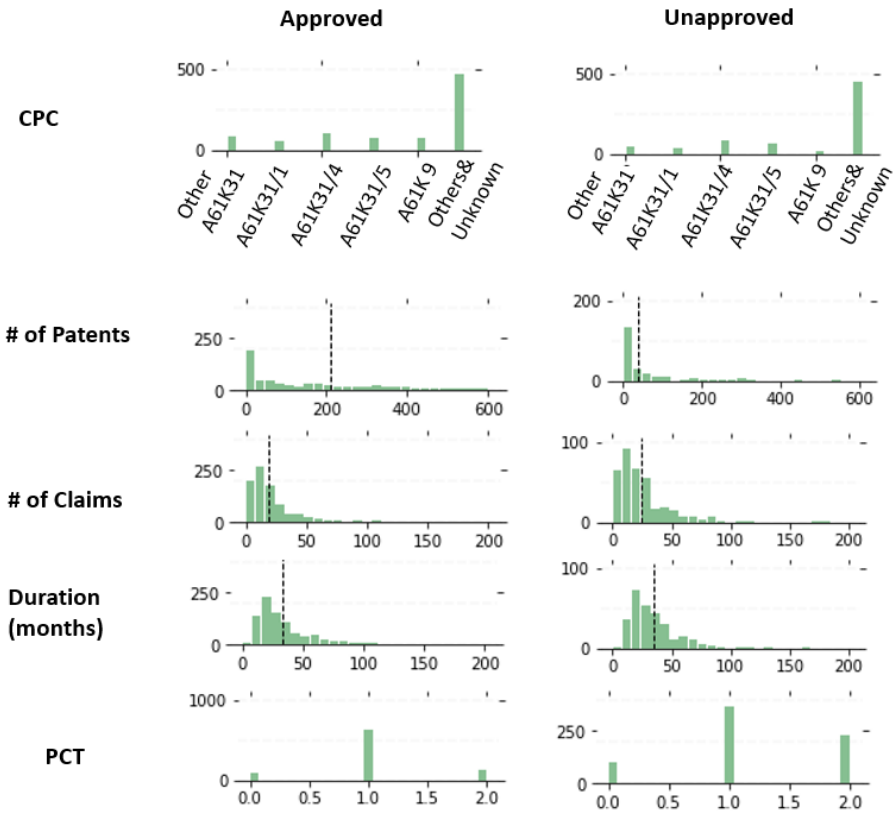


Figure 4.18. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Neoplasms" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

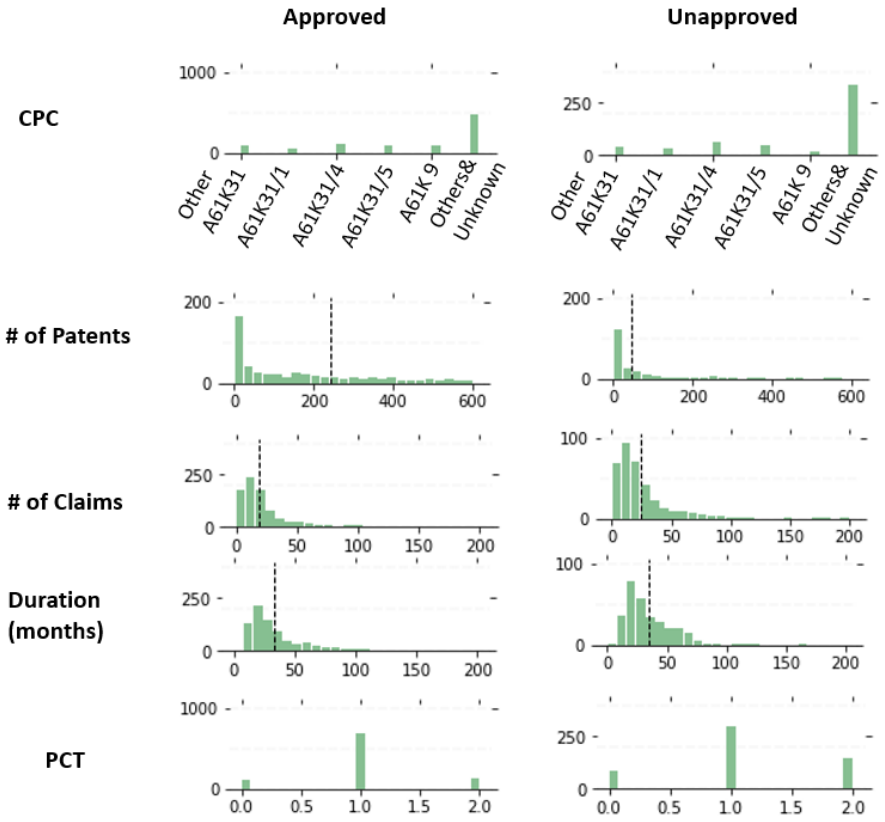


Figure 4.19. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Dermatological Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

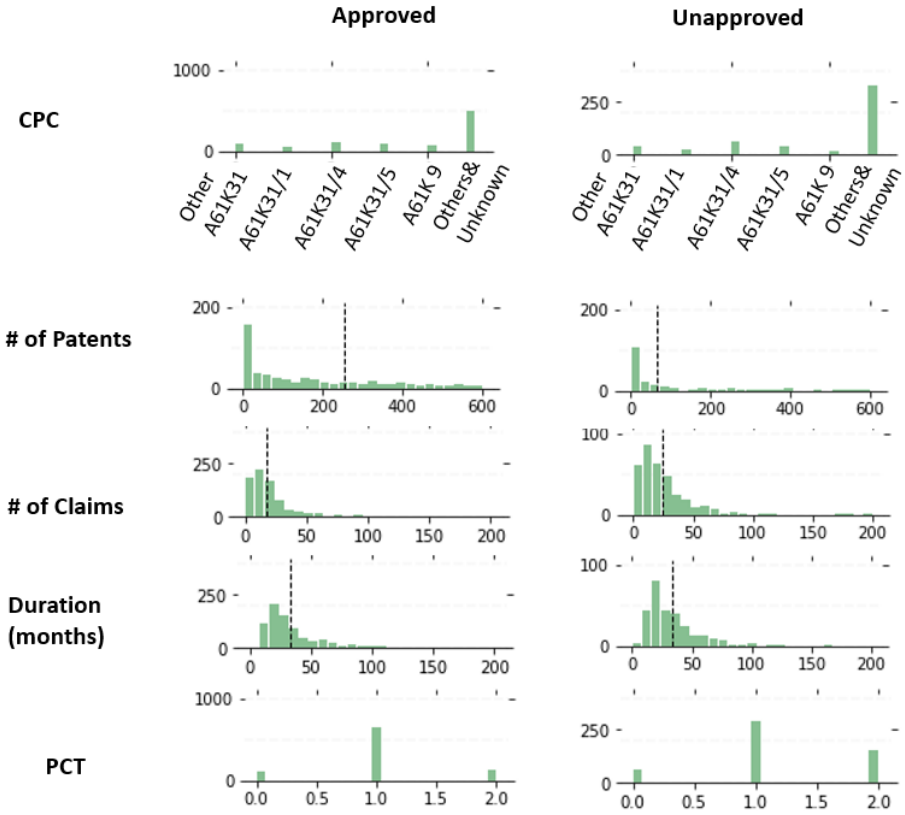


Figure 4.20. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Urinary Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

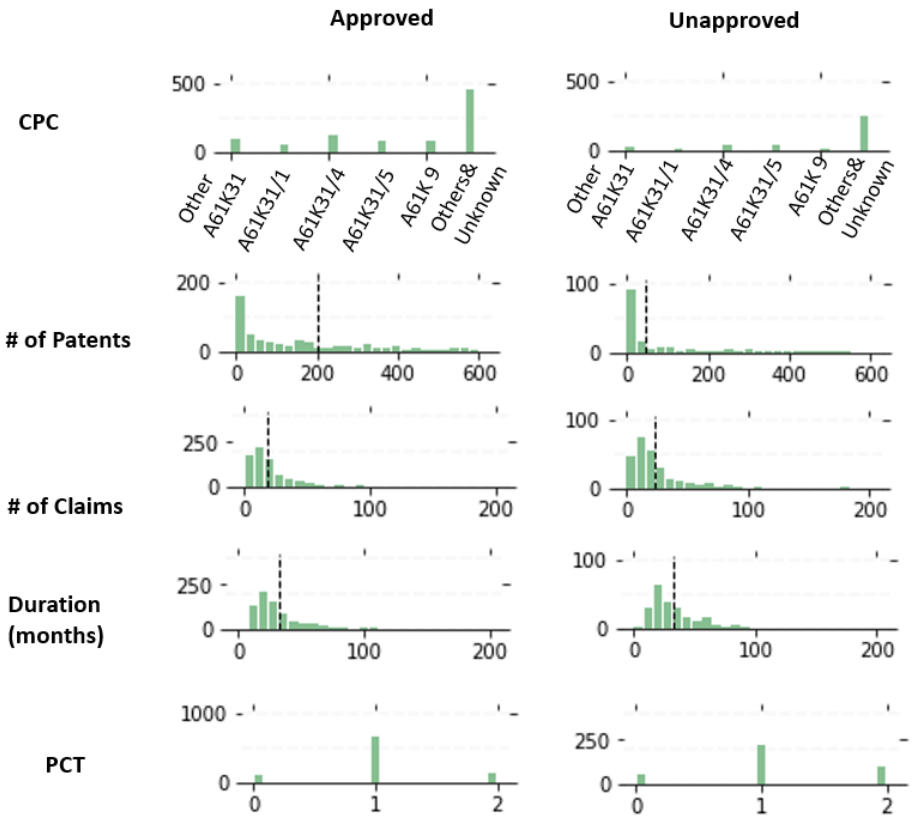


Figure 4.21. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Heart Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

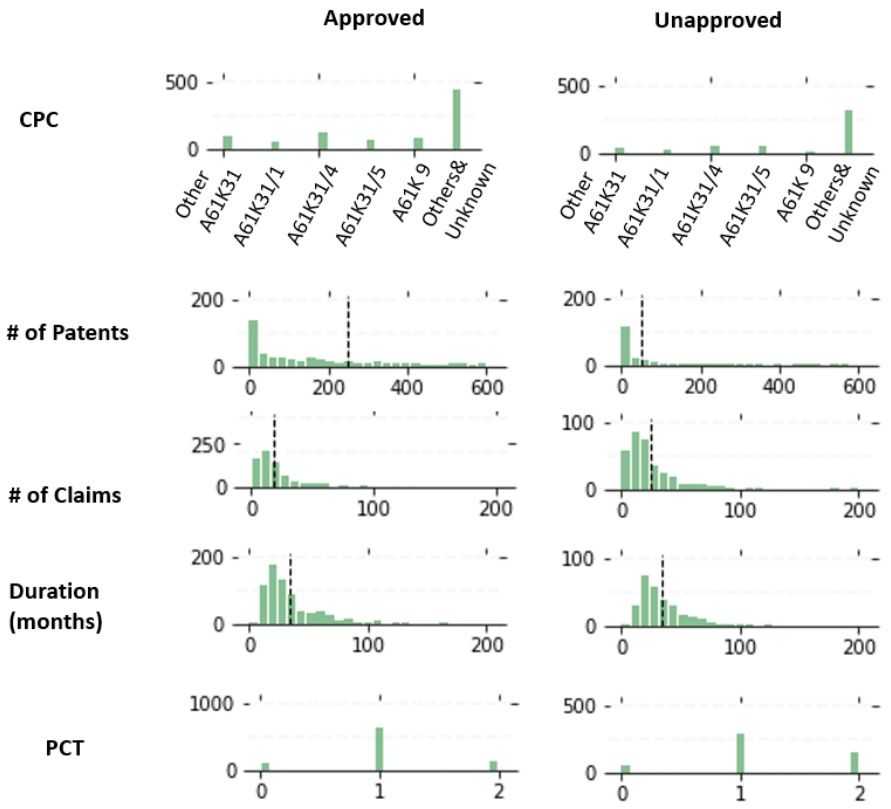


Figure 4.22. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Immunological Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

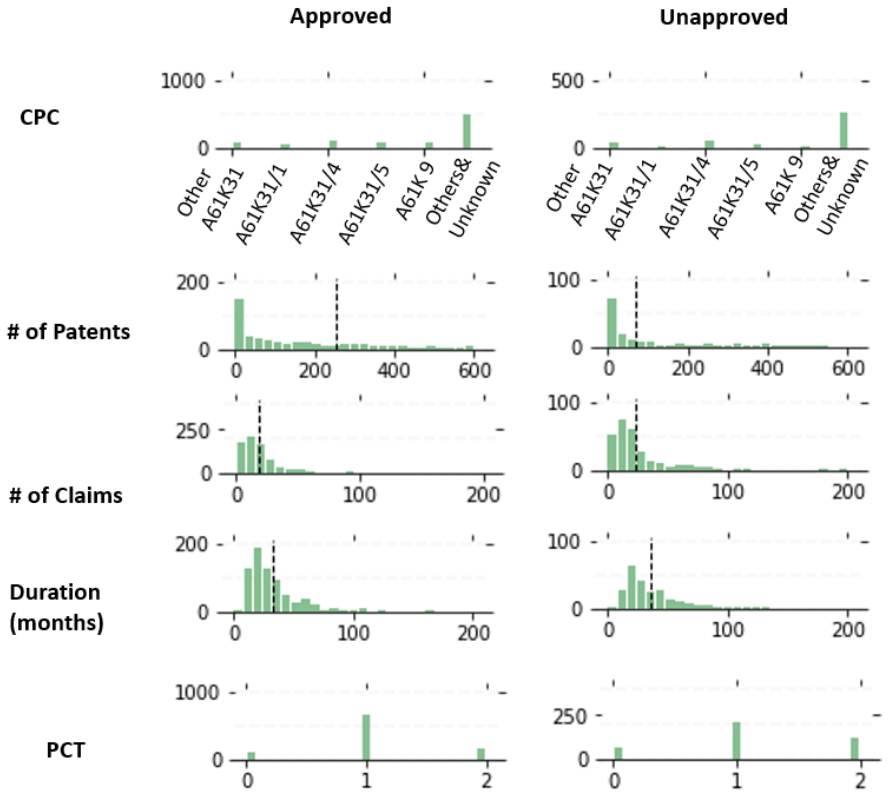


Figure 4.23. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Anti-infective Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

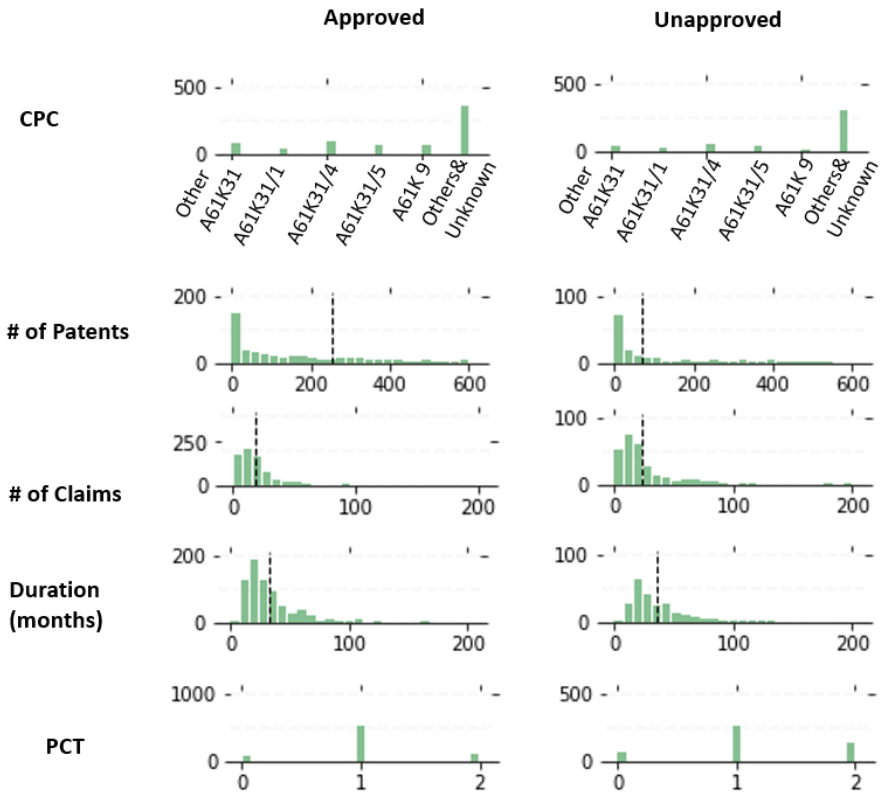


Figure 4.24. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Respiratory Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

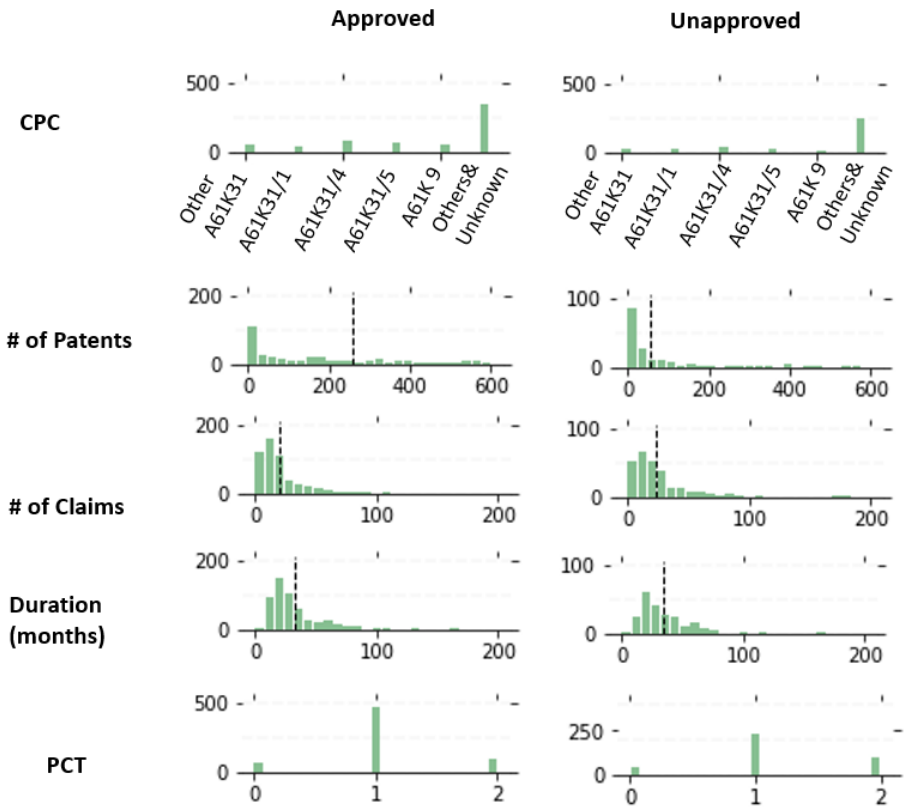


Figure 4.25. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Hormonal Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

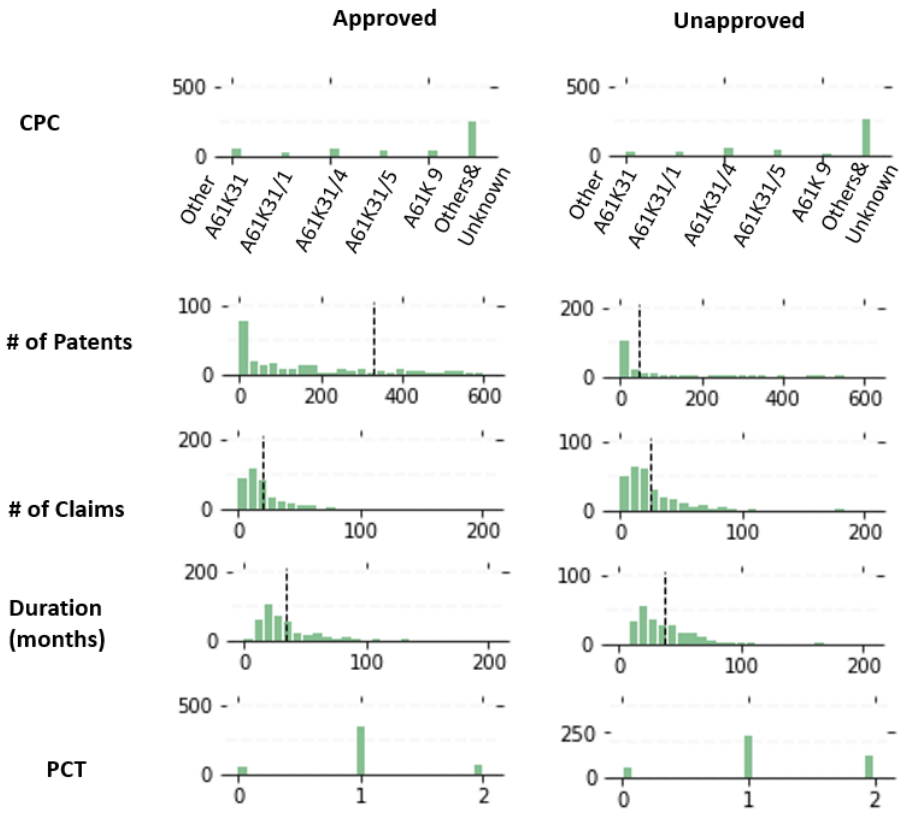


Figure 4.26. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Blood Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

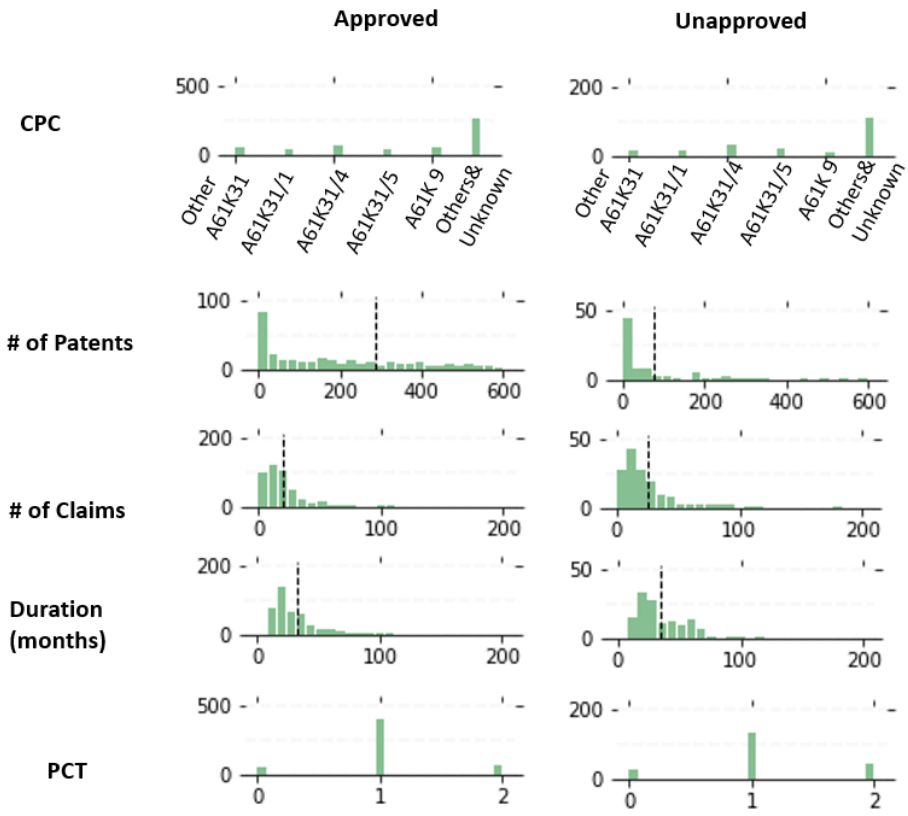


Figure 4.27. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Musculoskeletal Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

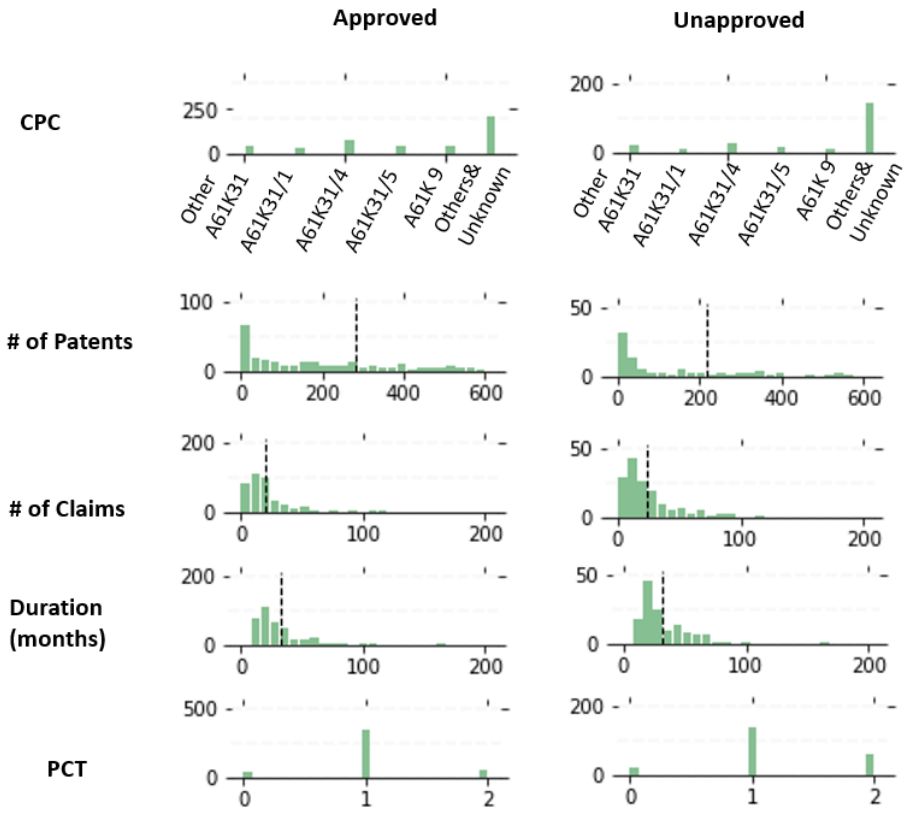


Figure 4.28. Patent-related feature distributions of regulatorily approved and unapproved drugs for the "Sensory Diseases" class. Abbreviations; #: number, CPC: Cooperative Patent Classification, PCT: the presence of Patent Cooperation Treaty application, PCT (#0): no PCT application, PCT (#1): presence of PCT application, PCT (#2): unknown.

APPENDIX D

CLINICAL TRIAL-RELATED FEATURE DISTRIBUTIONS OF REGULATORILY APPROVED AND UNAPPROVED DRUGS

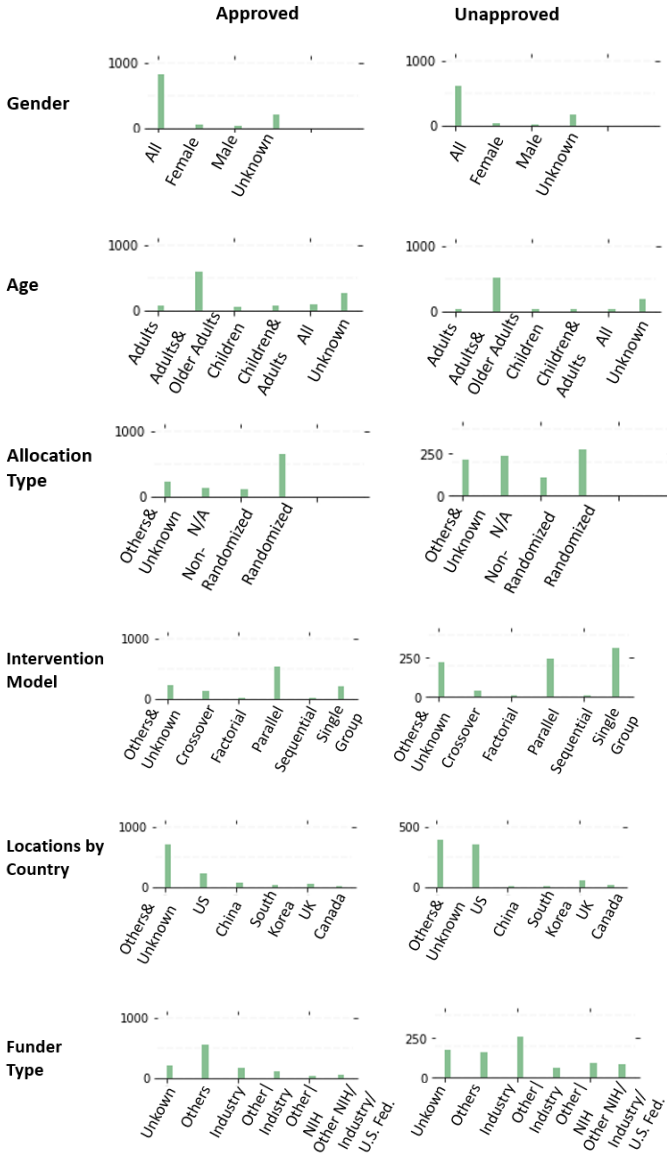


Figure 4.29. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the "Rare Diseases" class.

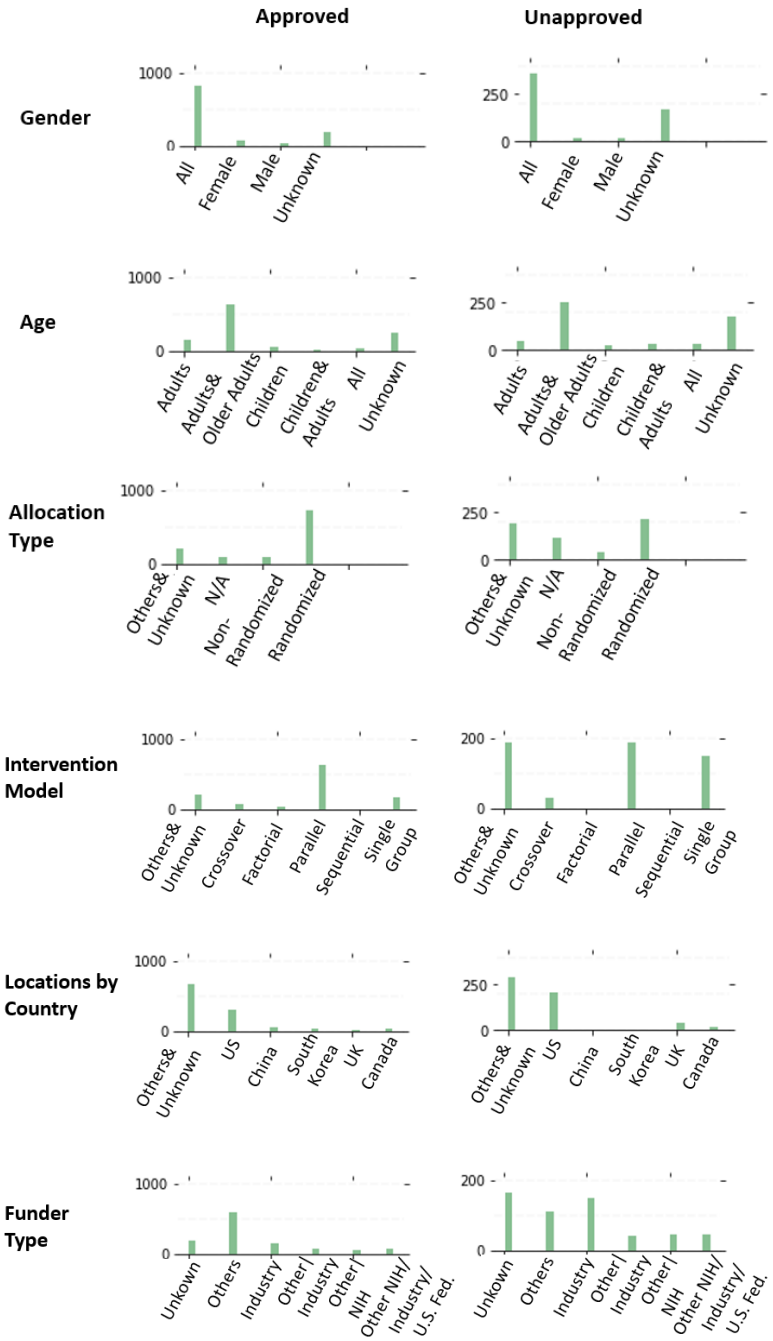


Figure 4.30. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the "Nervous Diseases" class.

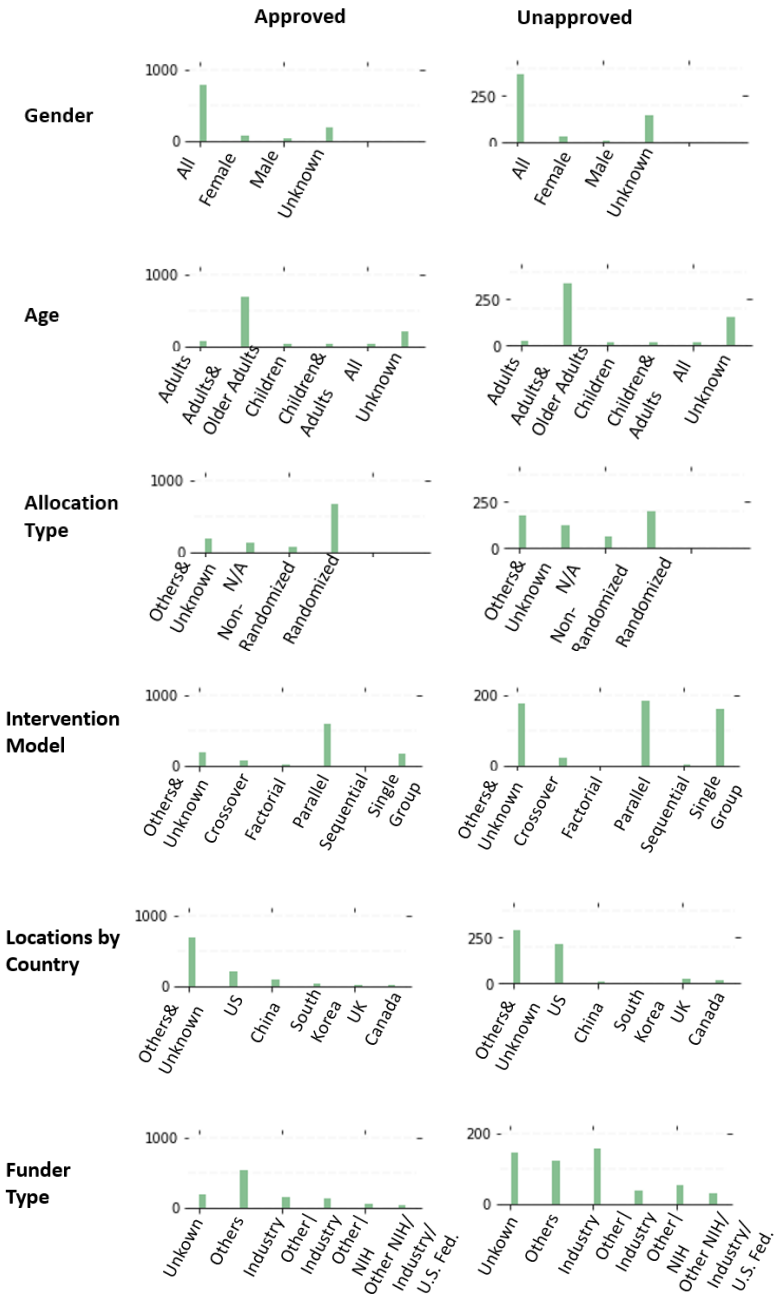


Figure 4.31. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the "Alimentary Diseases" class.

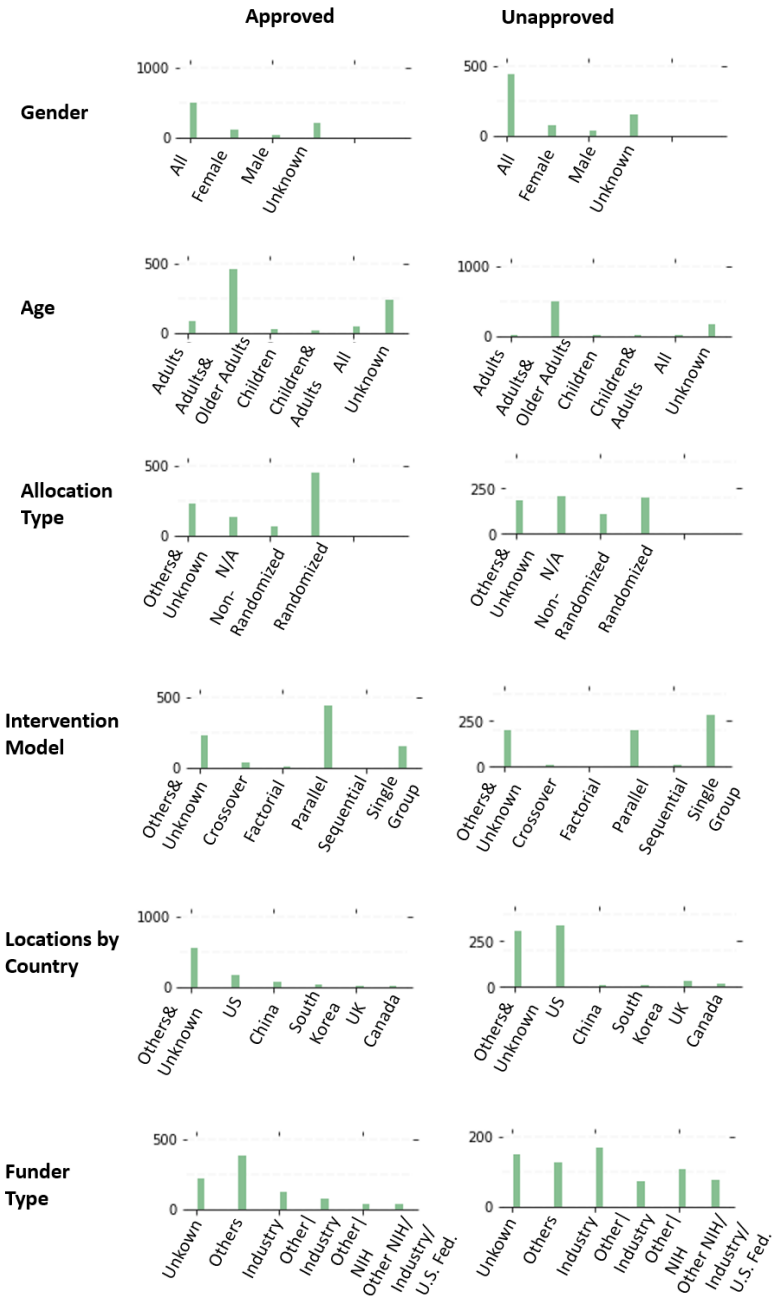


Figure 4.32. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Neoplasms” class.

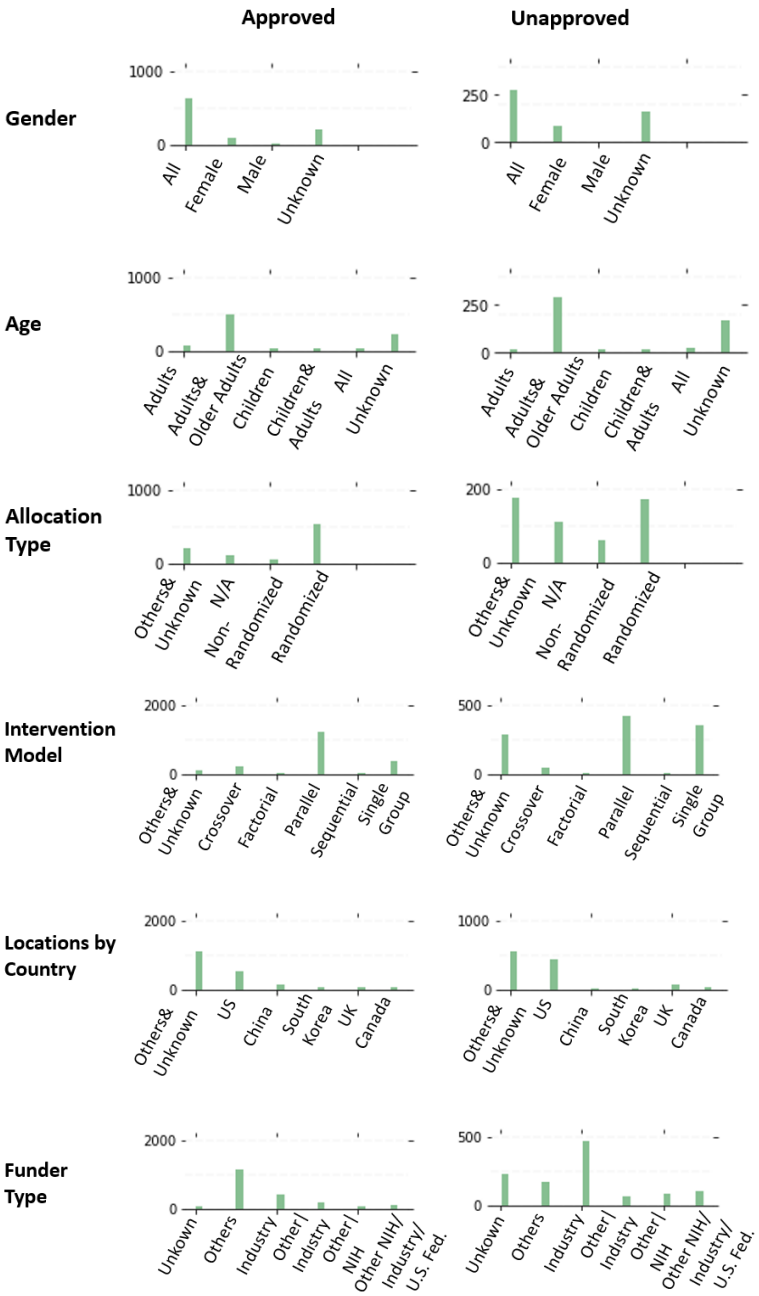


Figure 4.33. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Dermatological Diseases” class.

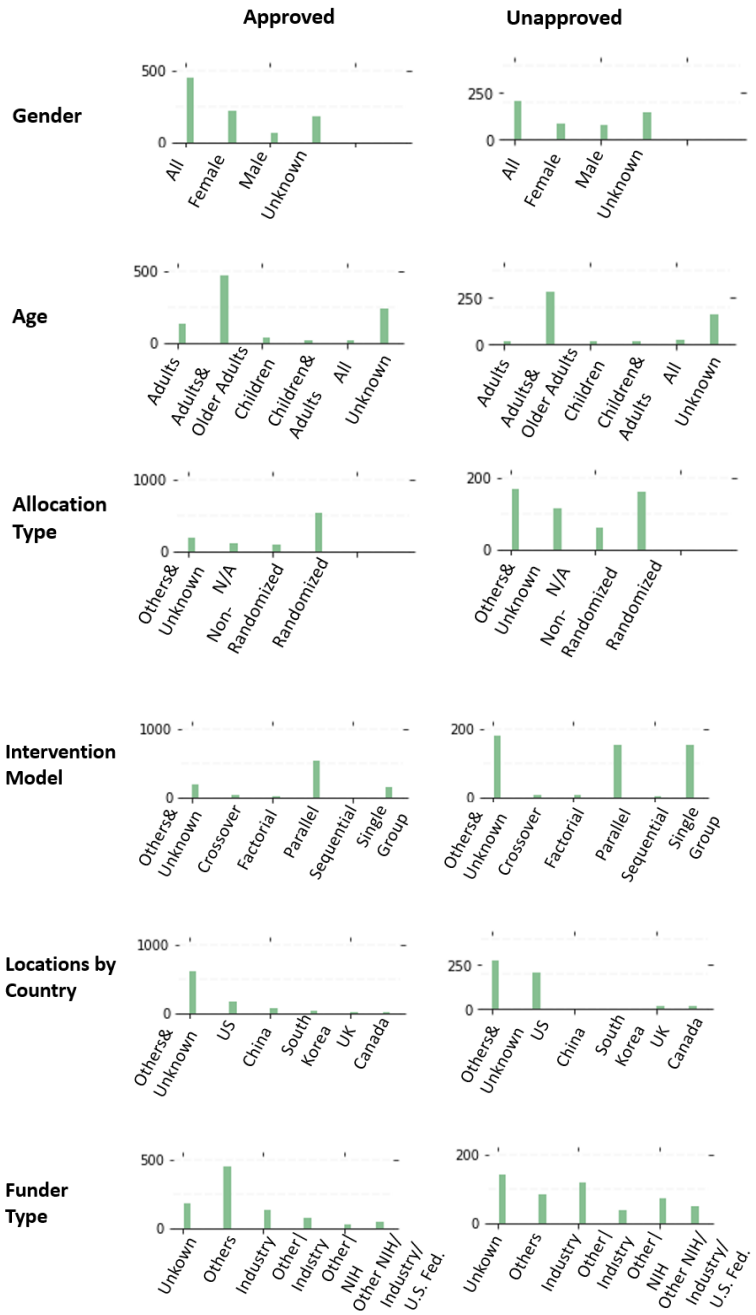


Figure 4.34. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the "Urinary Diseases" class.

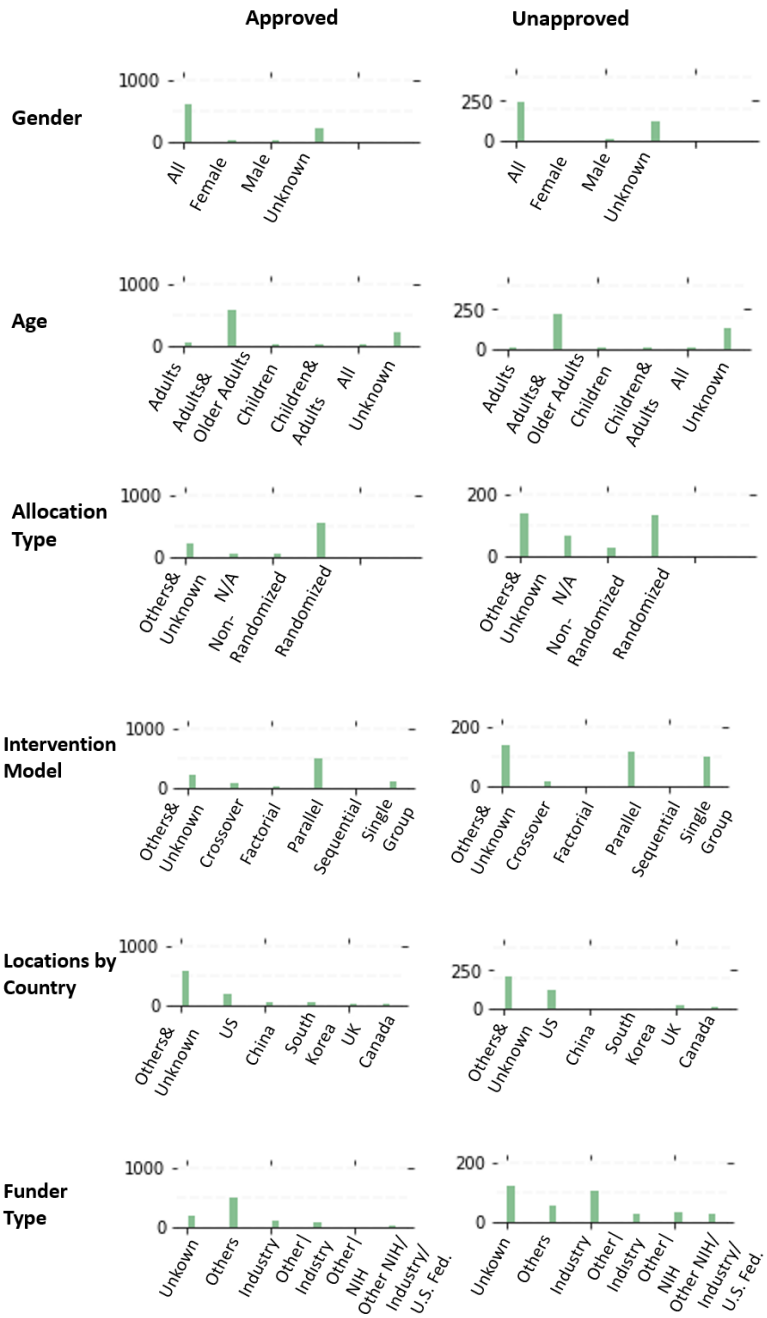


Figure 4.35. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Heart Diseases” class.

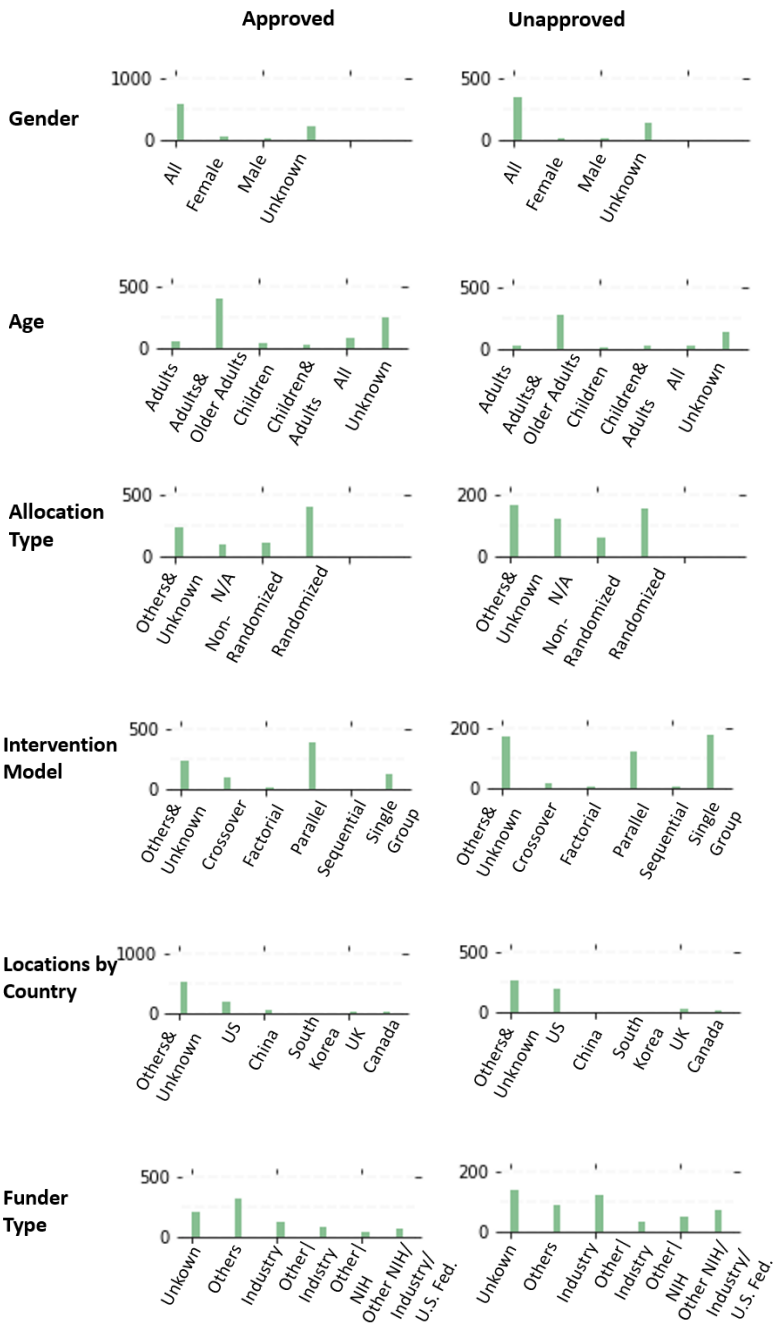


Figure 4.36. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Immunological Diseases” class.

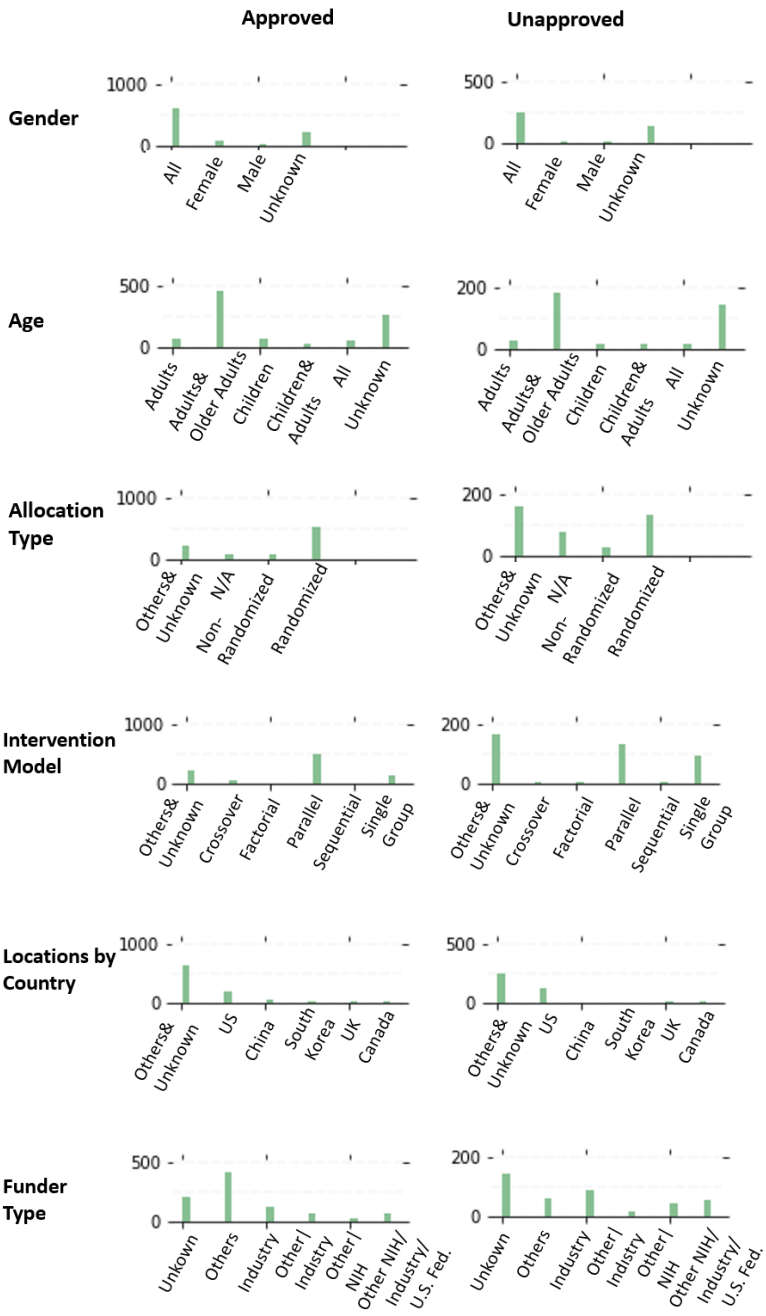


Figure 4.37. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Anti-infective Diseases” class.

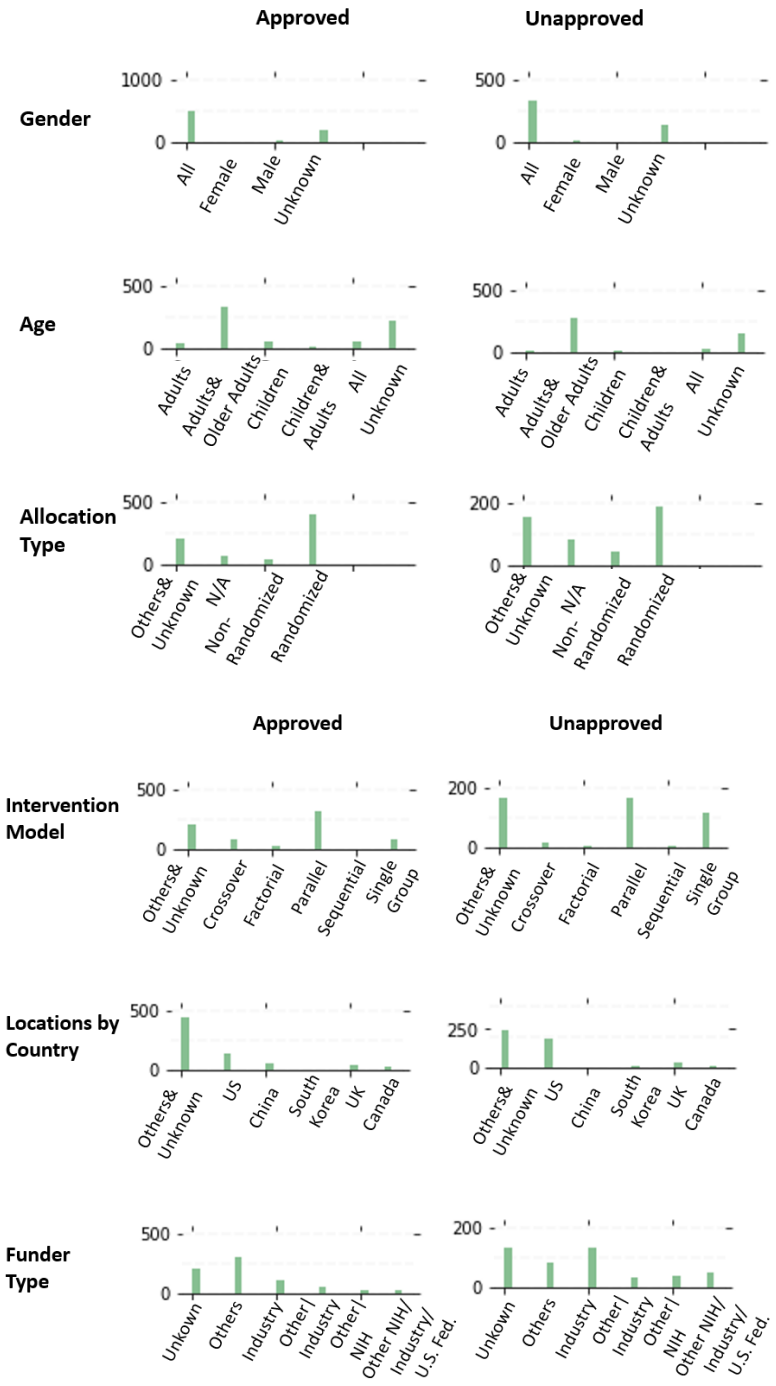


Figure 4.38. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Respiratory Diseases” class.

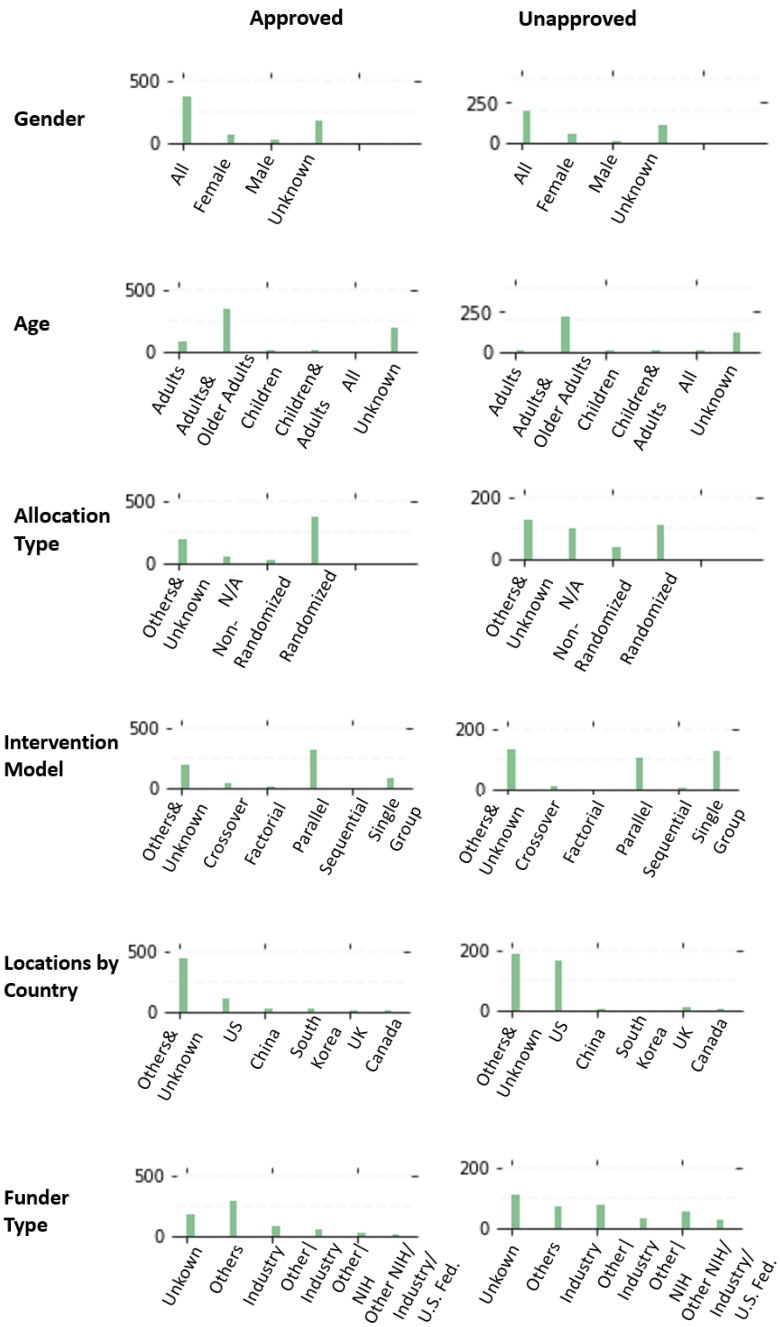


Figure 4.39. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the "Hormonal Diseases" class.

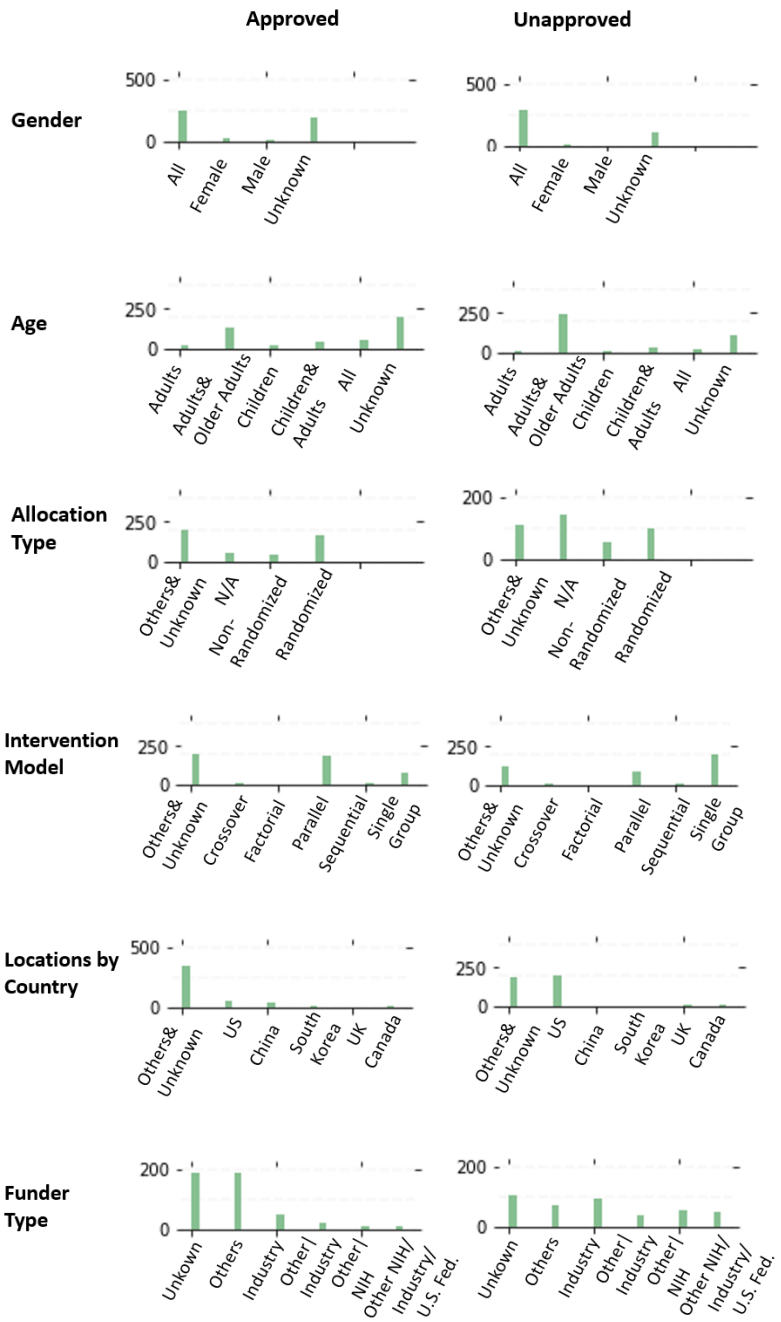


Figure 4.40. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Blood Diseases” class.

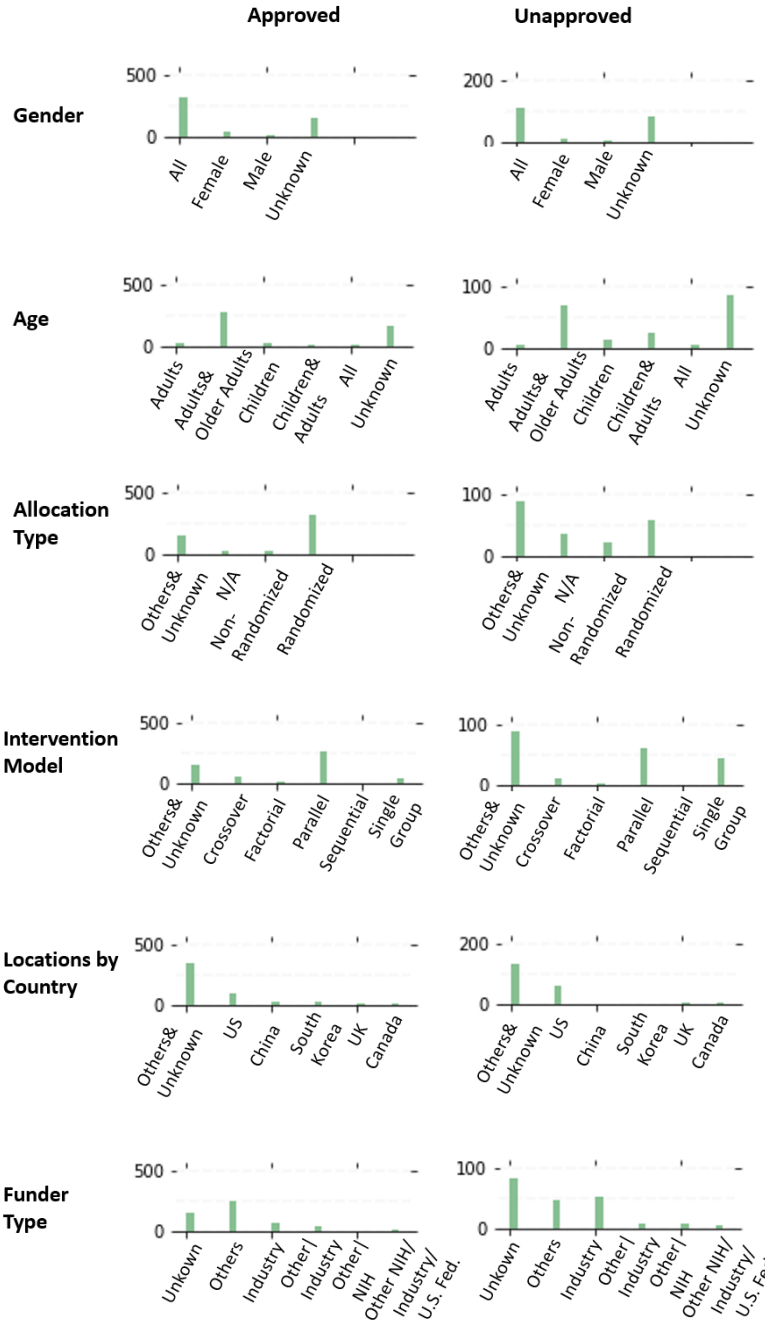


Figure 4.41. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the "Musculoskeletal Diseases" class.

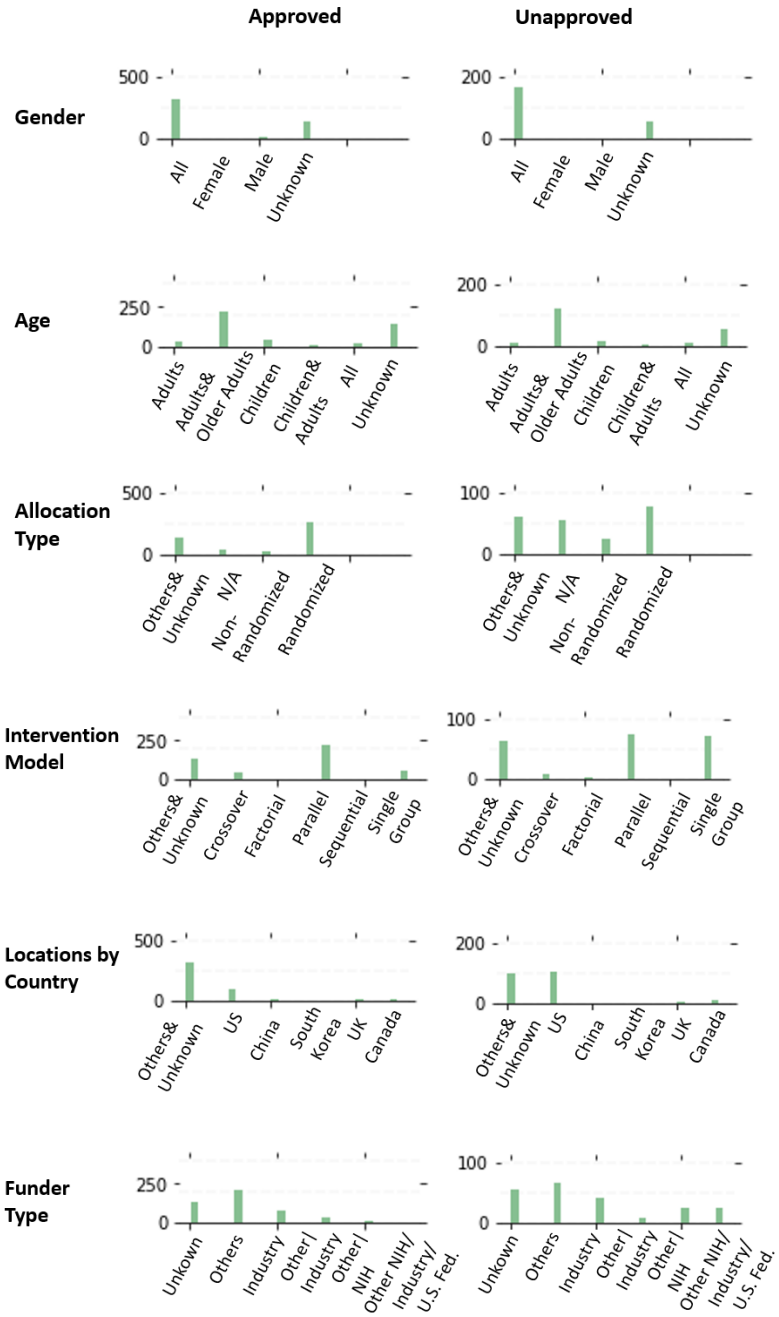


Figure 4.42. Clinical trial-related feature distributions of regulatorily approved and unapproved drugs for the “Sensory Diseases” class.

APPENDIX E

PHYSICOCHEMICAL FEATURE DISTRIBUTIONS OF REGULATORILY APPROVED AND UNAPPROVED DRUGS

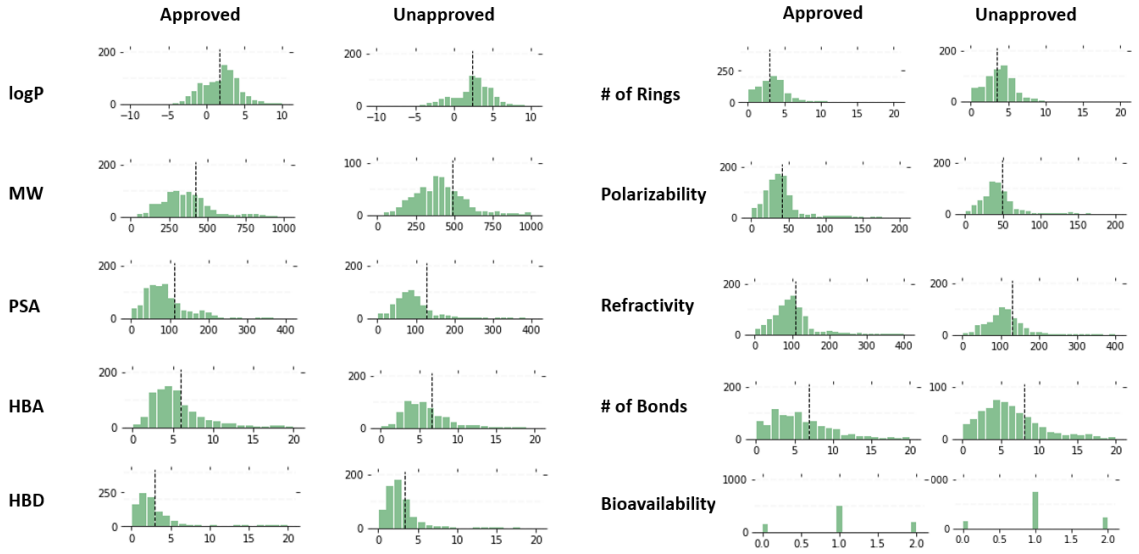


Figure 4.43. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Rare Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

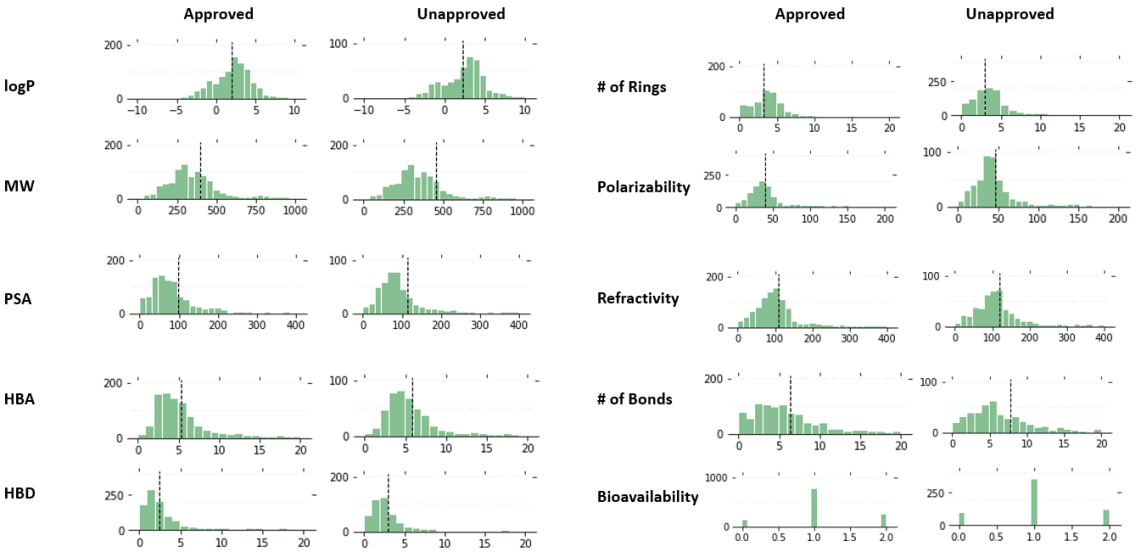


Figure 4.44. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Nervous Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

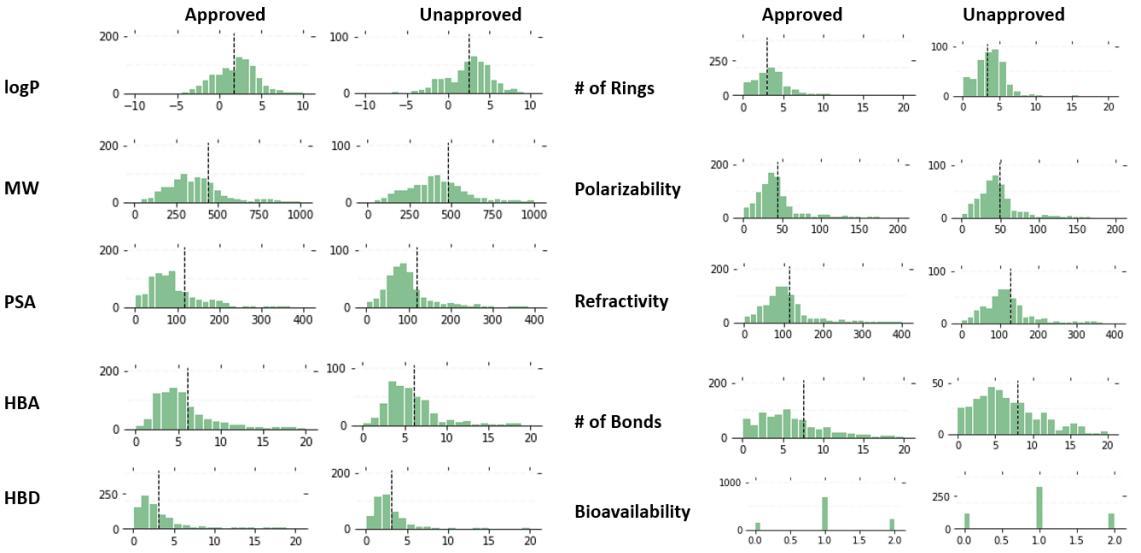


Figure 4.45. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Alimentary Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen

bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

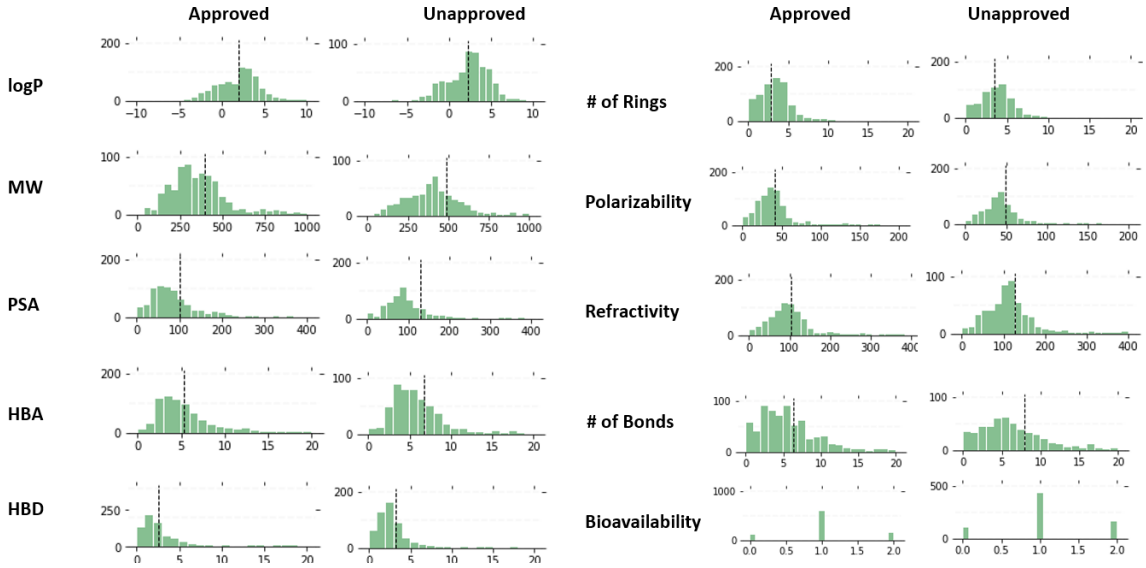


Figure 4.46. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Neoplasms" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

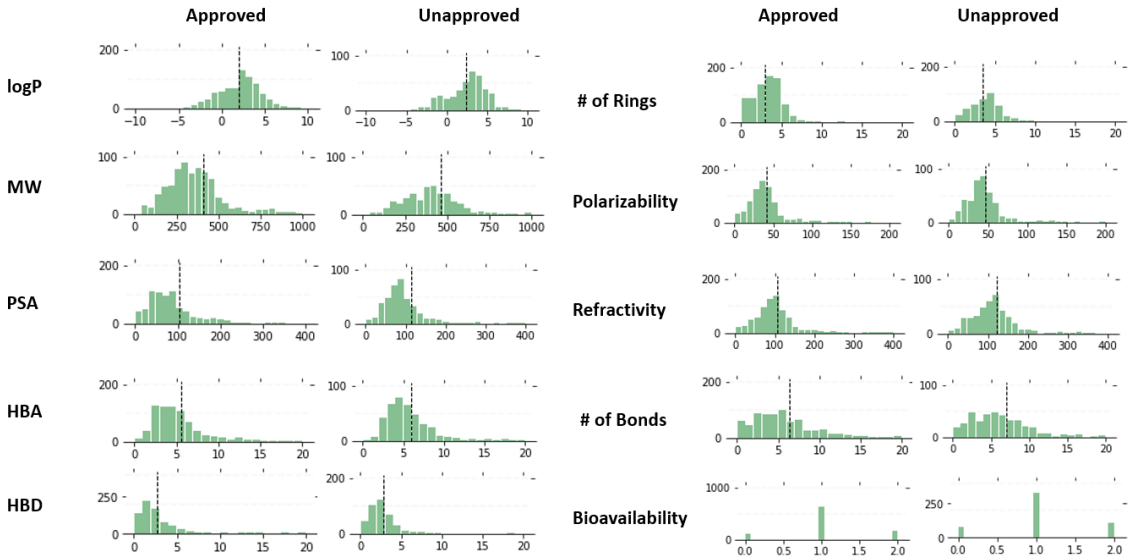


Figure 4.47. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Dermatological Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

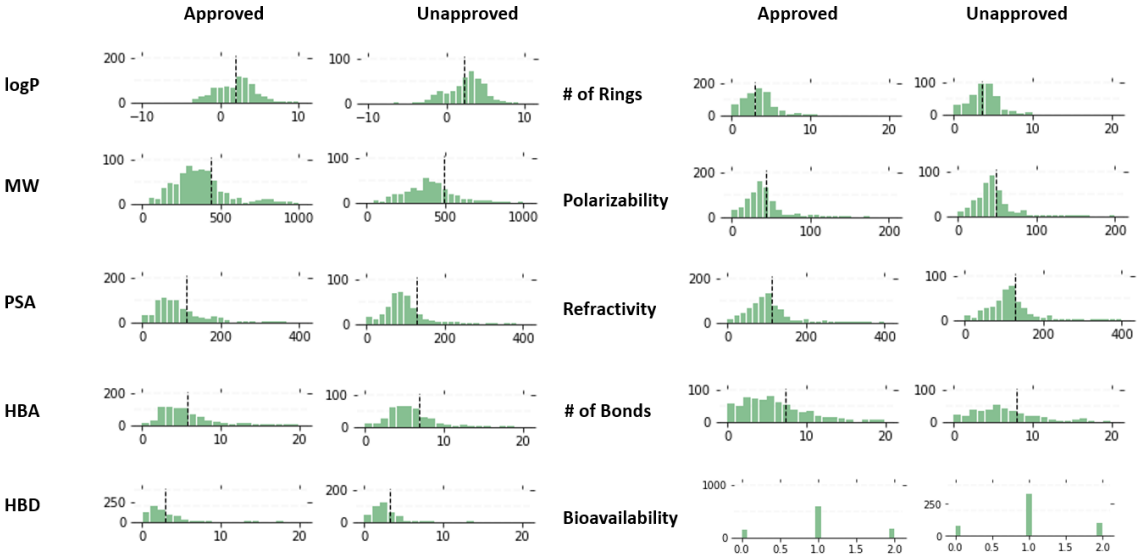


Figure 4.48. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Urinary Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen

bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

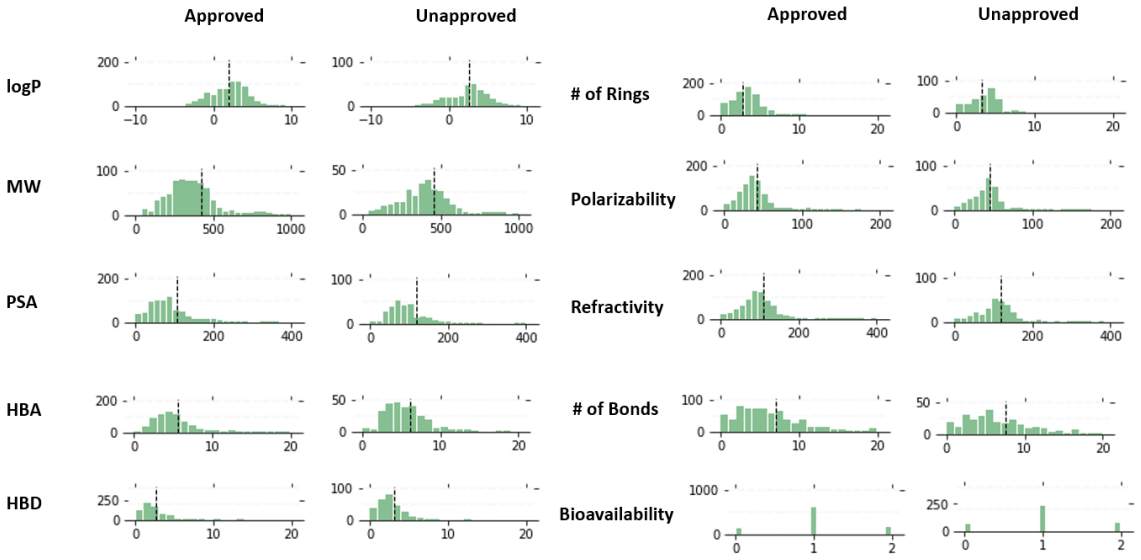


Figure 4.49. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Heart Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

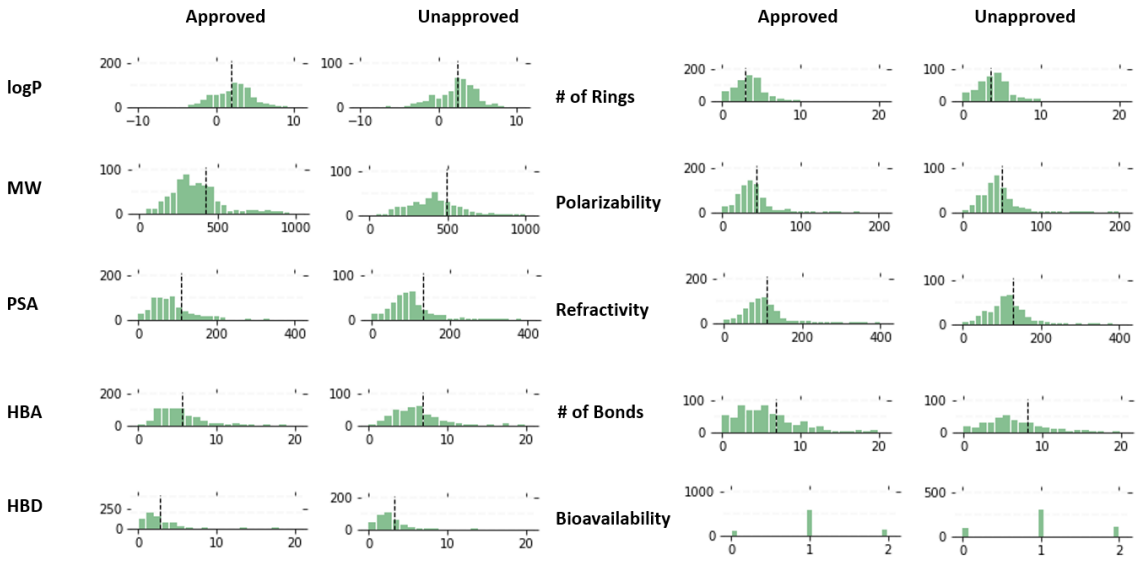


Figure 4.50. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Immunological Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

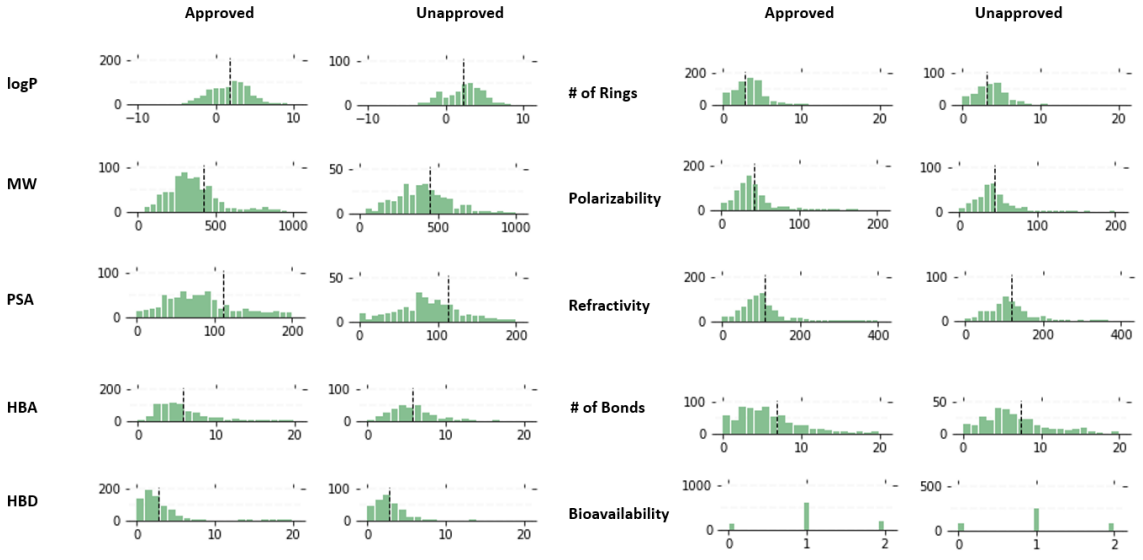


Figure 4.51. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Anti-infective Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA:

hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

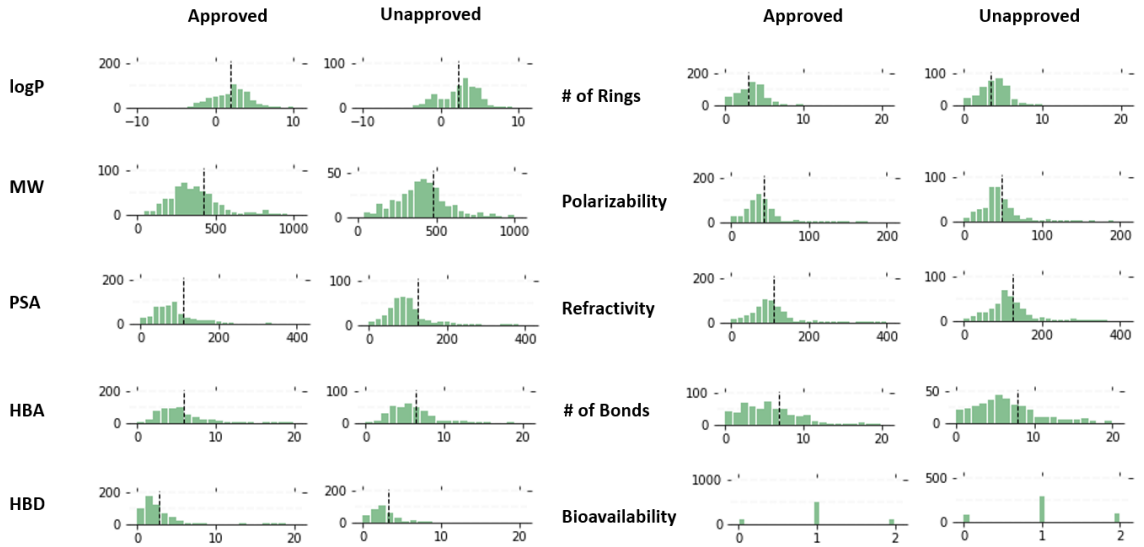


Figure 4.52. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Respiratory Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

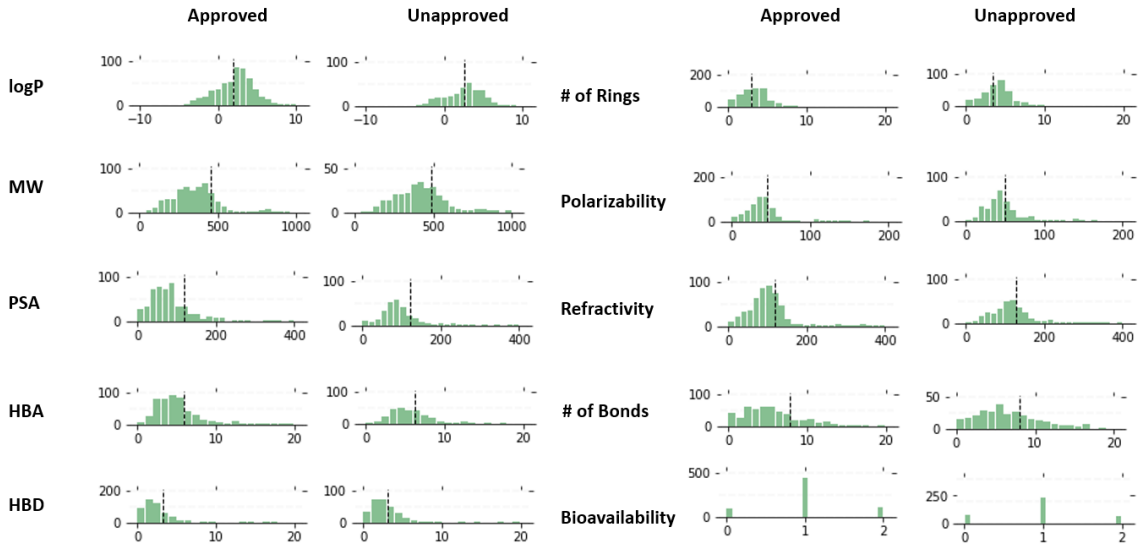


Figure 4.53. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Hormonal Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

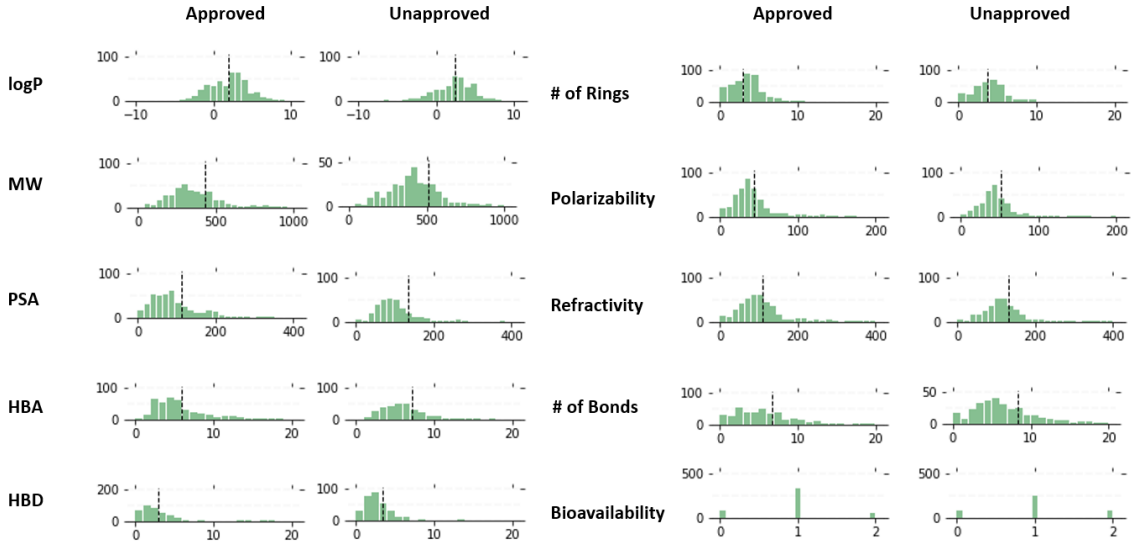


Figure 4.54. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Blood Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

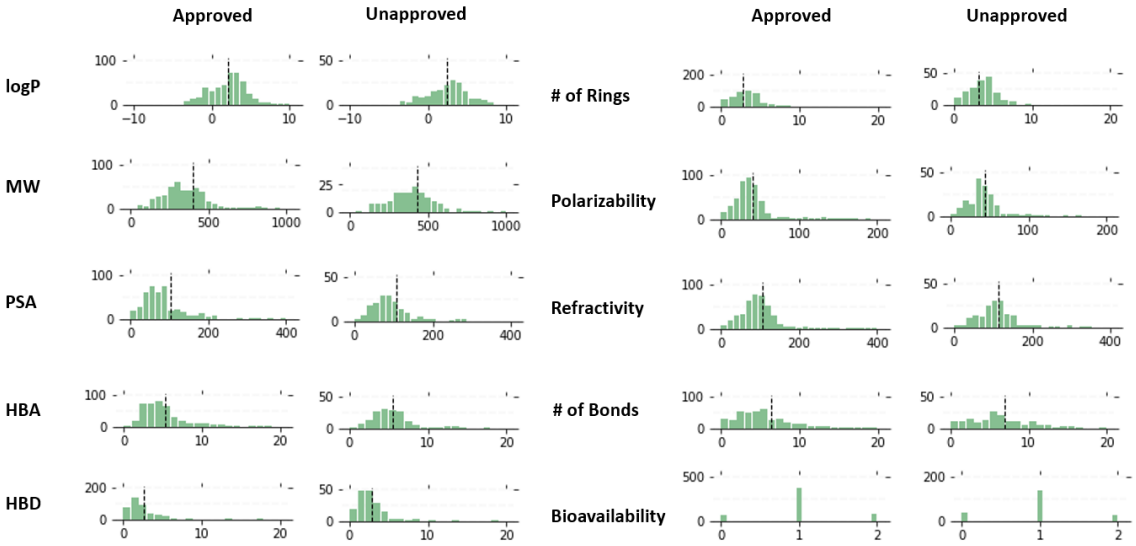


Figure 4.55. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Musculoskeletal Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

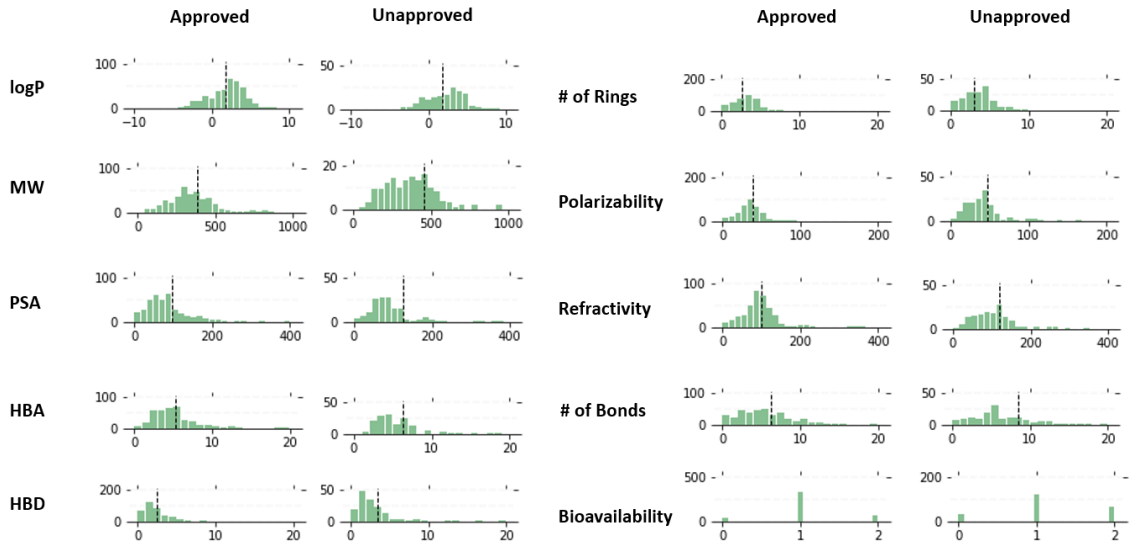


Figure 4.56. Physicochemical feature distributions of regulatorily approved and unapproved drugs for the "Sensory Diseases" class. Abbreviations; #: number, logP: lipophilicity measure, MW: molecular weight, PSA: polar surface area, HBA: hydrogen

bond acceptor, HBD: hydrogen bond donor, Bioavailability (#0): predicted bioavailability as 0, Bioavailability (#1): predicted bioavailability as 1, Bioavailability (#2): unknown. Dashed line corresponds to the mean value for numerical features.

CURRICULUM VITAE

Fulya ÇIRAY

Nationality : Turkish
Date of Birth : 15.02.1988
Place of Birth : Eskişehir, Turkey
Email : fulyaciray@gmail.com

EDUCATION

PhD Program — **Middle East Technical University**
02.2016 - 09. 2022 — Medical Informatics
(Cumulative GPA: 4.00 over 4.00)

MSc Program — **Heidelberg University (Germany)**
10.2011 - 07.2013 — Molecular and Cellular Biology
(Cumulative GPA: 3.67 over 4.00)

BSc Program — **Middle East Technical University**
09.2006 - 06.2011 — Molecular Biology and Genetics
(Cumulative GPA: 3.89 over 4.00)

WORK EXPERIENCE

Year	Place	Enrollment
2015-Present	TURKPATENT	Patent Examiner
2014-2015	METU, *Genetics*	Research and Teaching Assistant
2012-2013	Heidelberg University, *Molecular and Cellular Biology*	Research Assistant

FOREIGN LANGUAGES

	Reading	Writing	Speaking
English	Advanced	Advanced	Advanced
German	Intermediate	Intermediate	Intermediate

SCHOLARSHIPS

2014-2020: TÜBİTAK (2211-A)

2011-2013: HBIGS (Hartmut Hoffmann-Berling Molecular and Cellular Biology Program)

2006-2011: TÜBİTAK (2205)

PUBLICATIONS

Hideki Yokoyama, Birgit Koch, Rudolf Walczak, Fulya Ciray-Duygu, Juan Carlos Gonzales-Sanchez, Damien P. Devos, Iain W. Mattaj, Oliver J. Gruss (2014). The nucleoporin MEL-28 promotes RanGTP-dependent γ-tubulin recruitment and microtubule nucleation in mitotic spindle formation. *Nature Communications*, 5, 3270.

TEZ İZİN FORMU / THESIS PERMISSION FORM

ENSTİTÜ / INSTITUTE

Fen Bilimleri Enstitüsü / Graduate School of Natural and Applied Sciences ☐

Sosyal Bilimler Enstitüsü / Graduate School of Social Sciences ☐

Uygulamalı Matematik Enstitüsü / Graduate School of Applied Mathematics ☐

Enformatik Enstitüsü / Graduate School of Informatics ☐

Deniz Bilimleri Enstitüsü / Graduate School of Marine Sciences ☐

YAZARIN / AUTHOR

Soyadı / Surname : .....
Adı / Name : .....
Bölümü / Department : .....

TEZİN ADI / TITLE OF THE THESIS (**İngilizce** / English) : .....
.....
.....
.....
.....

TEZİN TÜRÜ / DEGREE**:** **Yüksek Lisans** / Master ☐ **Doktora** / PhD ☐

1. **Tezin tamamı dünya çapında erişime açılacaktır. /** Release the entire work immediately for access worldwide. ☐
2. **Tez iki yıl süreyle erişime kapalı olacaktır.** / Secure the entire work for patent and/or proprietary purposes for a period of **two year. *** ☐
3. **Tez altı ay süreyle erişime kapalı olacaktır.** / Secure the entire work for period of **six months**. ***** ☐

***** *Enstitü Yönetim Kurulu Kararının basılı kopyası tezle birlikte kütüphaneye teslim edilecektir.*
A copy of the Decision of the Institute Administrative Committee will be delivered to the library together with the printed thesis.

Yazarın imzası / Signature ..... **Tarih** / Date .....