Predicting Clinical Trial Outcomes Using Drug Bioactivities through Graph Database Integration and Machine Learning

Vidhya Murali[1¥], Pradyumna YM[2¥], Cassandra Königs[3], Meera Nair[4], Sethulekshmi[4], Prema Nedungadi[5], Gowri Srinivasa[2], Prashanth Athri[1*]

[1]Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Bengaluru, India

[2]PES Center for Pattern Recognition, Department of Computer Science and Engineering, PES University, Bengaluru, India

[3]Bioinformatics and Medical Informatics, Bielefeld University, Northrhine-Westphalia, Germany

[4]Amrita School of Biotechnology, Amritapuri, Amrita Vishwa Vidyapeetham, Kerala, India

[5]Department of Computer Science and Engineering, Amrita School of Engineering, Amritapuri, Amrita Vishwa Vidyapeetham, Kerala, India

[¥]Contributed equally

* Email: prashanthathri@gmail.com

Abstract

The ability to estimate the probability of a drug to receive approval in clinical trials provides natural advantages to optimizing pharmaceutical research workflows. Success rates of clinical trials have deep implications to costs, duration of development, and under pressure due to stringent regulatory approval processes. We propose a machine learning approach that can predict the outcome of the trial with reliable accuracies, using biological activities, physicochemical properties of the compounds, target-related features, and NLP-based compound representation. Biological activities have never been used as a predictive feature. We have extracted the drug-disease pair from clinical trials and mapped target(s) to that pair using multiple data sources. Empirical results demonstrate that ensemble learning outperforms independently trained, small-data ML models. We report results and inferences derived from a Random forest classifier with an average accuracy of 93%, and an F1 score of 0.96 for the 'Pass' class. 'Pass' refers to one of the two classes (Pass/ Fail) of all clinical trials and the model performed well in predicting the 'Pass' category. An analysis of the features demonstrates that bioactivity plays an important role in predicting the outcome of a clinical trial. A significant effort has gone into the production of the dataset that, for the first time, integrates clinical trial information with protein targets. All code to map these entities is available through this study, and all data are from publicly available sources. While our model identifies low-lying inferences when biological activities are included, the code to integrate biological activity and target information provide researchers with access to deep curated and proprietary clinical trial databases the ability to get deeper insights, better statistical significance, and capabilities to better predict trial failures.

KEYWORDS

bioactivity, clinical trial, data integration, ensemble algorithms, graph database, machine learning

1 INTRODUCTION

Clinical trials evaluate the efficacy and toxicity of a drug candidate. A molecule is released as a new drug after success in all four phases of clinical trials. While treatments come in many forms like multi-drug therapies, medical devices, vaccines, biologics, and others, here, we build a predictive model to estimate the clinical trial success of single-molecule therapeutics. Trials determine if pre-clinical stage molecules can metamorphose into a therapeutic, and form a large part of the budget required to transform a potential active molecule to a New Drug Application (NDA). Hence, the ability to predict clinical trial outcomes derisk R&D costs, assist in the early prioritization of active molecules, and reduce attrition rates in drug approval.

Source Code used in this study can be found at:
https://github.com/pathri/ClinicalTrials_Prediction

High costs in drug development are due to multiple complex contributors like scientific discovery processes, infrastructure, collaborations, clinical development duration, safety studies, efficacy studies, market research, and clinical trial costs, to name a few (Maxmen, 2016). Conducting clinical trials amidst knowing the efficacy and value of drugs, regulatory scrutiny, population, rising costs remains a core challenge for pharma (Martin et al., 2017, Lang & Siribaddana, 2012). Starting with a big positive shift away from the dip in approvals between 2015-17 (Dowden & Munro, 2019), the data published by FDA (the annual New Drug Approvals Report) suggest that an average of 46 drugs have been approved per year across the past 5 years. Nonetheless, the success rate of molecules to transition from Phase II to III is the least (Yamaguchi et al., 2021), and hence exemplifies the phase transition that carries the highest risk. Further, studies (Dowden & Munro, 2019) indicate that there are higher chances of a molecule being launched as a drug when it reaches Phase III.

1.1 Machine Learning to predict Clinical Trial Outcomes

Recent advancements in Artificial Intelligence (AI) and Machine Learning (ML) will modernize and transform clinical development because of data availability and the computational power that can efficiently handle this data (Shah et al., 2019). Specifically, studies using ML and quantitative modeling have been widely applied to clinical trials with varying goals and using diverse types of independent predictive variables. However, the annotation of clinical trial datasets is highly time-consuming and require domain expertise, due to which nearly all outcome prediction studies use commercially available, annotated clinical trial datasets (Lo et al., 2018, Feijoo et al., 2020, Zhavoronkov et al., 2020, Wong et al., 2019, Vergetis et al., 2020, Yamaguchi et al., 2021). A labyrinth set of studies evaluate various factors that affect trial outcomes and span a wide set of questions (see Feijoo et al., 2020 for a detailed review of ML and non-ML research). Yamaguchi et al (Yamaguchi et al., 2021), apart from implementing a logistic regression model that uses target, mechanism of action (MOA), modality, and application of drugs as features to predict clinical trial success, also provide a review that classifies literature based on the type of features used to predict clinical outcomes. We summarize the principal differentiation in Section 1.1.1 and refer the reader to the two recent studies (Feijoo et al., 2020, Yamaguchi et al., 2021 that provide exhaustive reviews of predictive algorithms.

1.1.1 Bioactivities and Drug Properties as Clinical Trial Outcome Predictors

In this study, we explore the predictive power of bioactivity, alongside features that represent the drug's chemical and pharmacological properties. Each trial has a candidate drug and an indication. Common protein targets between the drug and the indication are retrieved using the databases and methods described in Section 2. We show that using an aggregate activity value of common targets between drug and disease is a powerful indicator of clinical trial success, even with the challenges and ambiguity involved in bioactivity classes.

A feature set with bioactivity and, primarily, drug-based variables (see Table 1 for the full list) ensures that our model is not just a statistical retrospective study, but can be used in the early stages of drug discovery. This provides a probability value at the very early stages of drug discovery, and the opportunity to re-focus prioritization of hits. Since the bioactivity is specific to a drug-target pair, unlike any other studies in the past, researchers will have opportunities to modify drug design protocols or prioritize existing drug candidates based on the candidates' probability to succeed in clinical trials. As far as we know, this is the only study that correlates bioactivity to clinical trial success. Further, we have exclusively chosen open-source data sources to produce our data, and provide access to all the code used to generate our dataset and produce the ML models in an unrestricted open-source format. This ensures that other groups with access to larger (commercial) annotated clinical trial data can use our algorithms, and possibly, achieve higher predictive scores.

Each trial has a specific indication and a single drug candidate. As mentioned earlier, we consider single drug clinical trials. Trials with greater than one indication or those that present ambiguity in the assignment of a specific indication are excluded in this study. Targets that can be mapped to the drug and can be linked mechanistically to the disease associated with the trial, present an exclusive set of 'efficacy

targets' (Santos et al., 2017). It is generally accepted that an unambiguous assignment of a specific drug-target interaction responsible for the therapeutic activity is evasive. Nonetheless, we have used bioactivity data for drug-target association *only* when the target has been implicated through direct evidence with the disease in question. Many drugs have promiscuous mechanistic effects, and hence using efficacy targets as an indicator will include a statistical aggregating component across the available multi-target bioactivities (Alberga et al., 2018, Montaruli et al., 2019, Lenselink et al., 2017, Papadatos et al., 2015, Mendez et al., 2019, Mervin et al., 2015, Mohan, 2019). Statistical measures to enumerate bioactivities have been used previously, towards other goals (Montaruli et al., 2019, Lenselink et al., 2017, Shamsara, 2021). Nonetheless, due to stringent exclusion criteria (see Section 2), the intersection of two manually curated resources, DrugBank (Wishart et al., 2006) and CTD (Davis et al., 2021) to derive this list, resulted in 80% of the trials we have used in this dataset having a maximum of two targets in common with drug and indication as shown in Section S7.4 of Supporting Information. This implies the model will perform better on predicting trial outcomes of drugs with known mechanisms of action, in other words, validated efficacy targets.

2 METHODS AND MATERIALS

The foundation of every machine learning predictive model in the area of biomedical informatics is a robust integrated dataset. In our experience, the big data problem is significantly pronounced in the field of chemistry (and biology). As outlined in other studies (Searls, 2005, Ritchie et al., 2015, Swainston et al., 2017, Zeng et al., 2022), as well as our own experience, the challenges range from the mapping of individual entities (e.g., names of drugs between ChEMBL and DrugBank), mapping entities to valid experimental data (e.g., mapping drugs from one database to activity data of the same drug in another), and so on. The data pre-processing, preparation and engineering steps are approximately 80% of the invested time in most biomedical informatics studies, and this stems from data integration being a dense process requiring strong domain expertise in addition to data wrangling proficiency. Data scientists in the area of life sciences are part computer science and part domain experts (Tetko et al., 2016). Consequently, each research objective has to have its data integration strategy that is focused on serving the needs of that objective, and universal integration of all databases will not work in most ML-based biomedical research. In this context, we have previously synthesized the initial framework, CompoundDB4j (Murali et al., 2020), which contains integrated data of ChEMBL and DrugBank, to have all approved drugs in DrugBank accurately mapped to ChEMBL, such that all related annotations (e.g., bioactivities) of these compounds can be retrieved. The extended capabilities in the current instance of the framework include clinical trials (Tasneem et al., 2012) and experimentally validated disease-target associations (Davis et al., 2008) to predict clinical outcomes. The advantages of having this data in the Neo4j database are mentioned later in this section.

2.1 Clinical Trial Outcome as a Dependent Variable

The models (see Section 3) seek to predict the chances of a drug clearing at least Phase II of an FDA (Food and Drug Administration) trial. The bimodal dependent variable is constructed by classifying all trials with a 'Completed' Status in Phase III or Phase IV as a 'Pass' (0 value) categorization, and all trials that have the Status of 'Terminated', 'Suspended', etc. in Phase II/ III/ IV with a 'Fail' (1 value) categorization, as reported against the trial ID (NCT ID) in the AACT (accessed March 2021). This method of classification is not very different from earlier work (Gayvert et al., 2016). Since we categorize only 'Completed' trials from Phase III and up as a 'Pass', while the 'Fail' category includes negative trials from Phase II/ III/ IV, the model is expected to predict the possibility of the overall success of a particular single drug trial to move beyond Phase II, and not necessarily the ability of a compound to transition across specific phases. A separate model is presented that predicts the ability of a compound to clear Phase II. In this model, the 'Fail' category includes data from Phase II and above.

Apart from the trial outcome (*Status*), the *Intervention Name* and the *Condition / Disease* fields are retrieved from AACT, which are used to retrieve the associated values available from the other databases (see the following section).

2.2 Functional Summary of Data Integration

The following open-source data sets are used: ChEMBL (Gaulton et al., 2017), DrugBank (Wishart et al., 2006), UniProt (The UniProt Consortium, 2021), CTD (Davis et al., 2021), and AACT (Tasneem et al., 2012). AACT is used as the principal source of clinical trials. For each of the single-molecule/drug trials, we find the DrugBank record, which provides the target(s). While a drug-target search is also done on CTD, the disease/condition that is the focus of a trial is matched with CTD, so it can be mapped to genes (targets) associated with the disease. The intersection of this disease-target list with the drug's targets list will provide

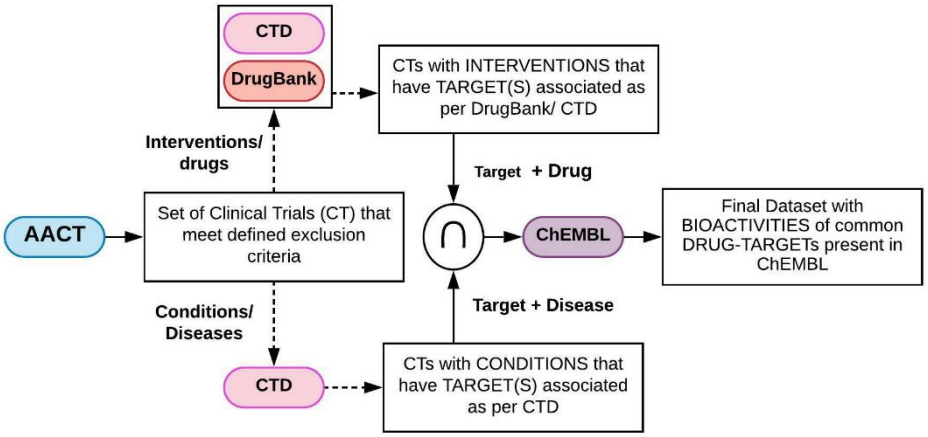

FIGURE 1 Data Integration

the final drug-target list of interest. This is then used to query ChEMBL to get the respective drug-target activity values. Finally, the drug's chemical properties are calculated using RDKit (Landrum, 2016).

The selection of a subset of the clinical trials based on the Phase and Status of a trial is explained in Section 2.1. All entities of interest (drugs, bioactivities, etc.) are not common across the databases. In other words, there is steady depletion in the number of trials that can be used in the data set since the corresponding feature necessary for this study may not be present across multiple databases (For example, a particular drug associated with a trial may be found in DrugBank, but not in CTD. In which case, the clinical trial record is dropped). As presented in Section S6 of Supporting Information, the final number of data points is an aggregate result of applying exclusion conditions and removing non-mappable points during data integration.

Figure 1 serves as a functional diagram to explain the integration. However, the actual integration is done as an extension to CompoundDB4j (Murali et al., 2020), which already integrates ChEMBL and DrugBank in a Neo4j (Neo4j) database. Graph databases enable the creation of complex queries that involve many connections, as compared to relational databases. As we are dealing with heterogeneous data sources, graph databases make pre-processing, filtering, etc. seamless and extremely adaptable to changing requirements of retrieval without needing to change fundamental storage paradigms (Robinson et al., 2015). The remaining databases are also integrated using cypher queries via intermediate CSV/TSV format subsets of them. The GitHub code can be used to completely regenerate the Neo4j integrated database, and also contains the cypher queries used to merge the individual datasets to get the final set of trials and the associated entities used to generate the final feature set. For a detailed programmatic workflow of the integration of the individual databases, the fields extracted from each, and the final construction of the data set, see Section S1 of Supporting Information.

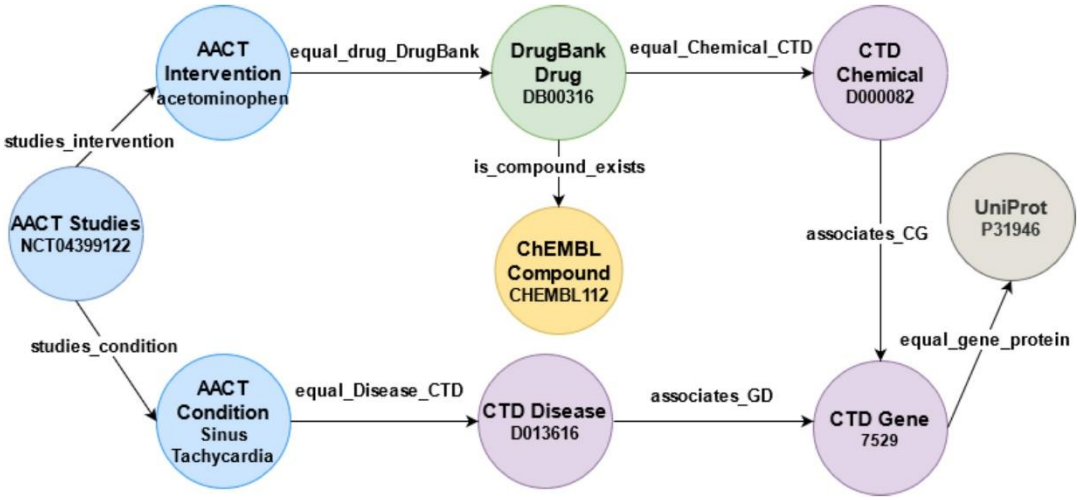

FIGURE 2 Illustration of graph connections from Integrated resource

2.3 Neo4j For Data Storage

Neo4j (Neo4j) is a top-ranked graph database based on various metrics (DB-Engines Ranking of Graph DBMS, Hoksza & Jelínek, 2015), and has been useful in bioinformatics applications towards the research and analysis of complex biological pathways (Balaur et al., 2017, Casaní et al., 2020), human metabolism data (Fabregat et al., 2018) and DNA interaction studies (D'Agostino et al., 2021), to name a few. We found Neo4j to be useful in ingesting data from diverse databases independently and then integrating them based on our current use case. Neo4j uses property-based graph models, and hence, nodes and edges represent one or many properties associated with the same. In the context under study, tables from each database are represented as node labels, and each node under these node labels store the associated metadata as properties. For example, node 'COMPOUND_RECORDS' contains properties - Record_ID, Compound_ID, MOLREGNO, Compound_Name etc. Relationships created directly map to ontologies and schema specific to the respective databases. Starting from XML and MySQL file formats of the latest versions for DrugBank and ChEMBL respectively, the data preparation, population, and fusion are done as in our previous study (Murali et al., 2020).

Similar protocols were used to ingest the other datasets, where all original download formats were converted to CSV, so cypher queries can then be used to import data into our Neo4j database. While CTD provides downloadable files in CSV format, in-house python scripts (ClinicalTrials_Prediction) are used to convert UniProt's XML and AACT's pipe-delimited formats to CSV formats. Subsequently, cypher queries are constructed and executed with periodic commit and APOC procedures (Baton & Van, 2017) to populate data in Neo4j. Only manually curated (Swiss-Prot) human proteins are extracted from UniProt. In AACT, the Studies, Conditions, and Interventions tables are used in this study, and these are merged using the *NCT_ID* unique identifier that is common across these tables. An illustration of intervention associated with a trial from AACT connected with DrugBank, from DrugBank to ChEMBL and CTD, and further, CTD to UniProt is shown as a connected graph in Figure 2.

2.4 Feature Derivations

Features that are used to train the models are listed in Table 1. This is a list of features that are used to build the ML models, and some of them are taken directly from the respective databases, while some are derived. An example of a derived feature is fingerprints. SMILES (Weininger, 1988) is a chemical notation to describe the structure of a compound. We extract SMILES from DrugBank and use this to get molecular fingerprints.

For the fields used from each data source, we refer the reader to Section S1.2 of Supporting Information. The final feature set can be categorized as molecular fingerprint-based, drug properties, compound-target association indices.

Category A (in Table 1) - Molecular encodings are an abstraction of compounds, and ideally, are expected to represent all the properties of the compound so this can be used in various cheminformatic computations (e.g., high throughput screening) without any loss of chemical and structural information. These machine-readable compound formats are used extensively in machine learning and AI-driven drug discovery (David et al., 2020). We have explored four molecular encodings and empirically evaluated which of them worked well for this dataset.

MACCS (Molecular ACCess System) Keys (Durant et al., 2002) are structural keys and fall under a class of ambiguous molecular notations seeking to fingerprint compounds based on the presence or absence of a specific preset functional group or structural component. The RDKit (Landrum, 2016) implementation of this is used in this study. The other type of molecular descriptor notation is the hashed fingerprint, where each compound is represented by features generated by the specific molecule and depends on its physicochemical and structural characteristics. Unlike structural keys, the features and the positions are not preset. Morgan fingerprints are an implementation of this class of fingerprints, also called circular fingerprints (Rogers & Hahn, 2010), built using the Morgan algorithm. This set of molecular features is also generated using RDKit in this study.

TABLE 1 Features extracted for Machine Learning

Description	Feature code	Source
Category A		
Molecular ACCess System based fingerprints	MACCS	RDKit
Morgan fingerprints with radius 1	Mol2vec	Jaeger et al., 2018
SMILES based fingerprints	smi2vec	Zhang et al., 2020
Category B		
Molecular weight	MLWT	RDKit
Hydrogen bond acceptor count	nHBA	RDKit
Hydrogen bond donor count	nHBD	RDKit
Topological polar surface area	TPSA	RDKit
Count of rings	nRNG	RDKit
Count of atoms	nA	RDKit
Count of rotatable bonds	nRB	RDKit
Gasteiger charge	GC	RDKit
Count of valence electrons	nVE	RDKit
Count of hetero atoms	nHetA	RDKit
Count of heavy atoms	nHevA	RDKit
Count of approved trials for an intervention	nAT	Calculated - AACT
Category C		
Count of common targets associated with a drug and disease	nCT	Calculated - DrugBank, CTD
Count of targets associated with a drug	nDT	Calculated - DrugBank, CTD
Standard value for bioactivity of a target	Stdval	ChEMBL
Binding affinity dissociation constant	Ki	ChEMBL

Natural Language Processing (NLP) techniques are widely used to produce continuous vector representations of words, which measure syntactic and semantic word similarities (Mikolov et al., 2013). This approach, called Word2Vec, seeks to group similar words in proximal vector spaces and learns low dimensional features from raw data. The continuous skip-gram model of Word2Vec with negative sampling

is used as it captures semantic information (Wan & Zeng, 2016). The skip-gram Word2Vec model generates vector $V(w) \in R^m$, where w represents the word and m represents the length of the word vector. In our case, the compound representation (Morgan or SMILES) is treated as a sentence and the associated substructures are treated as words. The model uses prediction of the context of a given word as it's an unsupervised task. Motivated by advanced representation learning models in the literature, we use SMILES2Vec (Goh et al., 2017) and Mol2vec (Jaeger et al., 2018) to explore the chemical space and characterize compound representations.

The same concept was extended to chemical fingerprints through an algorithm called Mol2vec (Jaeger et al., 2018), which uses the Morgan algorithm derived chemical substructures as *words* and the compounds as *sentences*. A pre-trained skip-gram model built with the corpus from 19.9 million compounds of ZINC original and ChEMBL is used. A radius 1, a window size of 10, and embedding vectors of length 300 are used in our study, and this reflects the optimizations adopted in the original study (Jaeger et al., 2018). The SMILES2Vec approach (Goh et al., 2017) obtains the embeddings for the compounds based on SMILES representation. We conformed to the best fitting hyperparameters of skip-gram (Wan & Zeng, 2016, Zhang et al., 2020) wherein the size of the embedding vectors is fixed at 100, while the size of the context window and the number of negative samples are fixed at 12 and 15 respectively. In the current work, we adopted the publicly available implementations, Mol2vec (Jaeger et al., 2018) and SMILES2Vec (Wan & Zeng, 2016, Zhang et al., 2020).

Recent works (Li & Fourches, 2020, Wang et al., 2019) have shown the effectiveness of large-scale pre-training followed by task-specific fine-tuning to obtain a task-specific language model. Pre-training on a larger dataset followed by fine-tuning on sparse data is observed to provide better generalization and performance in the area of QSAR modeling (Li & Fourches, 2020). Extrapolating this to our study, the original pre-trained model of Mol2vec with millions of compounds is fine-tuned on a smaller DrugBank-specific dataset of around 11k Morgan fingerprints. We refer to these features as the Mol2vec fine-tuned model (or *Mol2vec ft*).

Category B (in Table 1) - The feature representations from compound fingerprints are then combined with physiochemical descriptors. Features with source as RDKit in Table 1 are calculated based on SMILES for every trial in the constructed dataset.

The number of trials approved for a given intervention is calculated based on the overall data collected (Phase3, Phase4 Completed). We evaluate the importance of this feature in the later steps.

Category C (in Table 1) - The feature representations from compound fingerprints along with physicochemical descriptors are combined with target properties and are subjected for use in machine learning model development. In the development of drug treatments, the drug targets which interact with the disease-associated are considered (Sun et al., 2015). However, the exact mechanism is not completely clear between drug-target and treatment (Sun et al., 2015). This is still an important point, so identifying the common targets between drug and disease is important for clinical trials and hence we consider the intersection of targets between drug and disease and arrive at the count common targets, linked to each trial. We examine the drug targets from DrugBank and CTD one after the other, check whether it is available in disease targets from CTD and take only the targets which belong to both drug and disease. The distinct count of drug targets from DrugBank and CTD is calculated and serves as an additional feature.

For the common targets between Drug and Disease, bioactivity features are extracted from ChEMBL. The heuristics composed from literature for the binding assays linked to the target and the associated bioactivity values are explained in Section S6.2 of Supporting Information. We retrieve the values for Ki and Stdval for every common target linked in the constructed dataset, wherever applicable and available.

2.5 Data Exploration, Analysis, and Preprocessing

We performed an exploratory analysis of the features to gain insight into their distribution, the extent of correlation between the variables, and to understand their contribution to the classification performance. The insights from the experiments conducted are listed below**.**

- Section S7.1 of Supporting Information shows the Pearson's correlation coefficient of various

features. Pearson's correlation coefficient is a measure of linear correlation between two variables. The plot reveals that some features, such as molecular weight and the number of valence electrons or the number of atoms that have an obvious linear association, demonstrate a strong positive correlation.

- From the hue of the pair plots in Section S7.2 of Supporting Information, it is evident that the distributions of variables for trials with the label 'Pass' are not visually distinct from the distributions of the variables with the label 'Fail'.
- Section S7.3 shows a heat map of probability of passing, given a drug-disease pair. This probability is computed as the fraction of the data labeled 'Pass' for the 20 most frequent drug-disease pairs. Some entries are consistent with observations in the literature, such as a high probability of 'Pass' for Sorafenib for carcinoma (Keating & Santoro, 2009, Llovet et al., 2008, Lee et al., 2021) or Risperidone for anxiety disorders (Brawman et al., 2005, Simon et al., 2006).

The data is scaled with a min-max scalar to the range [0, 1]; this ensures each feature is in the same range. Next, to ensure the data is representative and that we work with samples with annotations that are amenable to interpretation, we plotted the number of clinical trials against the count of common targets (nCT) and inferred that nCT is less than 7 for most of the trials. This is shown in Section S7.4 of Supporting Information. Hence, for a meaningful design of the predictive models, we consider only the data for which nCT is less than or equal to 6 and contain nonblank ATC codes for the compound associated with a trial.

3 Machine learning models

This section explains the machine learning models built using the data extracted from the integrated data source, explained in Section 2.2. All the models are implemented in Python 3.6 with Sci-kit learn and its dependent packages (Pedregosa et al., 2011). The steps in the pipeline are shown in Figure 3. We begin with an exploratory analysis of the data and filtering of a meaningful subset as explained in Section 2.5. Next, we experiment with various representations of the features, particularly, various compound fingerprints detailed in Section 2.4. Next, we combine the best-performing fingerprints with compound-physiochemical features. We build models with and without the inclusion of bioactivity features to evaluate their utility. After experimenting with the combination of features, we compare the performance of multiple models towards predicting the outcome of clinical trials. Performance evaluation is done based on the most meaningful train-test split in the context of class imbalance and to tune the parameters on the training set for each of the models considered. The rest of this section details the rationale and nuances of the model construction process.

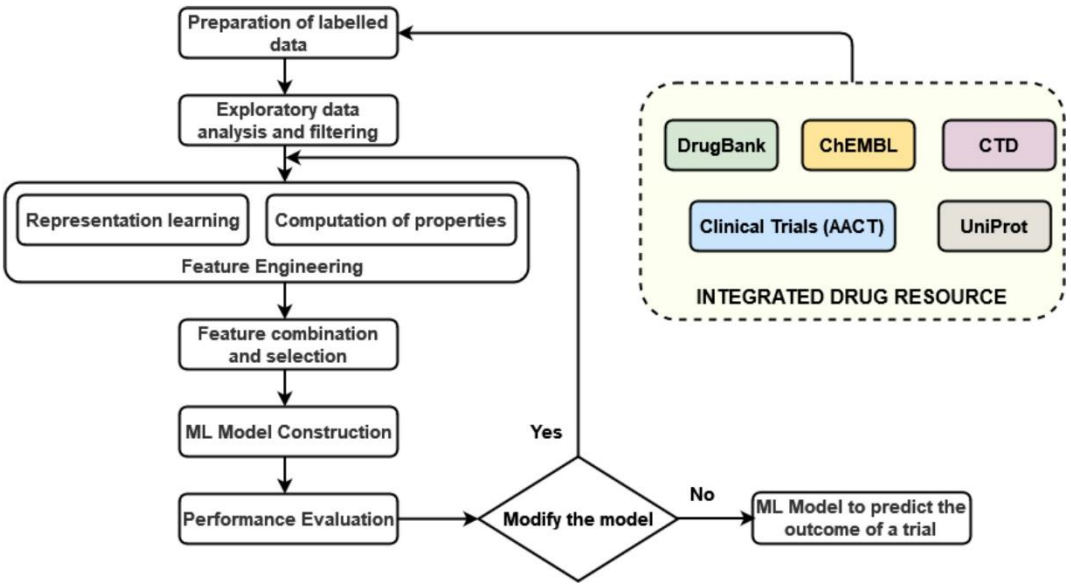

FIGURE 3 Machine learning pipeline

3.1 Rationale for model selection

Given the features (or a subset of features) in Table 1, the primary objective is to predict the success (or failure) of a clinical trial. Since the data has approximately 1290 instances and 18 features, including bioactivity, we treat this problem as one of 'small data' machine learning. 'Small data' generally refers to limited data collated for a purpose and is acknowledged to be more effective for drawing statistical inferences than the relatively noisy and more complex big data (Faraway & Augustin, 2018). Consequently, machine learning models used for prediction tasks on small data are more amenable to interpretation.

Among these models, the decision tree---a sequence of decisions on the features that leads to a prediction---is considered one of the most interpretable machine learning models (Breiman, 2001a, Guidotti et al., 2018). There are multiple algorithms for decision trees; we seek to leverage the power of multiple decision trees through ensembling.

3.2 Ensemble machine learning algorithms: Random forest

An ensembling of multiple decision trees on random subsets of the training data followed by a majority vote constitutes the Random forest (RF) (Breiman, 2001b). This prediction algorithm is not only popular and widely used but also serves as a benchmark for clinical trial research and has outperformed standard classifiers, such as Logistic Regression (Couronné et al., 2018). Recent studies on predicting the approval of clinical trials have demonstrated the efficacy of the RF algorithm (Gayvert et al., 2016, Lo et al., 2018, Feijoo et al., 2020, Vergetis et al., 2020, Ubels et al., 2020, Saurabh et al.,2020). Inspired by these precedents, we have used the RF classifier to predict an input clinical trial's success (or failure).

The Gini index is used as the splitting criterion to construct the decision trees (Raileanu & Stoffel, 2004). The sequence of features leading to the prediction indicates the importance of each feature, with each successive feature being less important in the sequence. Computed using the Gini index in our paper, the sequence of features facilitates interpreting the model and eliciting the most important features that contribute to making a prediction.

3.3 Ensemble machine learning algorithms: AdaBoost

AdaBoost is another successful algorithm from the category of 'Bagging-Boosting' classifiers, and are often seen as a 'random' forest of forests (Freund & Schapire, 1997, Wyner et al., 2017, Schapire, 2013). While the RF optimizes the Gini index for each tree, AdaBoost considers a stage-wise optimization for the full ensemble (Wyner et al., 2017). It has the ability not only to model nonlinear behavior, much like RF but also be effective for sparse datasets (Friedman et al., 2004).

The base estimator, a decision tree by default, which is used to build the ensemble and the number of estimators at which the boosting stops are the parameters of the AdaBoost algorithm that are empirically chosen.

3.4 Dealing with a class imbalance in the data and tuning of model parameters

There is a marked imbalance in the prediction classes: 86.8% of the data have the label 'Pass' and the remaining 13.2% of the data have the label 'Fail'. We use a class-balanced weighting scheme in the process of training the RF classifier, that supports a class-balanced weight for the loss function. The design of the AdaBoost algorithm that results in higher weights for misclassified samples does not warrant an explicit weight. Further, to tune the parameters of all models, we use a grid search of parameter values on the training data, for the combination that yields the highest prediction accuracy.

3.5 Towards prediction of successful trials with Clusters based on fingerprint similarity

The fact that fingerprints encode the structural aspects and similar molecules can exhibit the same biological

activity (Cereto et al., 2015, Muegge & Mukherjee, 2016) sets the groundwork and the reason to design the cluster-based cross-validation scheme. We use the K-means algorithm (Likas et al., 2003) to identify homogeneous subgroups or clusters within the data based on the similarity between their fingerprints. Based on this, we use random sampling to create validation sets for k-fold cross-validation.

We ensure that each of the *k* test sets contains data from each cluster, drawn randomly and we ensure the samples in the *k*th test set do not overlap with the samples in the *k*th training set. While the clustering provides a mechanism to arrive at representative train and test sets for validation, the skew in the distribution between 'Pass' and 'Fail' samples remains. To address the class imbalance, in addition to cluster-based random sampling with the data for Phase3 trials, we also study the effect the addition of Phase2, 'Fail' records has on the prediction accuracy.

3.6 Ensemble Towards prediction of failed trials using the Leave-p-out method

Identification and prediction of failures early in the development cycle of clinical trials are of great value in drug development (Feijoo et al., 2020). To facilitate the classification of 'Fail' samples, in this method, we

use *p* observations in the test set and all the remaining ones in the training set. In particular, we set the value of *p* as 2, wherein a pair of data points is set aside for validation, there is one sample from each class. Though the Leave-p-out scheme is computationally expensive due to repeated training, it is reliable and powerful for smaller datasets (Sammut & Webb, 2010, Hastie et al., 2009) and AUC estimations in a binary classifier (Airola et al., 2011). This approach aims to have a fair estimate in predicting the failed trials ignoring the computational cost. To inspect the behavior of the model on interventions that have more than one trial, we analyze the performance of recall when the intervention and all its trials are in the Test set against the scenario when only one intervention is in the Test set. We believe this analysis can provide deeper insights on the likely outcome of clinical trials for drugs in re-purposing studies.

4 Results

4.1 Integrated Framework

An integrated Neo4j graph database is created (Section 2.2). It contains 44,320,573 nodes and 131,916,935 relationships, which encompasses the data from multiple resources listed earlier. Our git repo (ClinicalTrials_Prediction) contains instructions on the installation, setup of the integrated data. To ensure

reproducibility, the data is made public (Murali, 2022a). The total count of nodes from individual databases is discussed in Section S1 (Supporting Information), the count of relations and relationship types spanning between the databases is presented in S2. Relationships across the databases, created out of complex data mapping and fusion operations, are shown with source and destination in S4. S5 contains Venn diagrams to depict the count of individual and common entities across the databases for Compound, Target, and Disease. This can offer useful insights on shared entities in the integrated data. The graph database provides visualization and data mapping, along with a flexible and extensible platform with an increased speed of execution. The scope of integration for our group, as well as other researchers, shall be extended to have a wide range of chemical databases (e.g., PubChem, KEGG, SIDER), and also non-chemical databases (e.g., USPTO). Browsing or establishing multiple federated queries across the databases is often cumbersome and computationally intensive. Hence, as a single source of information, this well-connected data can be used to discover linkage across heterogeneous resources and can serve as biological knowledge graphs. In addition, this can further facilitate machine learning studies on graphs (e.g., link prediction or community detection).

4.2 Machine learning models

We evaluated the performance of machine learning models discussed in Section 3. For a maximum depth of the individual estimators in the RF model set to 20 and the number of estimators set to 200 for RF and 250

for AdaBoost, respectively, we report Accuracy (ACC), True Positive Rate (TPR) also referred to as Sensitivity, the True Negative Rate (TNR) also known as Specificity, the Area Under the Curve (AUC) and the F1 score or the balanced F-measure (for Pass, Fail), the harmonic mean of precision and recall. A subset of curated data with all the features used for the development of models is made available in public share (Murali, 2022b).

4.2.1 Evaluation of Compound fingerprints

First, we evaluated the performance of models trained on fingerprints detailed in Section 2.4. We noticed that the RF model trained on Mol2vec fingerprints and SMILES2Vec fingerprints yield 0.92 (ACC). MACCS fingerprints yield 0.82 (ACC), and Morgan fingerprints are slightly better than MACCS 0.83 (ACC). It has to be noted that the fine-tuned model of Mol2vec also yielded 0.92 (ACC). This can be attributed to the fine-tuning being limited to a small number of representative compounds from the DrugBank. The best performing fingerprint i.e., Mol2vec, is used further in all the subsequent experiments.

4.2.2 Deriving the feature importance scores

Next, we study the importance of various features measured by the mean reduction in the Gini Index of the RF classifier. We present the results of feature importance scores associated with all physiochemical properties of compounds in Figure 4. The relative importance of various features when the Stdval, the

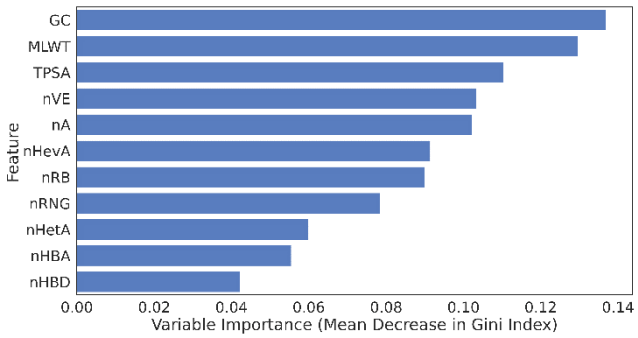

FIGURE 4 Mean decrease in Gini impurity for physio chemical features of Compound

bioactivity feature, is included is shown in Figure 5. In this figure, we report the average score of the Mol2vec fingerprints as one feature. Stdval (median, min), GC, Mol2vec, MLWT are observed to have greater importance, while nHevA seems to have a far less impact on predicting the outcome of a clinical trial.

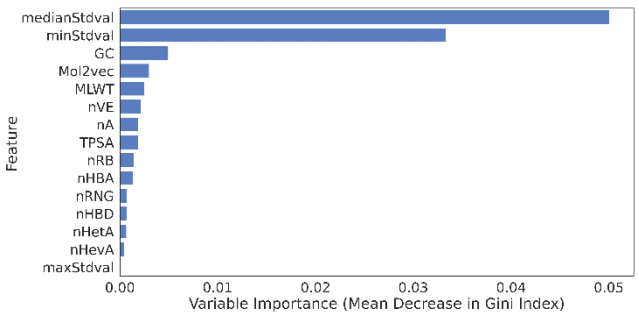

FIGURE 5 Mean decrease in Gini impurity for Compound and Bioactivity features

The top three therapeutic groups identified in the dataset are Group N--Nervous System, Group C--Cardiovascular system and Group A--Alimentary tract and metabolism. We present the feature importance for these categories in Figures 6, 7, and 8 respectively. Through all these figures, we find that independent of therapeutic area, bioactivity features, Stdval (min, and median values), are

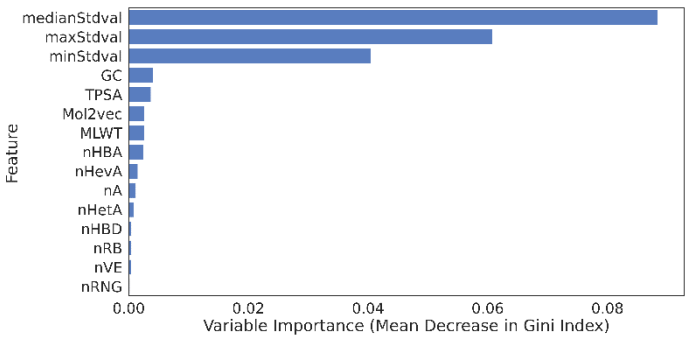

FIGURE 6 Mean decrease in Gini impurity for ATC Group N - Nervous System

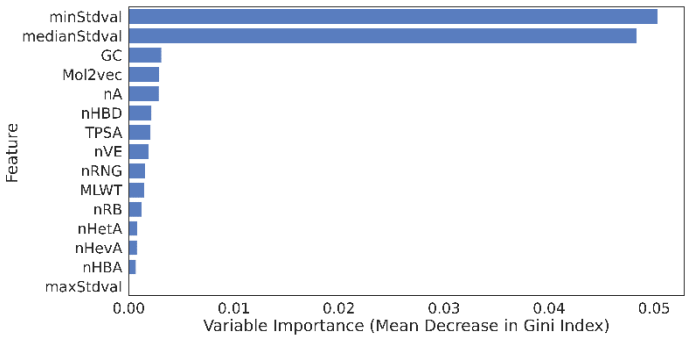

FIGURE 7 Mean decrease in Gini impurity for ATC Group C - Cardiovascular system

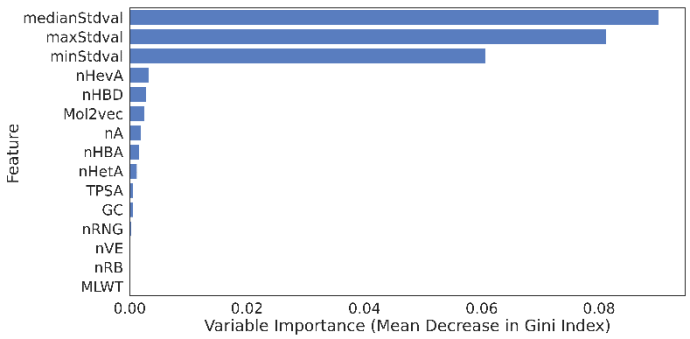

FIGURE 8 Mean decrease in Gini impurity for ATC Group A - Alimentary tract and metabolism

ranked at the top. The maximum value of Stdval seems to have a strong bearing in the prediction of the outcome of clinical trials for the therapeutic classes of the nervous system and the alimentary tract and metabolism.

4.2.3 | Cluster-based cross validation results

Since the primary task of the ML models is to predict the outcome of a clinical trial, we present the results of the cluster-based scheme discussed in Section 3.5. The clusters identified by the k-means algorithm are depicted in Section S8.1 of Supporting Information. There are six clusters in all and about four of these six have roughly the same number of data points. The results of 5-fold cross-validation are summarized in Table 2. We find that the highest accuracy is obtained using the RF classifier; the various performance measures computed on the predictions by the RF are 0.93 (ACC), 0.99 (TPR), 0.58 (TNR), 0.96 (F1-Pass) and 0.68 (F1-Fail) and the AdaBoost model results in 0.87 (ACC), 0.96 (TPR), 0.24 (TNR) and 0.93 (F1-Pass) and 0.32 (F1-Fail). The Receiver Operating curve (ROC) for the entire dataset (marked 'Whole Dataset' in the legend) is shown in brown in Figure 9 and the area under the receiver operating curve, AUC is 0.67. A plot of the histogram of the ATC groups within each cluster (Section S8.2 of Supporting Information) reveals that ATC group C or ATC group N or ATC group A is the most dominant ATC group in these clusters. A plot of the ROC curves for the top three therapeutic categories - Nervous system (labeled 'ATC group N'), the Cardiovascular system (labeled 'ATC group C') and the Alimentary tract and metabolism (labeled 'ATC group A') are shown in Figure 9. In particular, the cluster-based approach results in accurate predictions for the trials labeled 'Pass' (F1-score of 0.96 for 'Pass') using the RF classifier.

TABLE 2 Performance evaluation of Cluster-based and Leave-p-out based methods

Method	Classifier	Accuracy	TPR	TNR	F1-Pass	F1-Fail
Cluster-based	RF	0.93	0.99	0.58	0.96	0.68
	AdaBoost	0.87	0.96	0.24	0.93	0.32
Leave-p-out based	RF	0.74	0.80	0.68	0.74	0.64
	AdaBoost	0.66	0.68	0.64	0.60	0.57

Consistent with our understanding of the importance of the bioactivity feature, Stdval, models trained without Stdval show a slight drop (2%) in the accuracy. Further, we find no significant difference between using the median value of bioactivity *vis-a-vis*, the corresponding min-median-max values. The results obtained with

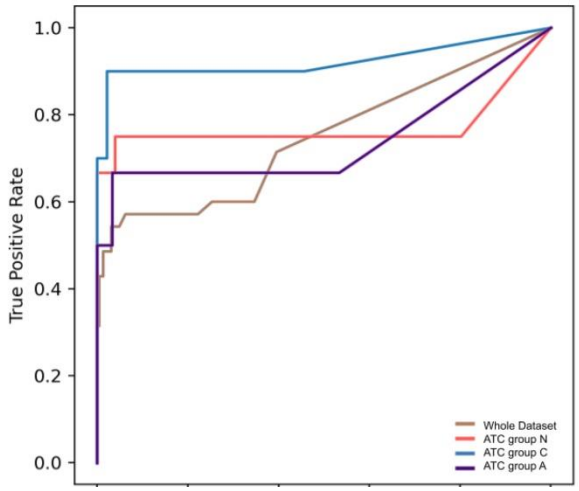

FIGURE 9 Receiver Operating Curve for cluster-based cross-validation

Ki (instead of Stdval) as the bioactivity feature are reported in Section S8.6.1 of Supporting Information. The performances of other machine learning models trained with all the features are presented in Section S8.3 of Supporting Information. Results with the inclusion of the instances labeled 'Fail' from Phase2 trials are reported in Section S.8.7.1 of the Supporting Information.

4.2.4 Leave-p-out based cross-validation results

The results of the leave-p-out cross-validation scheme discussed in Section 3.6, with a focus on predicting trials labeled 'Fail', we achieve 0.74 (ACC), 0.80 (TPR), 0.68 (TNR), 0.74 (F1-Pass), and 0.64 (F1-Fail) with the RF classifier. The model trained on AdaBoost resulted in 0.66 (ACC), 0.68 (TPR), 0.64 (TNR), 0.60 (F1-Pass) and 0.57 (F1-Fail) as reported in Table 2. The RF classifier outperforms AdaBoost for this experiment as well. The results with and without the bioactivity feature remain consistent with the cluster-based experiments, with a minor drop in the overall prediction accuracy on excluding Stdval: 0.72 (ACC) for RF and 0.65 (ACC) for the AdaBoost classifier. While the use of min-median-max values of Stdval achieves a good prediction accuracy, we note a similar performance when the median value is used as a feature, also consistent with our earlier observation.

Another interesting evaluation of the ML models is their performance on an intervention name not seen during the training phase through the exclusion of all its trials in training against the performance of the models when the models have seen some instances of the intervention in training. A few cases we consider from the data for these experiments are discussed below.

- Intervention, *'Methylphenidate'*, for Neurological disorders, has been used widely and seen in multiple trials such as *'NCT01244269'*, *'NCT02638168'*, *'NCT01825577'*, *'NCT00852059'*. We notice that when all these records are in the test set, the performance was around 0.39 (mean ACC) and 0.75 (mean F1-Fail) whereas, if we have one trial in the test set and rest in the training set with other interventions, we obtain 1.0 for ACC and F1-Fail.
- Intervention, *'Metoprolol'*, for cardiovascular diseases, is seen in trials such as '*NCT00806390'*, *'NCT00182039'*, *'NCT03371823'*. We notice that the test set with all the trials associated with this intervention resulted in 0.65 (mean ACC) and 0.23 (mean F1-Fail) whereas when one trial of this intervention is in the test set, we obtain 1.0 for both ACC and F1-Fail.

These observations lend credence to the hypothesis that the Leave-p-out scheme shows improved performance in predicting the failure rate when the training set is exposed to instances of the intervention applied to similar or other conditions (diseases) or targets.

When we analyze the performance for 'Fail' for top therapeutic categories, the Alimentary tract and metabolism show a better performance with a mean Recall of 0.85 than Cardiovascular (0.80) and Nervous System (0.60). The mean recall score for other therapeutic categories is presented in Section S8.4 of the Supporting Information. With this cross-validation scheme, the performance of other classifiers is shown in Section S8.5 of the Supporting Information. In Section S8.6.2 of Supporting Information, we report the results when Ki is used as a bioactivity feature. Results of the models trained on data set with the inclusion of Phase2 trials labeled 'Fail', are available in Section S.8.7.2 of the Supporting Information.

5 DISCUSSION

We have shown that including the bioactivity of the drug against the putative targets common to the disease considered in the clinical trial in question provides enhanced predictive power to our Random forest classifier. Early-stage prediction of the outcome of clinical trials will help drug discovery researchers to better estimate the probability of success based on tangible activity measurements. Bioactivity, as represented by Stdval, has a significant effect on the model's ability to predict trial outcomes. Further, it is shown to provide a significant advantage in predicting trials of top therapeutic categories (ATC Groups - N, C, A), as described in the results. This emphasizes the importance of considering the mechanism of action data when estimating trial success through our or any similar models, researchers may wish to use. It is important to note that we have not used (due to lack of access) larger clinical trial datasets, but expect that it will lead to

significantly better predictive power, especially for the clinical trial's 'Fail' category.

Our data, derived from the public clinical trial source, over-represents the 'Pass' category for obvious reasons. But, since we provide the code to integrate drug-disease pair to its component target bioactivity, we expect other stakeholders like pharmaceutical companies and research agencies that have access to curated clinical trial datasets and a large body of in-house 'Fail' category trials to perform much better. Our work introduces this possibility for the first time and enables any researchers to repeat the study. It should be noted that the creation of the dataset is a significant effort, and we have enabled researchers to streamline that part of the workflow. In other words, ‘small data’ models presented in this study provide significant predictions to provide value towards predicting the possibility of a 'Pass'. Nonetheless, larger datasets will only enhance the predictive power of both classes of trials.

The models, alongside the elaborate code to integrate drug-disease pairs of clinical trials to putative common targets through validated data sources, in this proposed framework, can aid various drug research groups in accessing early indicators of success. This optimizes the release of Investigational New Drugs (IND) or New Drug Application (NDA) by narrowing down the number of candidate drugs in scaffolds being considered and limiting the cost, time, and other technical challenges.

6 CONFLICT OF INTEREST

There is no conflict of interest among the authors.

7 FUNDING SOURCE

The research was supported by funding provided by the Department of Biotechnology, Government of India (Grant Reference Number: BT/PR16476/BID/7/631/ 2016).

8 DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in figshare at:
https://doi.org/10.6084/m9.figshare.18631901 (Murali, 2022a).
https://doi.org/10.6084/m9.figshare.18646118 (Murali, 2022b). Complete data used to train the ML models are available from the corresponding author upon reasonable request.

9 REFERENCES

Airola, A., Pahikkala, T., Waegeman, W., De Baets, B., & Salakoski, T. (2011). An experimental comparison of cross-validation techniques for estimating the area under the ROC curve. *Computational Statistics & Data Analysis*, *55*(4), 1828-1844. https://doi.org/10.1016/j.csda.2010.11.018

Alberga, D., Trisciuzzi, D., Montaruli, M., Leonetti, F., Mangiatordi, G. F., & Nicolotti, O. (2018). A new approach for drug target and bioactivity prediction: the multifingerprint similarity search algorithm (MuSSeL). *Journal of chemical information and modeling*, *59*(1), 586-596. https://doi.org/10.1021/acs.jcim.8b00698

Balaur, I., Mazein, A., Saqi, M., Lysenko, A., Rawlings, C. J., & Auffray, C. (2017). Recon2Neo4j: applying graph database technologies for managing comprehensive genome-scale networks. *Bioinformatics*, *33*(7), 1096-1098. https://doi.org/10.1093/bioinformatics/btw731

Baton, J., & Van Bruggen, R. (2017). *Learning Neo4j 3. x: Effective data modeling, performance tuning, and data visualization techniques in Neo4j*. Packt Publishing Ltd.

Brawman-Mintzer, O., Knapp, R. G., & Nietert, P. J. (2005). Adjunctive risperidone in generalized anxiety disorder: a double-blind, placebo-controlled study. *Journal of Clinical Psychiatry*, *66*(10), 1321-1325. https://doi.org/10.4088/jcp.v66n1016

Breiman, L. (2001a). Random forests. *Machine learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L. (2001b). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, *16*(3), 199-231. https://doi.org/10.1214/ss/1009213726

Casaní-Galdón, S., Pereira, C., & Conesa, A. (2020). Padhoc: a computational pipeline for pathway reconstruction on the fly. *Bioinformatics*, *36*(Supplement_2), i795-i803. https://doi.org/10.1093/bioinformatics/btaa811

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., & Pujadas, G. (2015). Molecular fingerprint similarity search in virtual screening. *Methods*, *71*, 58-63. https://doi.org/10.1016/j.ymeth.2014.08.005

ClinicalTrials_Prediction, Retrieved from https://github.com/pathri/ClinicalTrials_Prediction

Couronné, R., Probst, P., & Boulesteix, A. L. (2018). Random forest versus logistic regression: a large-scale benchmark experiment. *BMC bioinformatics*, *19*(1), 1-14. https://doi.org/10.1186/s12859-018-2264-5

D’Agostino, D., Liò, P., Aldinucci, M., & Merelli, I. (2021). Advantages of using graph databases to explore chromatin conformation capture experiments. *BMC bioinformatics*, *22*(2), 1-16. https://doi.org/10.1186/s12859-020-03937-0

David, L., Thakkar, A., Mercado, R., & Engkvist, O. (2020). Molecular representations in AI-driven drug discovery: a review and practical guide. *Journal of Cheminformatics*, *12*(1), 1-22. https://doi.org/10.1186/s13321-020-00460-5

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., Wiegers, J., Wiegers, T. C., & Mattingly, C. J. (2021). Comparative toxicogenomics database (CTD): update 2021. *Nucleic acids research*, *49*(D1), D1138-D1143. https://doi.org/10.1093/nar/gkaa891

Davis, A. P., Murphy, C. G., Rosenstein, M. C., Wiegers, T. C., & Mattingly, C. J. (2008). The Comparative Toxicogenomics Database facilitates identification and understanding of chemical-gene-disease associations: arsenic as a case study. *BMC medical genomics*, *1*(1), 1-12. https://doi.org/10.1186/1755-8794-1-48

DB-Engines Ranking of Graph DBMS, https://db-engines.com/en/ranking/graph+dbms

Dowden, H., & Munro, J. (2019). Trends in clinical success rates and therapeutic focus. Nature Reviews Drug Discovery, 18(7), 95-497
Durant, J. L., Leland, B. A., Henry, D. R., & Nourse, J. G. (2002). Reoptimization of MDL keys for use in drug discovery. *Journal of chemical information and computer sciences*, *42*(6), 1273-1280. https://doi.org/10.1021/ci010132r
Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., ... & Hermjakob, H. (2018). Reactome graph database: Efficient access to complex pathway data. *PLoS computational biology*, *14*(1), e1005968. https://doi.org/10.1371/journal.pcbi.1005968
Faraway, J. J., & Augustin, N. H. (2018). When small data beats big data. *Statistics & Probability Letters*, *136*, 142-145. https://doi.org/10.1016/j.spl.2018.02.031
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*(2), 414-421. https://doi.org/10.1016/j.drudis.2019.12.014
Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, *55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504
Friedman, J., Hastie, T., Rosset, S., Tibshirani, R., & Zhu, J. (2004). Discussion of boosting papers. *Annual Statistics*, *32*, 102-107.
Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., ... & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic acids research*, *45*(D1), D945-D954. https://doi.org/10.1093/nar/gkw1074
Goh, G. B., Hodas, N. O., Siegel, C., & Vishnu, A. (2017). Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034*.
Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, *51*(5), 1-42. https://doi.org/10.1145/3236009
Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learnin. *Cited on*, 33.
Hoksza, D., & Jelínek, J. (2015, September). Using Neo4j for mining protein graphs: a case study. In *2015 26th International Workshop on Database and Expert Systems Applications (DEXA)* (pp. 230-234). IEEE. https://doi.org/10.1109/DEXA.2015.59
Jaeger, S., Fulle, S., & Turk, S. (2018). Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, *58*(1), 27-35. https://doi.org/10.1021/acs.jcim.7b00616
Keating, G. M., & Santoro, A. (2009). Sorafenib. *Drugs*, *69*(2), 223-240. https://doi.org/10.2165/00003495-200969020-00006
Landrum, G. (2016). RDKit: Open-Source Cheminformatics Software.,. https://doi.org/10.5281/zenodo.3732262
Lang, T., & Siribaddana, S. (2012). Clinical trials have gone global: is this a good thing? PLoS medicine, 9(6), e1001228. https://doi.org/10.1371/journal.pmed.1001228
Lee, S. K., Jang, J. W., Nam, H., Sung, P. S., Kim, H. Y., Kwon, J. H., ... & Yoon, S. K. (2021). Sorafenib for advanced hepatocellular carcinoma provides better prognosis after liver transplantation than without liver transplantation. *Hepatology International*, *15*(1), 137-145. https://doi.org/10.1007/s12072-020-10131-0
Lenselink, E. B., Ten Dijke, N., Bongers, B., Papadatos, G., Van Vlijmen, H. W., Kowalczyk, W., ... & Van Westen, G. J. (2017). Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *Journal of cheminformatics*, *9*(1), 1-14. https://doi.org/10.1186/s13321-017-0232-0
Li, X., & Fourches, D. (2020). Inductive transfer learning for molecular activity prediction: Next-Gen QSAR Models with MolPMoFiT. *Journal of Cheminformatics*, *12*(1), 1-15. https://doi.org/10.1186/s13321-020-00430-x
Likas, A., Vlassis, N., & Verbeek, J. J. (2003). The global k-means clustering algorithm. *Pattern recognition*, *36*(2), 451-461. https://doi.org/10.1016/S0031-3203(02)00060-2
Llovet, J. M., Ricci, S., Mazzaferro, V., Hilgard, P., Gane, E., Blanc, J. F., ... & Bruix, J. (2008). Sorafenib in advanced hepatocellular carcinoma. *New England journal of medicine*, *359*(4), 378-390. https://doi.org/10.1056/NEJMoa0708857
Lo, A. W., Siah, K. W., & Wong, C. H. (2018). Machine learning with statistical imputation for predicting drug approvals. *Available at SSRN 2973611*. https://dx.doi.org/10.2139/ssrn.2973611
Martin, L., Hutchens, M., Hawkins, C., & Radnov, A. (2017). How much do clinical trials cost. Nat Rev Drug Discov, 16(6), 381-382. https://doi.org/10.1038/nrd.2017.70
Maxmen, A. (2016). Busting the billion-dollar myth: how to slash the cost of drug development. Nature News, 536(7617), 388. https://doi.org/10.1038/536388a
Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., De Veij, M., Félix, E., ... & Leach, A. R. (2019). ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, *47*(D1), D930-D940. https://doi.org/10.1093/nar/gky1075
Mervin, L. H., Afzal, A. M., Drakakis, G., Lewis, R., Engkvist, O., & Bender, A. (2015). Target prediction utilising negative bioactivity data covering large chemical space. *Journal of cheminformatics*, *7*(1), 1-16. https://doi.org/10.1186/s13321-015-0098-y
Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
Mohan, C. G. (Ed.). (2019). *Structural bioinformatics: applications in preclinical drug discovery process* (Vol. 27). Springer.
Montaruli, M., Alberga, D., Ciriaco, F., Trisciuzzi, D., Tondo, A. R., Mangiatordi, G. F., & Nicolotti, O. (2019). Accelerating drug discovery by early protein drug target prediction based on a multi-fingerprint similarity search. *Molecules*, *24*(12), 2233. https://doi.org/10.3390/molecules24122233
Muegge, I., & Mukherjee, P. (2016). An overview of molecular fingerprint similarity search in virtual screening. *Expert opinion on drug discovery*, *11*(2), 137-148. https://doi.org/10.1517/17460441.2016.1117070
Murali, V., Königs, C., Deekshitula, S., Nukala, S., Santhi, M. D., & Athri, P. (2020). CompoundDB4j: Integrated Drug Resource of Heterogeneous Chemical Databases. *Molecular informatics*, *39*(9), 2000013. https://doi.org/10.1002/minf.202000013
Neo4j, https://neo4j.com
Papadatos, G., Gaulton, A., Hersey, A., & Overington, J. P. (2015). Activity, assay and target data curation and quality in the ChEMBL database. *Journal of computer-aided molecular design*, *29*(9), 885-896. https://doi.org/10.1007/s10822-015-9860-5
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.
Raileanu, L. E., & Stoffel, K. (2004). Theoretical comparison between the gini index and information gain criteria. *Annals of Mathematics and Artificial Intelligence*, *41*(1), 77-93. https://doi.org/10.1023/B:AMAI.0000018580.96245.c6
Ritchie, M. D., Holzinger, E. R., Li, R., Pendergrass, S. A., & Kim, D. (2015). Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, *16*(2), 85-97. https://doi.org/10.1038/nrg3868
Robinson, I., Webber, J., & Eifrem, E. (2015). *Graph databases: new opportunities for connected data*. " O'Reilly Media, Inc.".
Rogers, D., & Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of chemical information and modeling*, *50*(5), 742-754. https://doi.org/10.1021/ci100050t
Sammut, C., & Webb, G. I. (2010). Leave-one-out cross-validation. *Encyclopedia of machine learning*, 600-601.
Santos, R., Ursu, O., Gaulton, A., Bento, A. P., Donadi, R. S., Bologa, C. G., ... & Overington, J. P. (2017). A comprehensive map of molecular drug targets. *Nature reviews Drug discovery*, *16*(1), 19-34. https://doi.org/10.1038/nrd.2016.230
Saurabh, R., Nandi, S., Sinha, N., Shukla, M., & Sarkar, R. R. (2020). Prediction of survival rate and effect of drugs on cancer patients with somatic mutations of genes: An AI-based approach. Chemical Biology & Drug Design, 96(3), 1005-1019. https://doi.org/10.1111/cbdd.13668
Schapire, R. E. (2013). Explaining adaboost. In *Empirical inference* (pp. 37-52). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41136-6_5
Searls, D. B. (2005). Data integration: challenges for drug discovery. *Nature reviews Drug discovery*, *4*(1), 45-58. https://doi.org/10.1038/nrd1608

Shah, P., Kendall, F., Khozin, S., Goosen, R., Hu, J., Laramie, J., ... & Schork, N. (2019). Artificial intelligence and machine learning in clinical development: a translational perspective. *NPJ digital medicine*, *2*(1), 1-5. https://doi.org/10.1038/s41746-019-0148-3

Shamsara, J. (2021). Evaluation of the performance of various machine learning methods on the discrimination of the active compounds. Chemical Biology & Drug Design, 97(4), 930-943. https://doi.org/10.1111/cbdd.13819

Simon, N. M., Hoge, E. A., Fischmann, D., Worthington, J. J., Christian, K. M., Kinrys, G., & Pollack, M. H. (2006). An open-label trial of risperidone augmentation for refractory anxiety disorders. *Journal of Clinical Psychiatry*, *67*(3), 381-385. https://doi.org/10.4088/jcp.v67n0307

Sun, J., Zhu, K., Zheng, W. J., & Xu, H. (2015, December). A comparative study of disease genes and drug targets in the human protein interactome. In *BMC bioinformatics* (Vol. 16, No. 5, pp. 1-9). BioMed Central. https://doi.org/10.1186/1471-2105-16-S5-S1

Swainston, N., Batista-Navarro, R., Carbonell, P., Dobson, P. D., Dunstan, M., Jervis, A. J., ... & Breitling, R. (2017). biochem4j: Integrated and extensible biochemical knowledge through graph databases. *PloS one*, *12*(7), e0179130. https://doi.org/10.1371/journal.pone.0179130

Tasneem, A., Aberle, L., Ananth, H., Chakraborty, S., Chiswell, K., McCourt, B. J., & Pietrobon, R. (2012). The database for aggregate analysis of ClinicalTrials. gov (AACT) and subsequent regrouping by clinical specialty. *PloS one*, *7*(3), e33677. https://doi.org/10.1371/journal.pone.0033677

Tetko, I. V., Engkvist, O., Koch, U., Reymond, J. L., & Chen, H. (2016). BIGCHEM: challenges and opportunities for big data analysis in chemistry. *Molecular informatics*, *35*(11-12), 615-621. https://doi.org/10.1002/minf.201600073

The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, *Nucleic Acids Research*, Volume 49, Issue D1, 8 January 2021, Pages D480–D489, https://doi.org/10.1093/nar/gkaa1100

Ubels, J., Schaefers, T., Punt, C., Guchelaar, H. J., & de Ridder, J. (2020). RAINFOREST: a random forest approach to predict treatment benefit in data from (failed) clinical drug trials. *Bioinformatics*, *36*(Supplement_2), i601-i609. https://doi.org/10.1093/bioinformatics/btaa799

Van Bruggen, R. (2014). *Learning Neo4j*. Packt Publishing Ltd.

Vergetis, V., Liaropoulos, G., Georganaki, M., Dimakakos, A., Skaltsas, D., Gorgoulis, V. G., & Tsirigos, A. (2020). A Machine Learning approach for assessing drug development risk. *bioRxiv*. https://doi.org/10.1101/2020.10.08.331926

Wan, F., & Zeng, J. M. (2016). Deep learning with feature embedding for compound-protein interaction prediction. *bioRxiv*, 086033. https://doi.org/10.1101/086033

Wang, K., Gao, X., Zhao, Y., Li, X., Dou, D., & Xu, C. Z. (2019, September). Pay attention to features, transfer learn faster CNNs. In *International Conference on Learning Representations*.

Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, *28*(1), 31-36. https://doi.org/10.1021/ci00057a005

Wishart, D. S., Knox, C., Guo, A. C., Shrivastava, S., Hassanali, M., Stothard, P., ... & Woolsey, J. (2006). DrugBank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, *34*(suppl_1), D668-D672. https://doi.org/10.1093/nar/gkj067

Wong, C. H., Siah, K. W., & Lo, A. W. (2019). Estimation of clinical trial success rates and related parameters. *Biostatistics*, *20*(2), 273-286. https://doi.org/10.1093/biostatistics/kxy072

Wyner, A. J., Olson, M., Bleich, J., & Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, *18*(1), 1558-1590.

Yamaguchi, S., Kaneko, M., & Narukawa, M. (2021). Approval success rates of drug candidates based on target, action, modality, application, and their combinations. Clinical and Translational Science. https://doi.org/10.1111/cts.12980

Zeng, X., Tu, X., Liu, Y., Fu, X., & Su, Y. (2022). Toward better drug discovery with knowledge graph. *Current opinion in structural biology*, *72*, 114-126. https://doi.org/10.1016/j.sbi.2021.09.003

Zhang, Y. F., Wang, X., Kaushik, A. C., Chu, Y., Shan, X., Zhao, M. Z., ... & Wei, D. Q. (2020). SPVec: a Word2vec-inspired feature representation method for drug-target interaction prediction. *Frontiers in chemistry*, *7*, 895. https://doi.org/10.3389/fchem.2019.00895

Zhavoronkov, A., Kudrin, R., Tutubalina, E., & Kuzmina, A. Multimodal AI Engine for Clinical Trials Outcome Prediction: Prospective Case Study of Big Pharma for Q2 2020.

Murali, V(2022a), *Annexure_ClinicalTrials_Prediction_IntegratedData* figshare. https://doi.org/10.6084/m9.figshare.18631901

Murali, V(2022a). Annexure_ClinicalTrials_Prediction_MLSampleData. figshare. https://doi.org/10.6084/m9.figshare.18646118