Journal Pre-proof

Molecular Descriptor Analysis of Approved Drugs Using Unsupervised Learning for Drug Repurposing

Sita Sirisha Madugula, Lijo John, Selvaraman Nagamani, Anamika Singh Gaur, Vladimir V. Poroikov, G. Narahari Sastry

PII: S0010-4825(21)00650-8

DOI: https://doi.org/10.1016/j.compbiomed.2021.104856

Reference: CBM 104856

To appear in: *Computers in Biology and Medicine*

Received Date: 5 July 2021

Revised Date: 24 August 2021

Accepted Date: 6 September 2021

Please cite this article as: S.S. Madugula, L. John, S. Nagamani, A.S. Gaur, V.V Poroikov, G.N. Sastry, Molecular Descriptor Analysis of Approved Drugs Using Unsupervised Learning for Drug Repurposing, *Computers in Biology and Medicine*, https://doi.org/10.1016/j.compbiomed.2021.104856.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

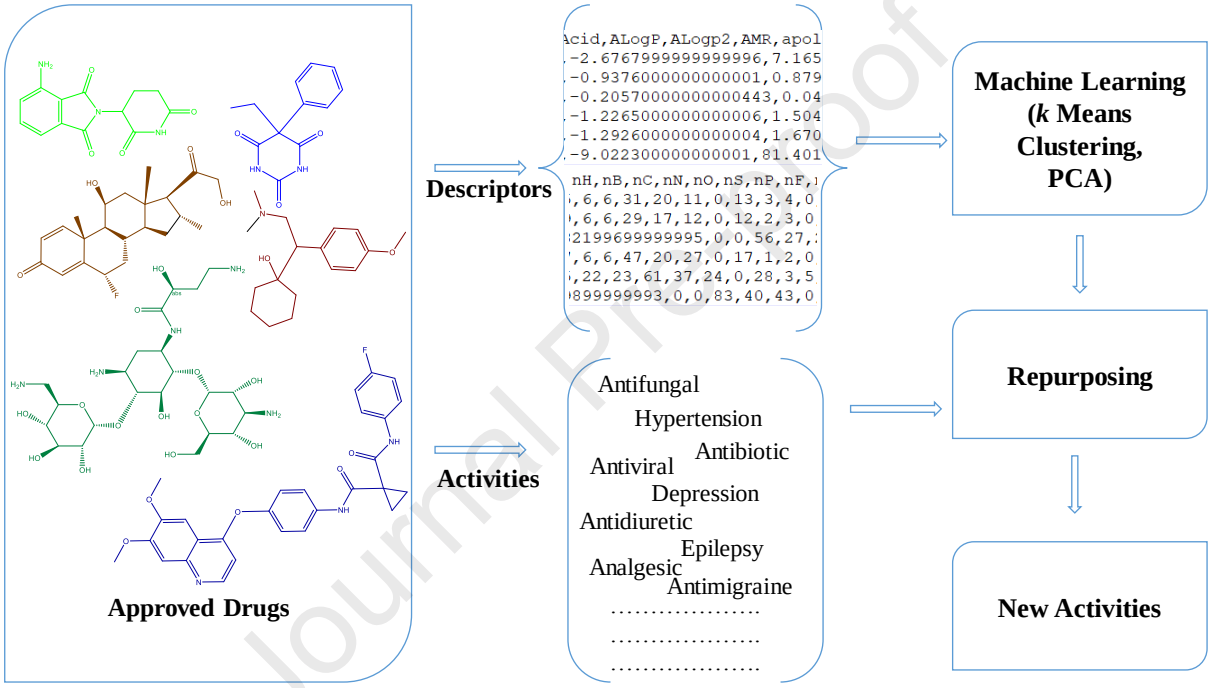
© 2021 Elsevier Ltd. All rights reserved.

Journal Pre-proof

Graphical abstract

Molecular Descriptor Analysis of Approved Drugs Using Unsupervised Learning for Drug Repurposing

Sita Sirisha Madugula[1,2], Lijo John[1,2], Selvaraman Nagamani[3], Anamika Singh Gaur[1,2,3], Vladimir V Poroikov[4], G. Narahari Sastry[*2,3]



Drug repurposing through an integrated unsupervised learning and quantitative structure activity relationships.

Journal Pre-proof

Journal Pre-proof

Molecular Descriptor Analysis of Approved Drugs Using Unsupervised Learning for Drug Repurposing

Sita Sirisha Madugula[1,2], Lijo John[1,2], Selvaraman Nagamani[2,3], Anamika Singh Gaur[1,2,3], Vladimir V Poroikov[4], G. Narahari Sastry[*2,3]

[1]*Centre for Molecular Modeling, CSIR-Indian Institute of Chemical Technology, Hyderabad-500007, India*

[2]*Academy of Scientific and Innovative Research (AcSIR), Ghaziabad-201002, India*

[3]*Advanced Computation and Data Sciences Division, CSIR – North East Institute of Science and Technology, Jorhat, Assam – 785 006, India.*

[4]*Laboratory for Structure-Function Drug Design, Institute of Biomedical Chemistry, Moscow, 119121, Russia*

Corresponding Author

Email: gnsastry@gmail.com; gnsastry@neist.res.in

Journal Pre-proof

Abstract

Machine learning and data-driven approaches are currently being widely used in drug discovery and development due to their potential advantages in decision-making based on the data leveraged from existing sources. Applying these approaches to drug repurposing (DR) studies can identify new relationships between drug molecules, therapeutic targets and diseases that will eventually help in generating new insights for developing novel therapeutics. In the current study, a dataset of 1671 approved drugs is analyzed using a combined approach involving unsupervised Machine Learning (ML) techniques (Principal Component Analysis (PCA) followed by *k*-means clustering) and Structure-Activity Relationships (SAR) predictions for DR. PCA is applied on all the two dimensional (2D) molecular descriptors of the dataset and the first five Principal Components (PC) were subsequently used to cluster the drugs into nine well separated clusters using *k*-means algorithm. We further predicted the biological activities for the drug-dataset using the PASS (Predicted Activities Spectra of Substances) tool. These predicted activity values are analyzed systematically to identify repurposable drugs for various diseases. Clustering patterns obtained from *k*-means showed that every cluster contains subgroups of structurally similar drugs that may or may not have similar therapeutic indications. We hypothesized that such structurally similar but therapeutically different drugs can be repurposed for the native indications of other drugs of the same cluster based on their high predicted biological activities obtained from PASS analysis. In line with this, we identified 66 drugs from the nine clusters which are structurally similar but have different therapeutic uses and can therefore be repurposed for one or more native indications of other drugs of the same cluster. Some of these drugs not only share a common substructure but also bind to the same target and may have a similar mechanism of action, further supporting our hypothesis. Furthermore, based on the analysis of predicted biological activities, we identified 1423 drugs that can be repurposed for 366 new indications against several diseases. In this study, an integrated approach of unsupervised ML and SAR analysis have been used to identify new indications for approved drugs and the study provides novel insights into clustering patterns generated through descriptor level analysis of approved drugs.

Keywords: Unsupervised learning, dimensionality reduction, clustering, approved drugs, drug repurposing, PASS

Journal Pre-proof

1. Introduction

Drug repurposing (DR) is the process of identifying new therapeutic applications for existing drugs [1]. Over the past few years, pharmaceutical industries have hugely invested in the repositioning of approved and withdrawn drugs as traditional drug development is an extremely expensive, laborious, time-consuming, and highly failure-prone avenue [2-6]. DR especially finds its application in rare and neglected diseases where there are very few or no drugs available for treatment [7]. The US-FDA has provided a list of approved pharmaceuticals that can be promising drug candidates for repurposing in rare diseases [8]. Nonetheless, it has always been an endeavour to identify drugs that are equipotent to such orphan drugs. DR also finds its application in infectious diseases such as tuberculosis [9-15], HIV and other communicable diseases where multi-drug resistance is a major problem [16-18]. Currently, DR is also used to identify promising drugs for non-communicable diseases like cancer [19], neurological [20], inflammatory bowel disease [21] and cardiovascular diseases [22-23]. For this purpose, both experimental and computational DR approaches have been used to identify potential candidates for several diseases [24]. Experimental drug repurposing approaches majorly include proteomics techniques [25-26] and *in vitro* high throughput screenings [27] whereas chemoinformatics, data-driven and statistical methods involving gene-target-disease level associations and structural analysis of existing drugs are the computational approaches [28-32]. Also, the advancement in computational drug discovery processes has proved helpful in identifying potential lead molecules in various studies [33-42]. Moreover, the availability of enormous data related to the physicochemical and pharmacological properties of existing drugs and clinical trial information of prospective drug molecules has further aided in identifying promising candidates for DR [43-46]. Hence the 21st century pharmaceutical science is largely dependent on the synergistic use of data-driven and experimental approaches to drug repositioning.

Over the past few decades, several dozens drugs have been repurposed successfully for many new indications outside the scope of their native therapeutic application using experimental, *in silico* and data driven approaches [47]. It is therefore imperative to instil newer data analytical methods to make a substantial effort in designing effective therapeutics. Statistical and data driven approaches are mostly dependent on the structure of drug molecules [48]. There are many freely available drug databases like DrugBank [49], DrugCental [50], PubChem [51], Therapeutic Target Database (TTD) [52], CenterWatch [53], United States Food and Drug Administration (US-FDA) [54] which provides physicochemical and pharmacological profiles of approved drugs across all the major regulatory bodies like FDA, Health Canada, EMA, etc.

Significant efforts have been made to analyze these drugs using statistical and machine learning approaches based on the available enormous data [55]. Consequently, descriptor analysis of

molecules using unsupervised or supervised machine learning techniques may provide a valuable tool to establish various structure-activity relationships. In the current study, we have employed unsupervised ML techniques such as PCA and *k*-means clustering in combination with predictive modeling using PASS tool [56-59] to identify (a) repurposable candidates for various diseases and (b) repurposable indications for the approved drugs. Our approach here is to perform a full-fledged analysis of molecular descriptors of drug molecules followed by clustering them into different clusters to identify similar molecules sharing common molecular features which form the basis for DR. The results were further verified based upon the predicted biological activity estimates obtained by PASS.

2. Methodology

2.1. Preparation of Approved Drugs Library

A set of 2092 approved drug molecules was downloaded from the DrugBank database [49]. These drugs were systematically pre-processed for further analysis. First, molecules containing ions and salts which were unsuitable for further calculations were removed and the remaining molecules were retained from the dataset. We further filtered out compounds used in cosmetics, antiseptics, and sanitizers, which do not form the mainstream therapeutics leading to a dataset of 1671 approved drug molecules. The 1671 approved drugs are further used to calculate the (a) 1444 2D descriptors available in PaDEL [60] which are analysed using PCA and *k*-means clustering (b) all the predicted biological activities in PASS 2017 to identify drugs that can be repurposed for new therapeutic indications. The workflow depicting data curation and the different types of analysis carried out on the approved drugs is shown in **Figure 1**.

2.2. Unsupervised Machine Learning

2.2.1. Principal Component Analysis

In the current study, PCA is performed to reduce the high dimensional data to fewer PCs which are used further to cluster the drug dataset. Both PCA and *k*-means clustering is carried out in R version 3.6.2 operating environment [61a]. Considering the large dimensionality of the dataset, the variables are firstly pre-processed by removing the zero value variables followed by removal of the Near Zero Variance (NZV) variables using 'nearZeroVar' function of *Caret* package [61b]. This step removes constant and near constant variables across the dataset and retains variables that can explain maximum variance in the dataset. Moreover, NZV variables can lead to noise in the dataset, which deteriorates the quality of the model [61c].

The dataset is reduced to 1079 variables for 1671 drugs after preprocessing and the correlation

matrix of variables is shown in **Figure 2**. As seen from the correlation matrix, most of the variables are highly correlated and therefore a dimensionality reduction step is included in the workflow. The dataset is scaled after which PCA was performed using 'prcomp' function of the built-in *R Stats* package.

2.2.2. *k*-means Clustering

The first five PCs from PCA are used to cluster the dataset of 1671 approved drugs using *k*-means clustering algorithm where *k* represents the number of clusters. In this study, we have calculated the *k* value using three methods namely elbow, Nbclust and Silhouette methods. The elbow and Nbclust methods gave a similar result of k=9. However, silhouette method gave k-value as 2. Thus, *k* value is determined using the elbow method where the With-in Sum of Square (WSS) is calculated at each cluster and plotted as a function of the cluster number. The cluster number at which the addition of another cluster does not decrease the WSS value anymore is considered as the optimum number of clusters required to partition the dataset. WSS is determined for *k* values ranging from 1 to 15 (a commonly selected range) and plotted against the number of clusters (**Figure 5(a)**). In addition, the *k*-value is also determined using the 'fviz_nbclust()' function of *factoextra* R package [62], which also focuses on WSS to determine the optimum number of clusters. The WSS is determined by varying the *k* value from one to fifteen (widely used range) and plotted against the number of clusters. Upon varying the *k* value, it was noted that the WSS decreases continuously until cluster number 9 (**Figure 5 (b)**). However, the addition of further clusters does not decrease the WSS significantly. Therefore *k*-means clustering was performed using *k*-value = 9 under the widely used parameters of nstart=25 and iter.max=1000.

2.3. Biological Activity Prediction Using PASS Tool

PASS estimates biological activities of molecules using Multilevel Neighborhood of Atoms (MNA) descriptors, which are calculated based on the structural formulae, and Bayesian approach for analysis of structure-activity relationships using a training set of 1,025,468 biologically active substances [56-59, 63]. For the studied molecule and each predictable biological activity, PASS estimates two probabilities: Pa and Pi that reflect the likelihood of the belonging to the classes of "actives" and "inactives", respectively. PASS method is described in detail earlier [57, 58]. Average accuracy of prediction estimated for the whole PASS training set in leave-one-out and twenty-fold cross-validation exceeds 95%. The rest 5% could be explained by the approximations of the method used for SAR analysis, incompleteness of the training set, as well as the possible activity cliffs, which reflect the outliers in the particular chemical series. From the 1671 drugs given

as input, PASS generated predicted biological activities for 1592 drugs while 79 drugs could not be processed. These include drug molecules with high molecular weight (MW >1250Da), carbon atoms < 3, the molecular charge of +1, etc., as listed by PASS 2017 tool. Therefore, the analysis was carried out on the 1592 drugs and 5050 biological activities were generated for every drug using the default Pa > Pi settings. From these 5050 indications, 366 indications involved in disease conditions are identified and categorized into 22 broad categories for further analysis. The predicted biological activities for the 1592 drugs are analyzed exhaustively that led to the identification of drugs that can be repurposed for the 22 categories of diseases.

Journal Pre-proof

3. Results and Discussion

3.1. Principal Component Analysis

The first five PCs were selected from PCA based on scree plot, which shows the variance captured by each PC (**Figure 3**). From **Figure 3**, it is seen that the percentage of variance explained by the first three PCs is significantly high, which further reduced up to the fifth principal component. However, the difference between the percentage of variance explained by PC5 and PC6 is not significant, hence five PCs are chosen for this study. The PCA individual factor map of the drugs is shown in **Figure 4** while the biplot is shown in **Figure S1**. From the factor map, it is observed that most of the drugs are on the center of the two axes. A few drugs are far from the center towards the left of PC2 axis. Similarly, another group of drugs is clustering to the right side of the PC2 axis. Analysis of the factor map provides insights into the possible clustering patterns in the dataset, which can be further verified upon cluster analysis. We further investigated the percentage contribution of the variables towards PC1, as this component explains the maximum variance in the dataset. The percentage contribution of the first few variables towards PC1 is shown in **Figure S2**. In addition, the variables are also analysed based on cos2 values, which indicate their quality of representation on the first two PCs, as shown in **Figure S3** [64]. Therefore, PCA has successfully reduced many latent variables (i.e. 1444 descriptors) to fewer orthogonal PCs which are used further to cluster the drug molecules.

3.2. *k*-means clustering

The 1671 drugs were grouped into nine clusters and the plot and their respective size is shown in **Figure 7**. The 2D and 3D cluster plots of the nine clusters are shown in **Figure 6 (a)** and **(b)**, respectively. **Tables 1** show a few representative drugs from each of the nine clusters. We identified repurposable drugs from the drug clusters based on structural similarity. Cluster 8 contains a large number of broad-spectrum β-lactam antibiotics (i.e. cephalosporins) (**Table 1**), penicillin derivatives (penams) (**Table S1**) and monobactams since all the molecules shared a common β- lactam ring and thus have clustered together based on both structural and therapeutic similarity. Similarly, many antihypertensive drugs have been grouped in cluster 8 based on both structural and therapeutic similarities. Cluster 6 (**Table 1**) also contains drugs that are not only structurally but also therapeutically similar to each other. For example, a group of cyclooxygenase (COX) inhibitors called the Non-Steroidal Anti-inflammatory Drugs (NSAID) analgesics had grouped in cluster-6. Similarly, sulfonamide and sulfanilamide drugs have clustered together based on structural and therapeutic similarity. Likewise, a group of anti-anxiety and anticonvulsant drugs which are commonly referred to as neurological drugs have clustered together in cluster 2

(**Table S2**).

Cluster 9 (**Table 1**) is the largest cluster based on size, contains both the clustering patterns. It contains five imidazole-based antifungals used for Tinea infections that have clustered together based on structural and therapeutic similarity (**Table S3**). Likewise, five triazolobenzodiazepine drugs and their analogues used as anti-anxiety and anticonvulsant medicines have also clustered together based on structural and therapeutic similarity in cluster 9. This cluster also contains a subgroup of twenty-two drugs containing phenothiazine- based antipsychotic drugs, which have clustered together with antiarrhythmic and antimigraine drugs (Moricizine and Dimetotiazine respectively) and also share a common phenothiazine substructure. Hence therapeutically different drugs which are similar structurally have clustered together. As in cluster 9, cluster 2 also contains drugs based on structural similarity, which may or may not have the same therapeutic uses. Three lipoglycopeptide antibiotics, namely Vancomycin, Dalbavancin and Telavancin, have clustered together based on their structural and therapeutic similarity (**Table 1**).

Similarly, three vasoactive drugs Desmopressin, Felypressin, and Terlipressin, are antidiuretics grouped but they have structural and therapeutic similarity. Likewise, three echinocandin antifungal drugs, namely, Anidulafungin, Caspofungin, and Micafungin, are both structurally and therapeutically similar **(Table S4)**. Besides the above clustering patterns, we have also obtained drugs that have clustered together based on structural similarity despite having different therapeutic uses. These include Afamelanotide, Ceruletide, Gonadorelin, Goserelin, Sincalide, and Triptorelin, which have clustered together despite their different therapeutic uses. **Table S5** shows a few more subgroups of drugs of cluster 7, which have clustered together despite their different therapeutic indications. As in cluster 2, cluster 4 also contains drugs with different therapeutic uses but has grouped within the same cluster due to structural similarity. The drugs of cluster 4 are shown in **Table S6**. It is observed that two first and second-line antitubercular drugs Ethionamide and Isoniazid bearing a pyridine ring have clustered together in cluster 4 but not in cluster 3, where two other Rifamycin derivatives antitubercular drugs Rifampicin and Rifapentine, are found (**Table 1**). This indicates that clustering is directed by structural variation in the drugs wherein difference in the drug scaffold has led to the clustering of four antitubercular drugs in two different clusters. Other examples include Mercaptopurine and Tioguanine, which share a common substructure although they are used as anticancer and nephrological drugs, respectively (**Table S6**).

Cluster 7 contains several anesthetics and drugs used for neurological disorders including depression, seizures, Parkinson's disease, and insomnia based on structural and therapeutic similarity. Cluster 5 (**Table S7**) contains several steroid-based anti-inflammatory drugs like corticosteroids most of which bear a common steroid substructure. Also, this cluster contains a

series of opioid drugs used in extreme pain management. Cluster 1 (**Table 1**) is the smallest in size, contains four platinum-based antineoplastic drugs (alkylating agents), all of which have a tetra coordinated platinum group in common. Hence the drugs of this cluster have grouped based on both structural and therapeutic similarity. A similar pattern of grouping is observed in cluster 3). The first subgroup includes three Rifamycin derivative antibiotic drugs: Rifaximin, Rifampicin, and Rifapentine, sharing a common macrocyclic substructure. These drugs have different uses, with the former being used for Traveller's diarrhea while the latter two drugs are the widely used first-line and latent TB antitubercular drugs [65] (**Table 1**). The second subgroup includes anti-arrhythmic agents belonging to the cardiovascular group of drugs grouped based on their structural and therapeutic similarity. The next group includes a subgroup of the Tetracycline class of antibiotics having a structurally similar fused tetracyclic nucleus as the common substructure. Similarly, macrolide and aminoglycoside antibiotics, antiretroviral, anti-hepatitis-C drugs have also clustered into subgroups bearing their respective common scaffolds. The drugs and their respective scaffolds of cluster 3 are shown in **Table S8**. These results suggest that similarity among molecular structures of the drugs that may or may not have the same therapeutic use has led to the datasets efficient clustering. Such clustering, driven by the structural similarity between therapeutically dissimilar drugs can provide novel insights into identifying new uses for the existing drugs. The probability of the drugs that can be repurposed for new indications is further verified based on the predicted biological activities obtained from PASS analysis. Therefore, the study provides insights into how clustering of drugs based on their structural similarity can be explored as a tool for drug repurposing.

3.3. Analysis of Predicted Biological Activities

PASS generated 5050 predicted activities for 1592 drugs from which 366 indications that play a role in diseases or disease conditions are selected. Further to simplify the analysis, the 366 indications are broadly classified into 22 categories, as shown in **Figure 8**. The indications obtained for every drug include both original and repurposable ones. To identify repurposable indications, a cut-off of 0.5 Pa value is considered and all indications with Pa value ≥0.5 are selected for every drug. This step is executed using an Excel Visual Basic for Application (VBA) macro script (**Annexure-A, SI**) which automatically lists all the indications having Pa ≥0.5 for each of the 1592 drugs. In the next step, all the obtained indications are sorted one below the other using another VBA macro script (**Annexure-B, SI**) from which the repurposable indications for every drug are identified. Repurposable indications are identified based on two criteria (1) indications that are the same as the drug's original therapeutic indications are eliminated (2) already reported

repurposable indications are removed. As a result, all drugs with no repurposable indications left in the range of 0.5-1.0 after removing the original indications get subsequently eliminated. Hence out of 1592 drugs, 1423 unique drugs have obtained 13,741 repurposable indications.

These 13,741 indications belong to the 22 categories from which repurposable drugs for 12 clinically significant categories are selectively reported as shown in **Figure 9 (a-l)**. In the neurological category, the maximum number of drugs can be repurposed as antineurotic drugs (177), anti-inflammatory (145), sedative (113), analgesics (107) followed by Parkinson's disease (48), Alzheimer's disease (22), anticonvulsants (20), antidyskinetic (9) and Lateral sclerosis (5). Among the psychiatric diseases, the highest numbers of repurposable drugs are obtained for insomnia (82), mood disorders (57) followed by depression (49), schizophrenia (41), anxiety (35) and Attention-Deficit/Hyperactivity Disorder (ADHD) (5). Interestingly Deferiprone a thalassemia drug, showing a Pa value of 0.53 for acute neurological disorder in the present study, has reported clinical activity for Parkinson's disease in a clinical trial [66]. In the infectious disease category, the maximum number of drugs can be repurposed as antivirals for yellow fever (arbovirus) (267), HIV-RT inhibitor (91), aseptic meningitis, encephalitis (picornaviruses) (55), hepatitis (A,B and C) (18), herpes (12) and pox infections (12).

A further investigation is required to identify the targets for repurposing the drugs obtained in our study. In addition to the antivirals, the infectious disease category also includes antiprotozoals for *Plasmodium* infections like malaria (*Plasmodium falciparum* and *Plasmodium vivax*) (33), antidiarrheals (17), parasitic protozoan infections like Chagas disease (*Trypanosoma*) (8), Trichomoniasis (*Trichomonas*) (7), Leishmaniasis (*Leishmania)* (4) and antibacterials like antimycobacterial (*Mycobacterium*) (12) and *helicobacter* infections (6). Interestingly, among the twelve obtained antimycobacterial drugs, two drugs, namely Daunorubicin and Gatifloxacin, originally anticancer drugs have experimentally reported antitubercular activity [67-69]. From the literature, antimalarial drugs such as Artemisinin and its derivatives have been reported to have repurposable activity against Leishmaniasis and other parasitic diseases [70]. In our study, two antimalarial drugs, Artemether and Artesunate, are identified for repurposing against Leishmaniasis, with Pa of 0.93 and 0.87, respectively. Artesunate has shown clinically reported inhibitory activity against Cytomegalovirus (CMV) and Hepatitis has shown Pa values of 0.77 and 0.60 against CMV and Hepatitis-B in the current study, respectively [71]. Likewise, under the cancer category, a maximum number of drugs for repurposing are obtained for anticarcinogenic (136), pre-neoplastic (125), antimetastatic (84), antileukemic (37), antimutagenic (27), conditions antineoplastic antimetabolite (24), antineoplastic alkylator (16) and antineoplastic enhancer (16) conditions. Among the drugs obtained for the antimutagenic class are the phenothiazine-based CNS

drugs like Fluphenazine (Pa=0.54) and Chlorpromazine (Pa=0.55) which have reported antimutagenic activity in the literature [72-75]. Similarly, thiazine-based drugs like Acepromazine, Trifluoperazine, and Trifluoperazine are some other CNS drugs identified under antimutagenic conditions in our analysis. In the metabolic disorders category, the maximum number of repurposable drugs are obtained for antidiabetic (72), obesity (60), followed by type II diabetes (15) and hyperammonemia (8). In the systemic and hematological disease category, the maximum no of repurposable drugs is obtained for hyperthermia (580) followed by anemia (416), diuretics (72), vasodilators (70), antithrombotics (60) and hypertensives (20). Identifying repurposable drugs for rare diseases is another significant finding of the study. Rare diseases or orphan diseases are diseases that affect a very small percentage of the population (1 in 2000) [76]. We also identified drugs that can be repurposed for 12 rare diseases, shown in **Figure 9 (f)**. In the rare disease category, the maximum numbers of repurposable drugs are obtained for Adenomatous polyposis (142). This is followed by Crohn's (54) and Prion's disease [77] (22), which are prevalent worldwide and have no known drugs. This is followed by Wilson's disease (17), Multiple sclerosis (18), Muscular dystrophy (17), Cystic fibrosis (16), Sickle-cell anemia (11), Huntington's disease (6), Myasthenia gravis (4), Paget's disease (3) and Gaucher disease (1).

Interestingly, some of the drugs identified in the current study have reported experimental activities for their repurposable indications as discussed above. Further, we went on to identify drugs that can be repurposed for the maximum number of indications. For this purpose, a higher cut-off of 0.7 Pa value is considered to ensure that the best molecules are selected. Among the repurposable indications obtained above, all indications showing Pa value ≥ 0.7 and having a high count of total repurposable indications are selected and the top 20 drugs with their top repurposable indication and Pa value is reported in **Table 2**. *k*-means results provided subgroups of structurally similar drugs having different therapeutic uses. We hypothesized that such structurally similar but therapeutically different drugs can be potentially repurposed for one or more native indications of the members of the same cluster based on their predicted activity values obtained in PASS analysis. **Table 3** shows a list of 66 repurposable drugs from the nine clusters that are structurally similar but have different therapeutic indications identified in this study using our novel ML and SAR based combined approach. It is also interesting to see that among the 66 drugs of Table 3, some pairs of drugs which have different therapeutic uses not only share a common substructure but also a common target and may therefore operate through a common mechanism of action. For example, Amantadine and Memantadine which are used for Influenza and Alzheimer's disease respectively are known glutamate receptor antagonists. Similarly, Sulbactam and Clavulunate both target bacterial beta lactamase and have a similar mechanism of action. Everolimus which is used to

prevent organ rejection binds to the same target serine/Threonine-protein kinase mTOR as Sirolimus and Temsirolimus; all three of which have grouped within the same cluster. These results further ascertain the DR hypothesis considered in this study. In order support the findings, four drug molecules have been selected that has repurposable indications as antimycobacterial and docked against 20 antitubercular targets. The four drug molecules have docked well with good docking score support the findings in this study. Overall, our approach of combining ML techniques with SAR predictions for DR is based upon the reliable predictions made by PASS that have successfully contributed to identifying repurposable candidates confirmed through experimental studies [78-81]. These studies have shown that the prediction from the PASS analysis and data-driven approaches have the potential to identify alternative drug molecules for various disease conditions.

4. Conclusions

In the current study, molecular descriptors of approved drugs were analyzed using unsupervised ML techniques like PCA followed by *k*-means clustering in combination with biological activity predictions, as a combined approach for drug repurposing. PCA is employed as a dimensionality reduction tool to reduce the multi-collinear molecular descriptors into fewer low dimensional PCs that can potentially influence the clustering patterns of drugs. The application of PC instead of the latent variables to perform *k*-means has led to the clustering of drugs based on structural similarity. In every cluster, we obtained drugs that grouped based on structural and/or therapeutic similarity, sharing a common substructure. We showed that such therapeutically different drugs having a common substructure that has grouped in a cluster can be potentially repurposed for the native indications of the other members of the same cluster. Following this hypothesis, we identified 66 therapeutically different drugs from the nine clusters that could be repurposed for the native indications of the other members of the cluster, based upon their high Pa value obtained from PASS analysis. The exhaustive biological activity analysis in PASS led to the identification of 1423 unique repurposable candidates for 366 new disease indications considered in the study. Interestingly, many drugs that have appeared in our results have evidence of being clinically or experimentally repurposed. Through our analysis, we have also identified 20 top drugs that can be repurposed for the maximum number of indications within the considered 366 indications. Further, analysis of the PASS predictions has been useful in providing significant repurposable indications for every drug which can serve as a starting point before escalating a molecule for experimental DR studies. Hence, our combined approach helped in delineating the relationship between the drug molecular descriptors and their repurposable activities. This study serves as a milestone in the area

of DR towards *in silico* identification of a large number of repurposable indications for approved drugs using QSAR approach. Moreover, the study also features the application of machine learning techniques towards the structure-based clustering of approved drugs which will guide for optimizing the structures of new molecules/existing drugs to repurpose them for new indications.

Data availability

The scripts used to carry out PCA and *k*-means clustering are available on the git repository https://github.com/Sireesiru/Drug_repurposing.

Declaration of interest

None

Acknowledgments

The authors thank DST grant (project No: INT/RUS/RSF/12) and RSF grant (project No: 16-45-02012) for the financial support. GNS thanks DST for the award of J C Bose National fellowship. MSS thanks DST for the grant (project No: SR/WOS-A/CS-1091/2014). VVP thanks for the support by the Russian Federation Fundamental Research Program for the long-term period for 2021-2030.

Journal Pre-proof

References

1. T.T. Ashburn and K.B. Thor, Drug Repositioning: Identifying and Developing New Uses for Existing Drugs, Nat. Rev. Drug. Discov. 3 (2004) 673-683.
2. T.N. Raju, The Nobel chronicles. 1988: James Whyte Black, (b 1924), Gertrude Elion (1918-99), and George H Hitchings (1905-98), Lancet. 355 (2000) 1022.
3. F. Pammolli, L. Magazzini L and Riccaboni M, The Productivity Crisis in Pharmaceutical R&D, Nat. Rev. Drug Discov. 10 (2011) 428-438.
4. M.J. Waring, J. Arrowsmith, A.R. Leach, P.D. Leeson, S. Mandrell, R.M. Owen, G. Pairaudeau, W.D. Pennie, S.D. Pickett, J. Wang, O. Wallace and A. Weir, An Analysis of the Attrition of Drug Candidates from Four Major Pharmaceutical Companies, Nat. Rev. Drug Discov. 14 (2015) 475-486.
5. S. Pushpakom, F. Iorio, P.A. Eyers, K.J. Escott, S. Hopper, A. Wells, A. Doig, T. Guilliams, J. Latimer, C. McNamee, A. Norris, P. Sanseau, D. Cavalla and M. Pirmohamed. Drug Repurposing: Progress, Challenges and Recommendations, Nat. Rev. Drug Discov. 2019, 1 (2019) 41-58.
6. C.R. Chong, D.J. Sullivan, New Uses for Old Drugs, Nature. 448 (2007) 645-646.
7. E. Tambuyzer, Rare Diseases, Orphan Drugs and Their Regulation: Questions and Misconceptions, Nat. Rev. Drug Discov. 9 (2010) 921-929.
8. https://www.fda.gov/ForIndustry/DevelopingProductsforRareDiseasesConditions/
9. Z. Wei, S. Wei and S. Anton, Drug Repurposing Screens and Synergistic Drug-Combinations for Infectious Diseases, Br. J. Pharmacol. 175 (2018) 181-191.
10. S.K. Michael, P. Eric, P. Mark and H. Denton, An Analysis of FDA-approved Drugs for Infectious Disease: Antibacterial Agents, Drug. Discov. Today. 19 (2014)1283-1287.
11. S.R. Garcia, R.G.D. Rio, A.S. Villarejo, G.D. Sweet, F. Cunningham, D. Barros, L. Ballell, A.M. Losana, S.F. Bazaga and C.J Thompson, Repurposing Clinically Approved Cephalosporins for Tuberculosis Therapy, Sci. Rep. 6 (2016) 34293.
12. G.L. Law, J.T. Go, M.J. Korth, M.G. Katze, Drug repurposing: A Better Approach for Infectious Disease Drug Discovery?, Curr. Opin. Immunol. 5 (2013) 588-592.
13. A. Maitra, B. Sade, M. Shaik, D. Evangelopoulos, I. Abubakar, D.M. Timothy, L. Marc and B. Sanjib. Repurposing Drugs for Treatment of Tuberculsosis: A Role for Non-Steroidal Anti-Inflammatory Drugs, Br. Med. Bull. 118 (2016) 145-155.
14. J.D. Guzman, D. Evangelopoulos, A. Gupta, K. Birchall, S. Mwaigwisya, B. Saxty, T.D. MeHugh, S. Gibbons, J. Malkinson and S. Bhakta. Antitubercular Specific Activity of Ibuprofen and the Other 2-Arylpropanoic Acids Using the HT-SPOTi Whole-Cell Phenotypic Assay, BMJ Open. 3 (2013) e002672.
15. R. Gayatri, R. Nagasuma, S. Chandrac and S. Narayanaswamy, Recognizing drug targets using evolutionary information: implications for repurposing FDA-approved drugs against *Mycobacterium tuberculosis* H37Rv, Mol. BioSyst. 11 (2015) 3316-3205.
16. H.F. Chambers, D. Moreau, D. Yajko, C. Miick, C. Wagner, C. Hackbarth, S. Kocagöz, E. Rosenberg, W.K. Hadley and H. Nikaido, Antimicrob. Agents Chemother. Can penicillins and other beta-lactam antibiotics be used to treat tuberculosis?, 39 (1995) 2620-2624.
17. K.T. Andrews, G.Fisher and T.S.Skinner-Adams, Drug repurposing and human parasitic protozoan diseases, Int. J. Parasitol. Drugs Drug Resist. 4 (2014) 95-111.
18. N. Kumar, H. Sarma and G.N. Sastry, Repurposing of approved drug molecules for viral infectious diseases: a molecular modelling approach, J. Biomol. Struct. Dyn. 2 (2021) 1-17.
19. D. Carrella, I. Manni, B. Tumaini, R. Dattilo, F. Papaccio, M. Mutarelli, F. Sirci, C.A. Amoreo, M. Mottolese, M. Lezzi, L. Ciolli, V. Aria, R. Bosotti, A. Isacchi, F. Loreni, A. Bardelli, V.E. Avvedimento, D. Bernardo and L. Cardone, Computational drugs repositioning identifies

inhibitors of oncogenic PI3K/AKT/P70S6K-dependent pathways among FDA-approved compounds, Oncotarget. 7 (2016) 58743-58758.

20. A. Cabana, K. Pisarczyka, K. Kopacza, K. Kapuśniaka, K. Toumi, C. Rémuzatc and A. Kornfeldc, J. Mark. Filling the gap in CNS drug development: evaluation of the role of drug repurposing, Access. Health. Policy. 5 (2017) 1299833.
21. L. Grenier and P. Hu, Computational drug repurposing for inflammatory bowel disease using genetic information, Comput. Struct. Biotechnol. J. 17 (2019) 127-135.
22. US Preventive Services Task Force. Final recommendation statement. Aspirin use to prevent cardiovascular disease and colorectal cancer: preventive medication. US Preventive Services Task Force https:// www.uspreventiveservicestaskforce.org/Page/Document/ RecommendationStatementFinal/aspirin-to-preventcardiovascular-disease-and-cancer, 2017
23. G. Rena and C. C. Lang, Repurposing Metformin for Cardiovascular Disease, Circulation. 137 (2018) 422-424.
24. X. Hanqing, L. Jie, X. Haozhe and W. Yadong, W. Review of Drug Repositioning Approaches and Resources, Int. J. Biol. Sci. 14 (2018)1232-1244.
25. D. Brehmer, Z. Greff, K. Godl, S. Blencke, A. Kurtenbach, M. Weber, S. Muller, B. Klebl, M. Cotton, G. Keri, J. Wissing and H. Daub, Targets of Gefitinib, Cancer Res. 65 (2005) 379-382.
26. D.M. Molina, R. Jafari, M. Ignatushchenko, T. Seki, A. Larsson, C. Dan, L. Sreekumar, Y. Cao and P. Nordlund, Monitoring Drug Target Engagement in Cells and Tissues Using the Cellular Thermal Shift Assay, Science. 341 (2013) 84-87.
27. G.F. Wilkinson and K. Pritchard, In Vitro Screening for Drug Repositioning, J. Biomol. Screen. 2 (2015) 167-179.
28. (a) G. Jin and S.T.C. Wong, Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines, Drug Discov Today. 19 (2014) 637-644. (b) M.R. Hurle, L. Yang, Q. Xie, D.K. Rajpal, P. Sanseau and P. Agarwal, Computational Drug Repositioning: From Data to Therapeutics Clin. Pharmacol. Ther. 93 (2013) 335-341.
29. R.A. Hodos, B.A. Kidd, S. Khader, B.P. Readhead and J.T. Dudley, In Silico Methods for Drug Repurposing and Pharmacology, Wiley Interdiscip. Rev. Syst. Biol. Med. 8 (2016) 186-210.
30. C. Andronis, A. Sharma, V.V. Spyros and D.A. Persidis, Literature mining, ontologies and information visualization for drug repurposing, Brief Bioinform. 12 (2011) 357-368.
31. M. Joseph, J. Simon, W. Peter and W. Anil, An integrated data driven approach to drug repositioning using gene-disease associations, PLoS-One. 11 (2016) e0155811.
32. D. Vogrinc and T. Kunej, Drug repositioning: computational approaches and research examples classified according to the evidence level, Discoveries, 5 (2017) e75.
33. A.S. Reddy, S.P. Pati, P.P. Kumar, H.N. Pradeep and G.N. Sastry. Virtual screening in drug discovery - a computational perspective, Curr. Protein Pep. Sci. 8 (2007) 329-51.
34. H.K. Srivastava and G.N. Sastry, Molecular dynamics investigation on a series of HIV protease inhibitors: assessing the performance of MM-PBSA and MM-GBSA approaches, J. Chem. Inf. Model. 52 (2012) 3088-98.
35. P. Badrinarayan and G.N. Sastry, Virtual high throughput screening in new lead identification, Comb. Chem. High. Throughput. Screen. 14 (2011) 840-60.
36. G.K. Ravindra, G. Achaiah and G.N. Sastry, Molecular modeling studies of phenoxypyrimidinyl imidazoles as p38 kinase inhibitors using QSAR and docking, Eur. J. Med. Chem. 43 (2008) 830-8.
37. M.H. Bohari and G.N. Sastry, FDA approved drugs complexed to their targets: evaluating pose prediction accuracy of docking protocols, J. Mol. Model. 18 (2012) 4263-74.
38. P. Badrinarayan and G.N. Sastry, Virtual screening filters for the design of type II p38 MAP kinase inhibitors: a fragment-based library generation approach, J. Mol. Graph. Model. 34 (2012) 89-100.

Journal Pre-proof

39. C. Choudhury, U.D. Priyakumar and G.N. Sastry, Dynamics based pharmacophore models for screening potential inhibitors of mycobacterial cyclopropane synthase, J. Chem. Inf. Model. 55 (2015) 848-60.
40. A.S. Gaur, S. Nagamani, K. Tanneeru, D. Druzhilovskiy, A. Rudik, V.V. Poroikov and G.N. Sastry, Molecular Property Diagnostic Suite for Diabetes Mellitus ($MPDS^{DM}$): An Integrated Web Portal for Drug Discovery and Drug Repurposing, J. Biomed. Inform. 85 (2018) 114-125.
41. V. Jha, N.R. Rameshwaram, S. Janardhan, R. Raman, G.N. Sastry, V. Sharma, J.S. Rao, D. Kumar and S. Mukhopadhyay, Uncovering Structural and Molecular Dynamics of ESAT-6: β2M Interaction: Asp53 of Human β2-Microglobulin Is Critical for the ESAT-6: β2M Complexation, J. Immunol. 203 (2019) 1918-1929.
42. S. Nagamani, R. Sahoo, G. Muneeswaran, G.N. Sastry Data Science Driven Drug Repurposing for Metabolic Disorders, In silico Drug Design. (2019) 191-227.
43. U. Storz, How Approval History Is Reflected By a Corresponding Patent Filing Strategy. MAbs. 6 (2014) 820-837.
44. F. Iorio, T. Rittman, H. Ge and M. Menden, Transcriptional Data: A New Gateway To Drug Repositioning?, Drug Discov. Today. 18 (2013) 350-357.
45. J.T. Dudley, T. Deshpande and A.J. Butte, Exploiting drug-disease relationships for computational drug repositioning, Brief Bioinform. 12 (2011) 303-311.
46. K.A. Markey, R. Ottridge, J.L. Mitchell, C. Rick, R. Woolley, N. Ives, P. Nightingale and A.J. Sinclair, Assessing the Efficacy and Safety of an 11β-Hydroxysteroid Dehydrogenase Type 1 Inhibitor (AZD4017) in the Idiopathic Intracranial Hypertension Drug Trial, IIH: DT: Clinical Methods and Design for a Phase II Randomized Controlled Trial, JMIR. Res. Protoc. 6 (2017) e181.
47. C.H. Schein, Repurposing approved drugs on the pathway to novel therapies, Med. Res. Rev. 40 (2020) 586-605.
48. J.L.M. Franco, Drug Repurposing For Epigenetic Targets Guided By Computational Methods, Epi-Informatics Discovery and Development of Small Molecule Epigenetic Drugs and Probes, Elsevier Inc. 2016.
49. D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M.H. Hassanali, P. Stothard, Z. Chang and J. Woolsey, DrugBank: A Comprehensive Resource For In Silico Drug Discovery and Exploration, Nucleic Acids Res. 34 (2006) D668-D672.
50. O. Ursu, J. Holmes, J. Knockel, C. J. Bologa, J.J. Yang, S.L. Mathias, S.J. Nelson, T.I. Opera, DrugCentral: Online Drug Compendium, Nucleic Acids Res. 45 (2016) D932-D939.
51. S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B.A. Shoemaker, P.A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang and E.E. Bolton, PubChem 2019 Update: Improved Access to Chemical Data, Nucleic Acids Res. 47 (2016) D1102-D1109.
52. Y.H. Li, C.Y. Yu, X.X. Li, O.P. Zhang, J. Tang, Q. Yang, T. Fu, X. Zhang, X. Cui, G. Tu, Y. Zhang, S. Li, F. Yang, Q. Sun, C. Qin, X. Zeng, Z. Chen, Y.Z. Chen and F. Zhu, Therapeutic Target Database Update 2018: Enriched Resource for Facilitating Bench-to-Clinic Research of Targeted Therapeutics, Nucleic Acids Res. 46 (2018) D1121-D1127.
53. https://www.centerwatch.com
54. https://www.fda.gov
55. (a) M. Junshui, R.P. Sheridan, A. Liaw, G.E. Dahl and V. Svetnik, Deep Neural Nets as a Method for Quantitative Structure-Activity Relationships, J. Chem. Inf. Model. 55 (2015) 263-274 (b) H. Chen, O. Engkvist, Y. Wang, M. Olivecrona and T. Blaschke, The rise of deep learning in drug discovery, Drug Discov Today. 23 (2018) 1241-1250.(c) C. Lopez, S. Tucker, T. Salameh and C. Tucker, An Unsupervised Machine Learning Method for Discovering Patient Clusters Based on Genetic Signatures, J. Biomed. Inform. 85 (2018) 30–39. (d) Z. Zhao, J. Qin, Z. Gou, Y. Zhang and Y. Yang, Multi-Task Learning Models for Predicting Active Compounds, J. Biomed. Inform. 108 (2020) 103484. (e) R. Winter, F. Montanari, F.

Noé and D.A. Clevert. Learning Continuous and Data-Driven Molecular Descriptors by Translating Equivalent Chemical Representations, Chem Sci. 10 (2019) 1692-1701. (f) A. Mayr, G. Klambauer, T. Unterthiner and S. Hochreiter, Deeptox: toxicity prediction using deep learning, Front. Environ. Sci. 3 (2016) 80.

56. A. Lagunin, A. Stepanchikova, D. Filimonov and V.V. Poroikov, PASS: Prediction of Activity Spectra for Biologically Active Substances, Bioinformatics. 16 (2000) 747-748.
57. D.A. Filimonov, A.A. Lagunin, T.A. Gloriozova , A.V. Rudik, D.S. Druzhilovskiy, P.V. Pogodin and V.V. Poroikov, Prediction of the Biological Activity Spectra of Organic Compounds Using the Pass Online Web Resource, Chem. Heterocycl. Compd. 50 (2014) 444-457.
58. D.A. Filimonov, D.S. Druzhilovskiy, A.A. Lagunin, T.A. Gloriozova, A.V. Rudik, A.V. Dmitriev, P.V. Pogodin and V.V. Poroikov, Computer-aided prediction of biological activity spectra for chemical compounds: opportunities and limitations, Biomed. Chem. Res. Methods 1 (2018) e00004.
59. V.M. Bezhentsev, D.S. Druzhilovskiy, S.M. Ivanov, D.A. Filimonov, G.N. Sastry and V.V. Poroikov, Web Resources for Discovery and Development of New Medicines, Pharm. Chem. J. 51 (2017) 91-99.
60. C.W. Yap, PaDEL-descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints, J. Comput. Chem. 32 (2010)1466-1474.
61. (a) R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/ (b) M.J. Kunh, Building Predictive Models in R Using the caret Package Stat. Softw. 28 (2008) 1-26. (c) https://topepo.github.io/caret/pre-processing.html#nzv
62. K. Alboukadel and M. Fabian, Factoextra: Extract and Visualize the Results of Multivariate, Data Analyses, 2017.
63. V.V. Poroikov, D.A. Filimonov, T.A. Gloriozova, A.A. Lagunin, D.S. Druzhilovskiy, A.V. Rudik, L.A. Stolbov, A.V. Dmitriev, O.A. Tarasova, S.M. Ivanov, P.V. Pogodin, Computer-aided prediction of biological activity spectra for organic compounds: the possibilities and limitations, Russ. Chem. Bull. 68 (2019) 2143-2154.
64. H. Abdi and L.J. Williams. Principal Component Analysis WIREs, Comput, Stat. 2 (2010) 433-459.
65. (a) D.M. Rothstein, Rifamycins, Alone and in Combination, Cold Spring Harb. Perspect Med. 6 (2016) a027011. (b) https://www.cdc.gov/tb/topic/treatment/ltbi.html
66. D. Devos, C. Moreau, J.C. Devedjian, J. Kluza, M. Petrault, C. Laloux, A. Jonneaux, G. Ryckewaert, G. Garcon, N. Rouaix, A. Duhamel, P. Jissendi, K. Dujardin, F. Auger, L. Ravasi, L. Hopes, G. Grolez, W. Firdaus, B. Sablonnière, I. Strubi-Vuillaume, N. Zahr, A. Destee, J.C. Corvol, D. Poltl, M. Leist, C. Rose, L. Defebvre, P. Marchetti, I. Cabantchik and R. Bordet, Targeting Chelatable Iron as a Therapeutic Modality in Parkinson's Disease, Antioxid Redox Signal. 21 (2014) 195-210.
67. J.C. Rodriguez, M. Ruiz, M. Lopez and G. Royo, In Vitro Activity of Moxifloxacin, Levofloxacin, Gatifloxacin and Linezolid against *Mycobacterium Tuberculosis*, Int. J. Antimicrob. Agents. 20 (2002) 464-467.
68. J.C. Palomino and M. Anandi, Is Repositioning of Drugs a Viable Alternative in the Treatment of Tuberculosis?, Antimicrob. Chemother. 68 (2013) 275-283.
69. R. Rustomjee, C. Lienhardt, T. Kanyok, G.R. Davies, J. Levin, T. Mthiyane, C. Reddy, A.W. Sturm, F.A. Srigel, J. Allen, D.J. Coleman, B. Fourie and D.A. Mitchison, Gatifloxacin for TB (OFLOTUB) study team. A Phase II study of the Sterilising Activities of Ofloxacin, Gatifloxacin and Moxifloxacin in Pulmonary Tuberculosis, Int. J. Tuberc. Lung. Dis. 12 (2008) 128-138.

70. T.A. Katherine, F. Gillian, S. Tina and A. Skinner, Drug repurposing and human parasitic protozoan diseases, Int. J. Parasitol. Drugs. Drug. Resist. 4 (2014) 95-111.
71. E. Thomas, M.R. Romero, G.W. Dana, S. Thomas, J.G.M. Jose and M. Manfred, The Antiviral Activities of Artemisinin and Artesunate, Clin. Infect. Dis. 2008, 47, 804-811.
72. G. Kazimierz, B. Barbara, S. Katarzyna and L. Jerzy, Antimutagenic Activity of Fluphenazine in Short-Term Tests, Mutagenesis. 16 (2001) 31-38.
73. M.H. Chen, W.R. Yang, K.T. Lin, C.H. Liu, Y.W. Liu, K.W. Haung, P.M.H. Chang, J.M. Lai, C.N. Hsu, K.M. Chao, C.Y. Kao, C.Y.F. Huang, Expression-Based Chemical Genomics Identifies Potential Therapeutic Drugs in Hepatocellular Carcinoma, PLoS One. 6 (2011) e27186
74. E. Gocke, Review of the Genotoxic Properties of Chlorpromazine and Related Phenothiazines, Mutat. Res. 366 (1996) 9-21.
75. D.K. Joseph, M.C. Luz and D.S. Paul, The Effect of Substituted Phenothiazines on the Mutagenicity of Benzo[a]pyrene, Mutat Res-Fund. Mol M. 80 (1981) 259-264.
76. https://rarediseases.info.nih.gov/diseases/pages/31/faqs-about-rare-diseases
77. https://rarediseases.info.nih.gov/search?keyword=prion%20disease
78. V.V. Poroikov and D. Druzhilovskiy, Drug Repositioning: New Opportunities for Older Drugs. In: In Silico Drug Design, 1st Edition. Repurposing Techniques and Methodologies. Chapter 1. Editor: Kunal Roy. Elsevier, Academic Press (2019) 3-17.
79. A. Geronikaki, v. Kartsev, A. Petrou, M.G. Akrivou, I.S. Vizirianakis, F.M. Chatzopoulou, B. Lichitsky, S. Sirakanyan, M. Kostic, M. Smiljkovic, M. Soković, D. Druzhilovskiy and V.V. Poroikov, Antibacterial activity of griseofulvin analogues as an example of drug repurposing, Int. J. Antimicrob. Agents. 55 (2020) 105884-105895.
80. R.K. Goel, D.Y. Gawande, A.A. Lagunin and V.V. Poroikov, Pharmacological repositioning of Achyranthes aspera as an antidepressant using pharmacoinformatic tools PASS and PharmaExpert: a case study with wet lab validation, SAR QSAR Environ Res. 29 (2018) 69-81.
81. K. Lloyd, S. Papoutsopoulou, E. Smith, P. Stegmaier, F. Bergey, L. Morris, M. Kittner, H. England, D. Spiller, M.H.R. White, C.A. Duckworth, B.J. Campbell, V.V. Poroikov, V.A.P. Martins, V.A.P. Santos, A. Kel, W. Muller, D.M. Pritchard, C. Probert, M.D. Burkitt and SysmedIBD Consortium. Using systems medicine to identify a therapeutic agent with potential for repurposing in inflammatory bowel disease, Dis. Model. Mech. 13 (2020) 1-12.

Journal Pre-proof

List of tables and figures

Table 1. Representative drugs of clusters 1-9 along with their therapeutic indications and common substructure.

Table 2. List of the top 20 drugs and their top repurposable indications obtained in the study.

Table 3. Selected subgroups of drugs from the nine clusters having structurally similar but therapeutically different drugs that can be repurposed for one or more native indications of the other members of the same cluster.

Figure 1. The schematic workflow of the study.

Figure 2. The correlation matrix of the 1079 variables after pre-processing.

Figure 3. Scree plot showing the percentage of variance explained by the first ten principal components.

Figure 4. Individuals factor plot showing the scores of drugs on PC1 and PC2.

Figure 5 (a). Plot of the two techniques used to determine the number of clusters in *k*-means. (a) The plot of total within group sum of squares method (WSS) **(b).** Plot of optimum number of clusters using fviz_nbclust() function for clusters 2 to 15.

Figure 6 (a). 2D **(b)** 3D plot of the clusters of approved drugs obtained from the *k*-means clustering.

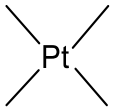
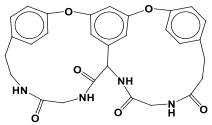
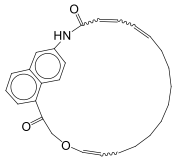
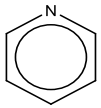
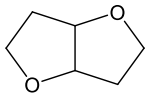
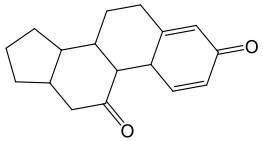
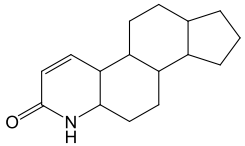
Figure 7. Plot of the nine clusters and their respective cluster size.

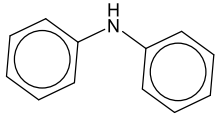
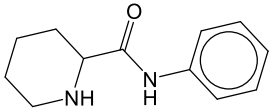
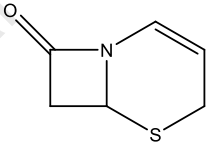
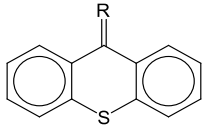
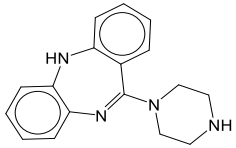
Figure 8. 366 diseases have been distributed into 22 major categories.

Figure 9 (a-l). Plots showing the distribution of repurposable drugs for 12 selected disease categories.

Journal Pre-proof

Table 1. Representative drugs of clusters 1-9 along with their therapeutic indications and common substructure.

Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4		Cluster 5	
Common sub-structure							
Drugs name	Cisplatin Nedaplatin Oxaliplatin Carboplatin	Dalbavancin Telavancin Vancomycin	*Rifaximin* Rifampicin Rifapentine	*Ethionamide* *Isoniazid* Pralidoxime	Isosorbide Dinitrate Isosorbide Mononitrate	*Clobetasone* Prednisone	Finasteride Dutasteride
Therapeutic indication	Antineoplastic (alkylating agent)	Lipoglycopeptide antibiotic	*Diarrhoea* Tuberculosis	*Tuberculosis* Organophosphate poisoning	Angina pectoris	*Eczema, psoriasis, dermatitis* Psoriatic arthritis, dermatomyositis	Benign prostatic hyperplasia

Cluster	Cluster 6	Cluster 7	Cluster 8	Cluster 9	
Common sub-structure					
Drugs name	Aceclofenac, Diclofenac, Meclofenamic acid, Mefenamic acid, Lumiracoxib, Alclofenac, Tolfenamic acid	Levobupivacaine Ropivacaine Mepivacaine Bupivacaine	Cefaclor, Cefadroxil, Cefalotin, Cefamandole, Cefapirin, Cefazolin, Cefdinir, Cefditoren, Cefepime, Cefixime, Cefmenoxime, Cefmetazole, Cefonicid, Cefoperazone, Ceforanide, Cefotaxime, Cefotetan, Cefotiam, Cefoxitin, Cefpiramide, Cefpodoxime, Cefprozil, Cefradine, Ceftaroline fosamil, Ceftazidime, Ceftibuten, Ceftizoxime, Ceftriaxone, Cefuroxime, Cephalexin, Cephaloglycin, Cephaloridine	*Thiothixene* *Chlorprothixene* Zuclopenthixol Flupentixol	*Clozapine* Olanzapine
Therapeutic indication	NSAID analgesic	Anesthetic	Cephalosporin antibiotic	*Schizophrenia* Schizophrenia, Bipolar disorder	*Schizophrenia, depression* Schizophrenia

Italicized drug names correspond to the italicized therapeutic indications

Journal Pre-proof

Journal Pre-proof

Table 2. List of the top 20 drugs and their top repurposable indications obtained in the study.

S. No.	Drug name	Drug's therapeutic indication	Top repurposing indication	Pa value	Total no. of repurposable indications of drug
1	Methyltestosterone	Testosterone Deficiency	Menorrhagia	0.95	26
2	Cholic acid	Rare Bile acid synthesis disorders, Zellweger Spectrum Disorders	Antihypercholesterolemic agent	0.95	23
3	Nitroglycerin	Vasodilator; Angina pectoris	Osteoarthritis	0.99	23
4	Fludrocortisone	Adrenocortical insufficiency in Addison's disease and treatment of salt-losing adrenogenital syndrome	Dermatitis	0.98	22
5	Arbutin	Prevent melanin formation	Anti-infective	0.96	21
6	Flumethasone	Corticosteroid-responsive dermatoses	Eye irritation	0.99	21
7	Fluoxymesterone	Hypogonadism	Antiallergic	0.96	21
8	Hydrocortamate	Anti-inflammatory to treat inflammation due to corticosteroid-responsive dermatoses	Respiratory analeptic	0.98	21
9	Aminocaproic acid	Treat severe bleeding caused by problems with the blood clotting system	Mucositis	0.91	20
10	Hydroxyprogesterone caproate	Avoid preterm birth	Menopausal disorders	0.96	20
11	Prednisolone	Conjunctivitis, rosacea, punctate keratitis, shingles and iritis	Anaemia	0.98	20
12	Testosterone	Hypogonadism	Alopecia	0.98	20
13	Testosterone-enanthate	Conditions associated with a deficiency or absence of endogenous testosterone	Antisecretoric	0.95	20
14	Testosterone-undecanoate	Treatment of testosterone deficiency	Antisecretoric	0.95	20
15	Dihydroergotamine	Migraine headache	Antiadrenergic	0.97	19
16	Levonordefrin	Hemorrhage	Cardiovascular analeptic	0.93	19
17	Metaraminol	Hypotension	Cardiovascular analeptic	0.92	19
18	Alfacalcidol	Vitamin D deficiency	Respiratory analeptic	0.98	18
19	Alprostadil	Congenital heart defects	Antisecretoric	0.98	18
20	Amcinonide	Corticosteroid-responsive dermatoses	Antiallergic	0.97	18

Table 3. Selected subgroups of drugs from the nine clusters having structurally similar but therapeutically different drugs that can be repurposed for one or more native indications of the other members of the same cluster.

S. No.	Drug name	Original indication (MOA)	Repurposable indication	Pa
1	Glucosamine	Osteoarthritis	Antineoplastic, alkylator	0.87
2	Streptozocin	Pancreatic cancer (DNA alkylation)	-	-
3	Amantadine	Influenza A virus, PD	Alzheimer's disease treatment	0.73
4	Memantine	Alzheimer's disease, dementia	Antiviral (Influenza A)	0.65
5	Levetiracetam	Epilepsy	Neurodegenerative disease	0.58
6	Piracetam	Senile dementia	-	
7	Clavulanate	Otitis	Skin infections	0.51
8	Sulbactam	Skin infections	-	
9	Temsirolimus	Antineoplastic	-	
10	Sirolimus	Immunosuppressant	Anticarcinogenic	0.95
11	Everolimus	Prevent organ rejection	Anticarcinogenic	0.96
12	Rifaximin	Traveller's diarrhea	Antituberculosic	0.98
13	Rifampicin	Tuberculosis	Diarrhea	0.98
14	Rifabutin	Tuberculosis	Diarrhea	0.86
15	DE	Migraine headache	-	
16	Ergotamine	Migraine headache	-	
17	Bromocriptine	Parkinson's disease	Antimigraine	0.99
18	CA	Prostatic carcinoma	Endometriosis treatment	0.73
19	Dydrogesterone	Endometriosis	-	
20	Medrogestone	Endometrial shedding	-	
21	MA	Breast and endometrial cancer	-	
22	Fludrocortisone	Adrenocortical insufficiency	Anticarcinogenic Menorrhagia treatment Menopausal disorders treatment	0.85 0.73 0.62
23	Hydrocortamate	Anti-inflammatory	Anticarcinogenic Menorrhagia treatment Menopausal disorders treatment	0.83 0.72 0.76
24	Hydrocortisone	Eczema, psoriasis and seborrheic dermatitis	Anticarcinogenic Menorrhagia treatment Menopausal disorders treatment	0.87 0.88 0.78
25	Progesterone	Infertility, avoid preterm birth	Anticarcinogenic	0.52
26	HC	Avoid preterm birth	Anticarcinogenic	0.81
27	Medrysone	Conjunctivitis and episcleritis	Anticarcinogenic Menorrhagia treatment Menopausal disorders treatment	0.79 0.92 0.82
28	MT	Breast cancer	-	
29	Testosterone	Breast cancer and hypogonadism	-	

Journal Pre-proof

S. No.	Drug name	Original indication (MOA)	Repurposable indication	Pa
30	ND	Breast cancer and Anaemia	-	
31	NP	Breast cancer and Anaemia	-	
32	Norethisterone	Endometriosis	Anticarcinogenic	0.69
33	FM	Hypogonadism	Anticarcinogenic	0.81
34	Tixocortol	Topical anti-inflammatory	Anticarcinogenic	0.76
35	Oxymetholone	Anaemia	Contraceptive	0.61
36	Drostanolone	Recurrent breast cancer	Contraceptive	0.58
37	Trilostane	Cushing's syndrome	Contraceptive	0.70
38	Norgestimate	Contraceptive	-	
39	Bromfenac	NSAID analgesic	-	
40	Dexketoprofen	NSAID analgesic	-	
41	Ketoprofen	NSAID analgesic	-	
42	Nepafenac	NSAID analgesic	-	
43	FA	Atherosclerosis	Anti-inflammatory	0.50
44	Zileuton	Asthma	-	
45	Stepronin	Expectorant	COPD	0.84
46	Chlorothiazide	Hypertension	Allergic reaction	0.55
47	Diazoxide	Hypertension, hypoglycemia	Antiemphysemic	0.55
48	Cytarabine	Antineoplastic antimetabolite	-	
49	Gemcitabine	Antineoplastic antimetabolite	-	
50	Zalcitabine	HIV	Antineoplastic antimetabolite	0.84
51	Floxuridine	antineoplastic antimetabolite	HSV	0.67
			Hepatitis B	0.56
			HIV	0.53
52	Idoxuridine	HSV	Antineoplastic antimetabolite	0.88
53	Stavudine	HIV	Antineoplastic antimetabolite	0.56
54	Telbivudine	Hepatitis-B	Antineoplastic antimetabolite	0.81
55	Trifluridine	HSV	Antineoplastic antimetabolite	0.82
56	Zidovudine	HIV	Antineoplastic antimetabolite	0.70
57	Alvimopan	Postoperative ileus	Analgesic	0.55
58	Anileridine	Narcotic analgesic	-	
59	Phenindamine	Allergic rhinitis and common cold	-	
60	Carbinoxamine	Allergic rhinitis and conjunctivitis	-	
61	Chlorphenamine	Allergic rhinitis and urticaria	-	
62	Bepotastine	Allergic conjunctivitis	-	
63	Bisacodyl	Constipation	Allergic reaction	0.50
64	DBP	Allergic conjunctivitis, hay fever	-	
65	Doxylamine	Allergies	-	
66	Disopyramide	Ventricular tachycardia, arrhythmia	Rhinitis treatment	0.67
			Allergic reaction	0.74

ALL: Acute Lymphoblastic Leukaemia, DE: Dihydroergotamine, CA: Cyproterone Acetate, MT: Methyltestosterone, DBP: Dexbrompheniramine, HC: Hydroxyprogesterone Caproate, MA: Megestrol acetate, NP: Nandrolone Phenpropionate, ND: Nandrolone Decanoate, FA: Fenofibric acid, FM: Fluoxymesterone, HIV: Human Immunodeficiency Virus, HSV: Herpes Simplex Virus, NSAID: Non-steroidal Anti-Inflammatory Drug

Journal Pre-proof

Journal Pre-proof

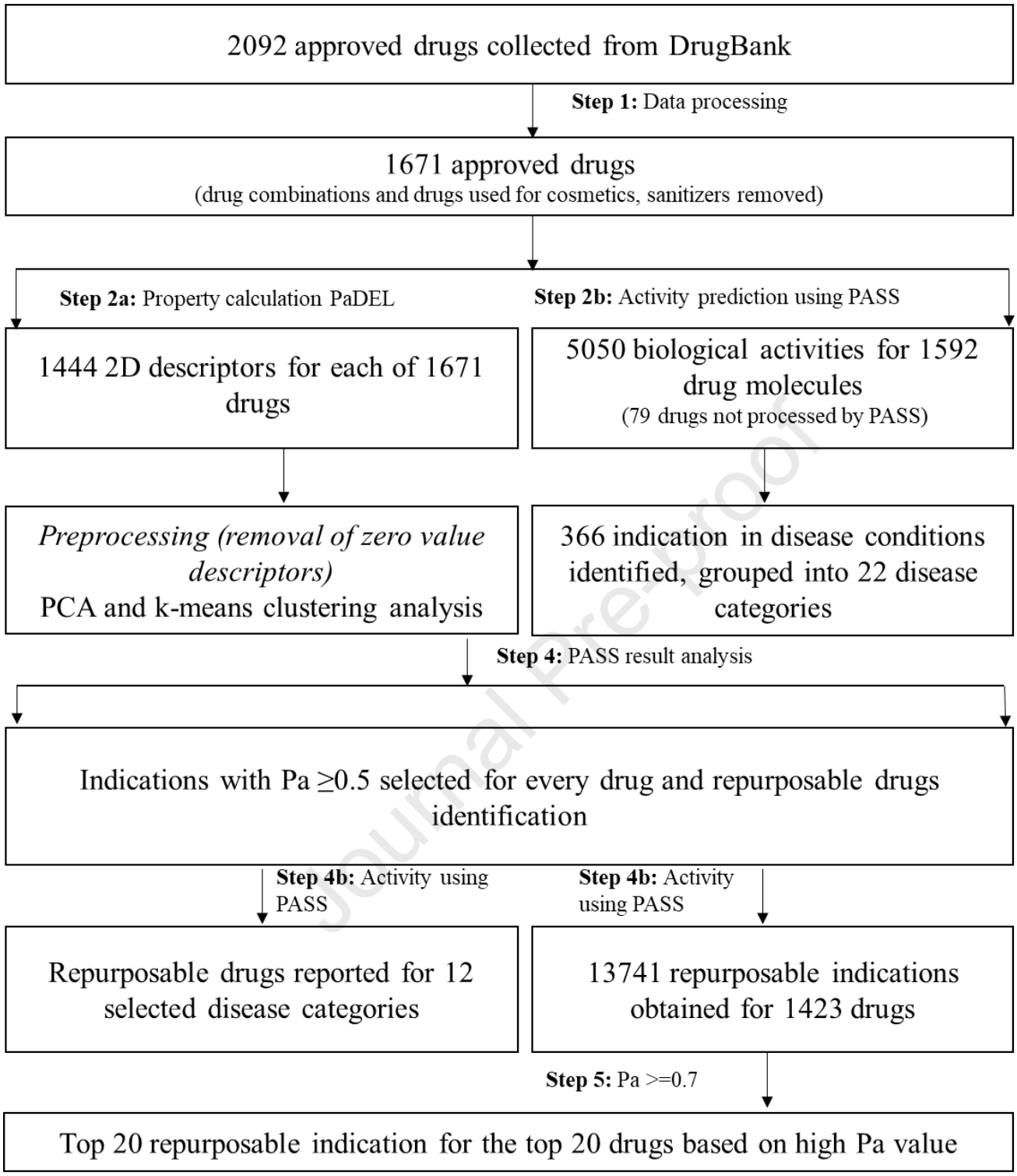


Figure 1. The schematic workflow of the study.

Figure 2. The correlation matrix of the 1079 variables after pre-processing.

Journal

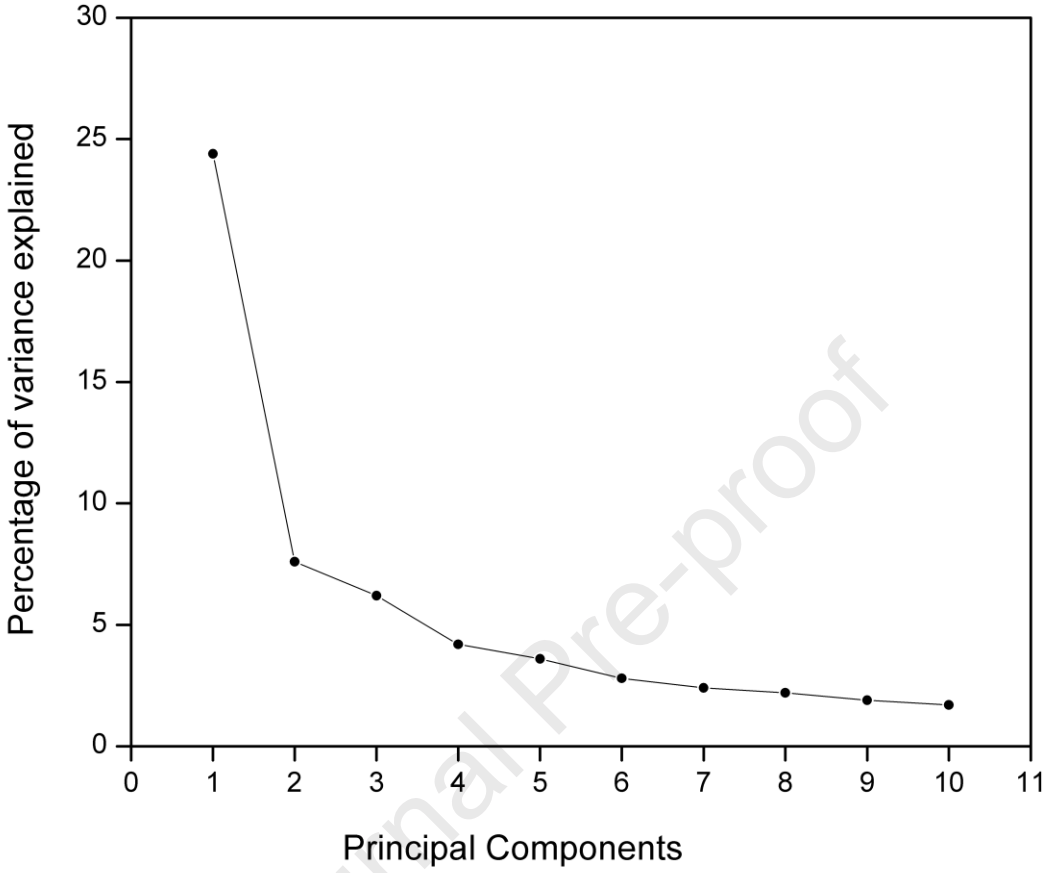


Figure 3. Scree plot showing the percentage of variance explained by the first ten principal components.

Journal Pre-proof

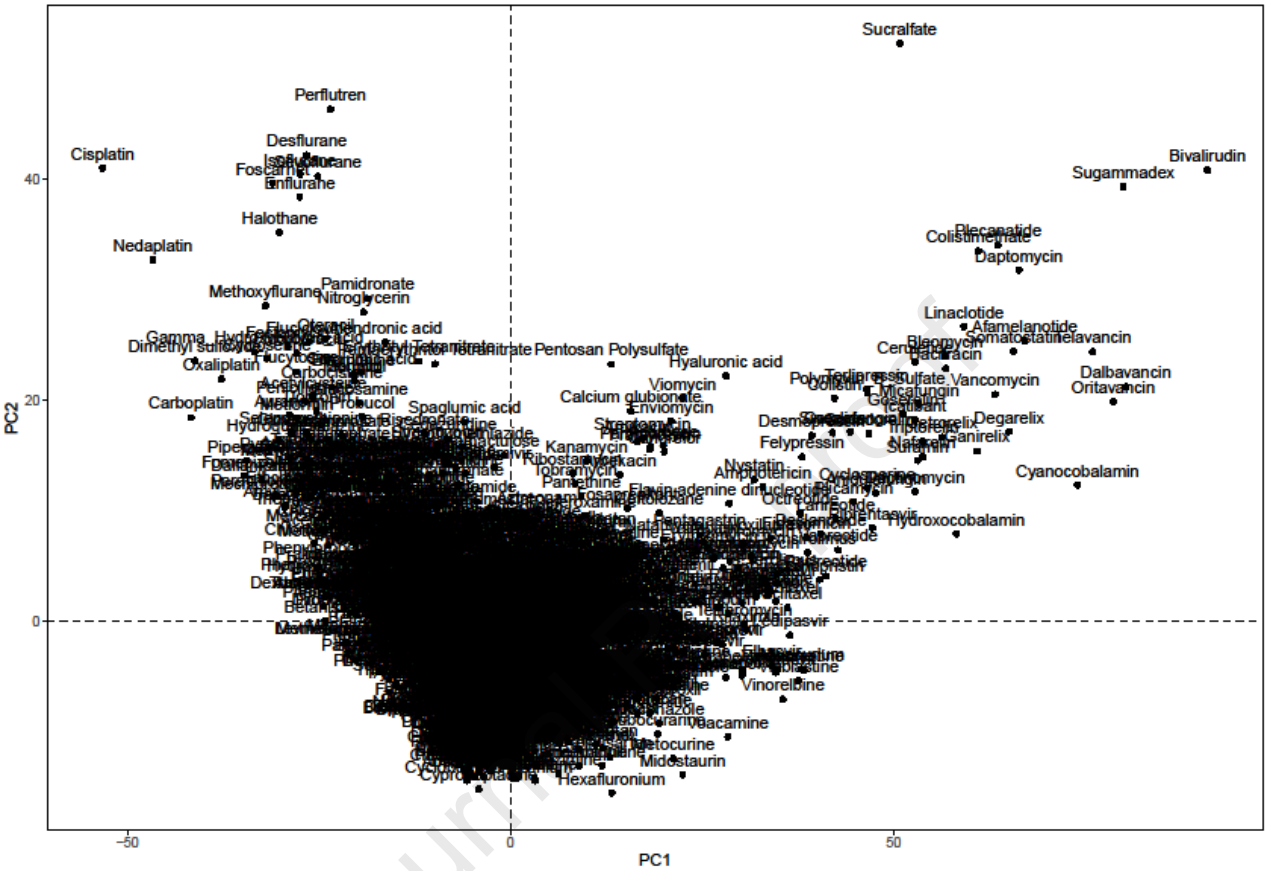


Figure 4. Individuals factor plot showing the scores of drugs on PC1 and PC2.

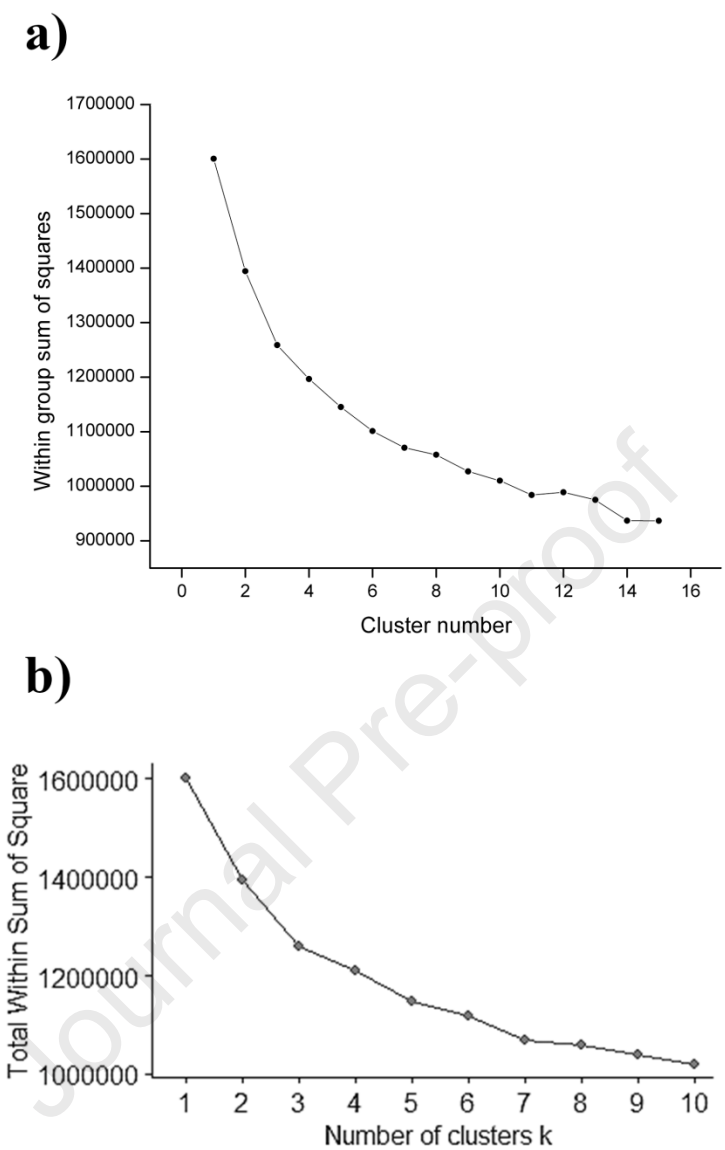


Figure 5 (a). Plot of the two techniques used to determine the number of clusters in *k*-means. (a) The plot of total within group sum of squares method (WSS) **(b).** Plot of optimum number of clusters using fviz_n bclust() function for clusters 2 to 15.

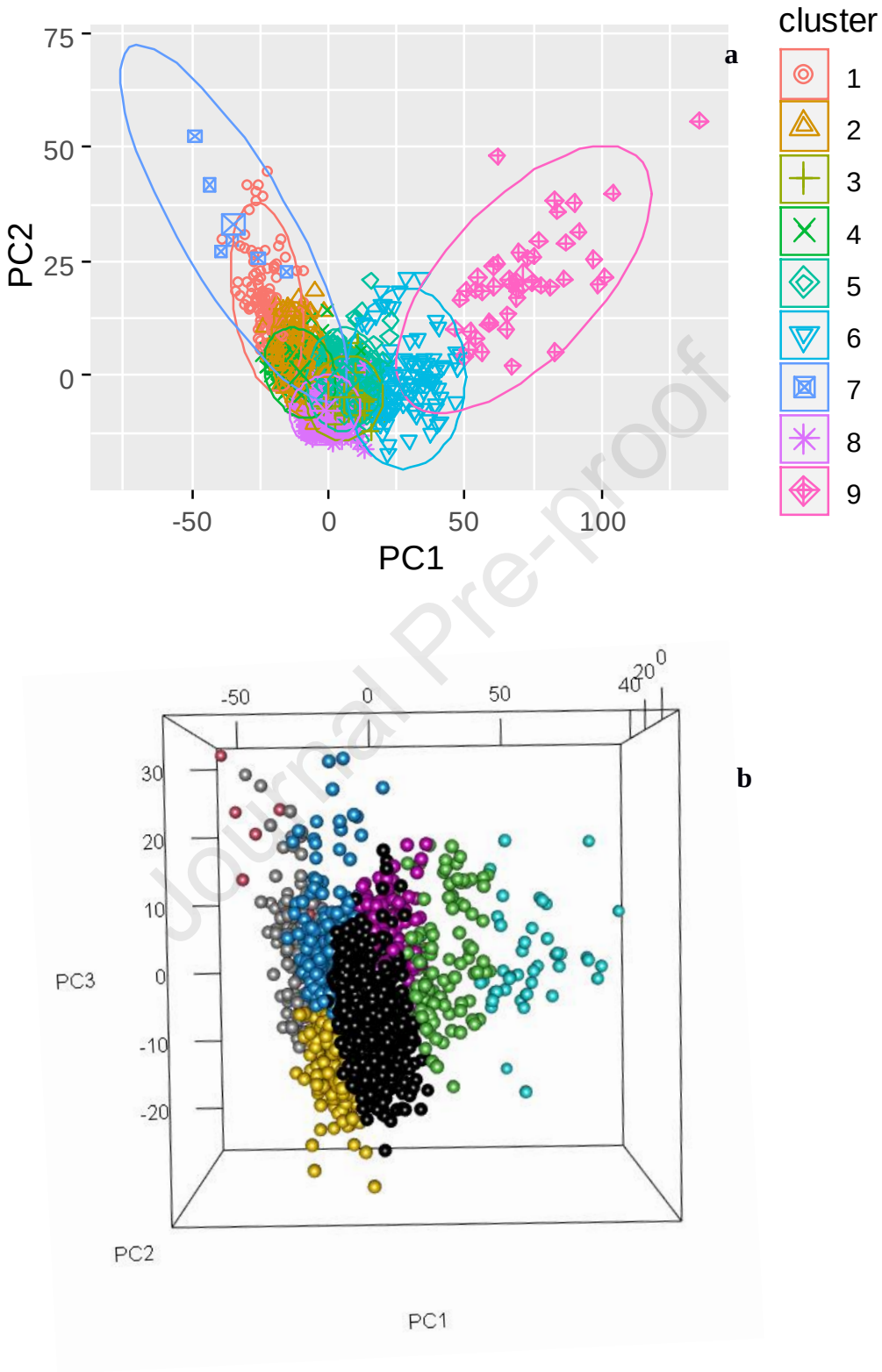


Figure 6 (a). 2D **(b)** 3D plot of the clusters of approved drugs obtained from the *k*-means clustering.

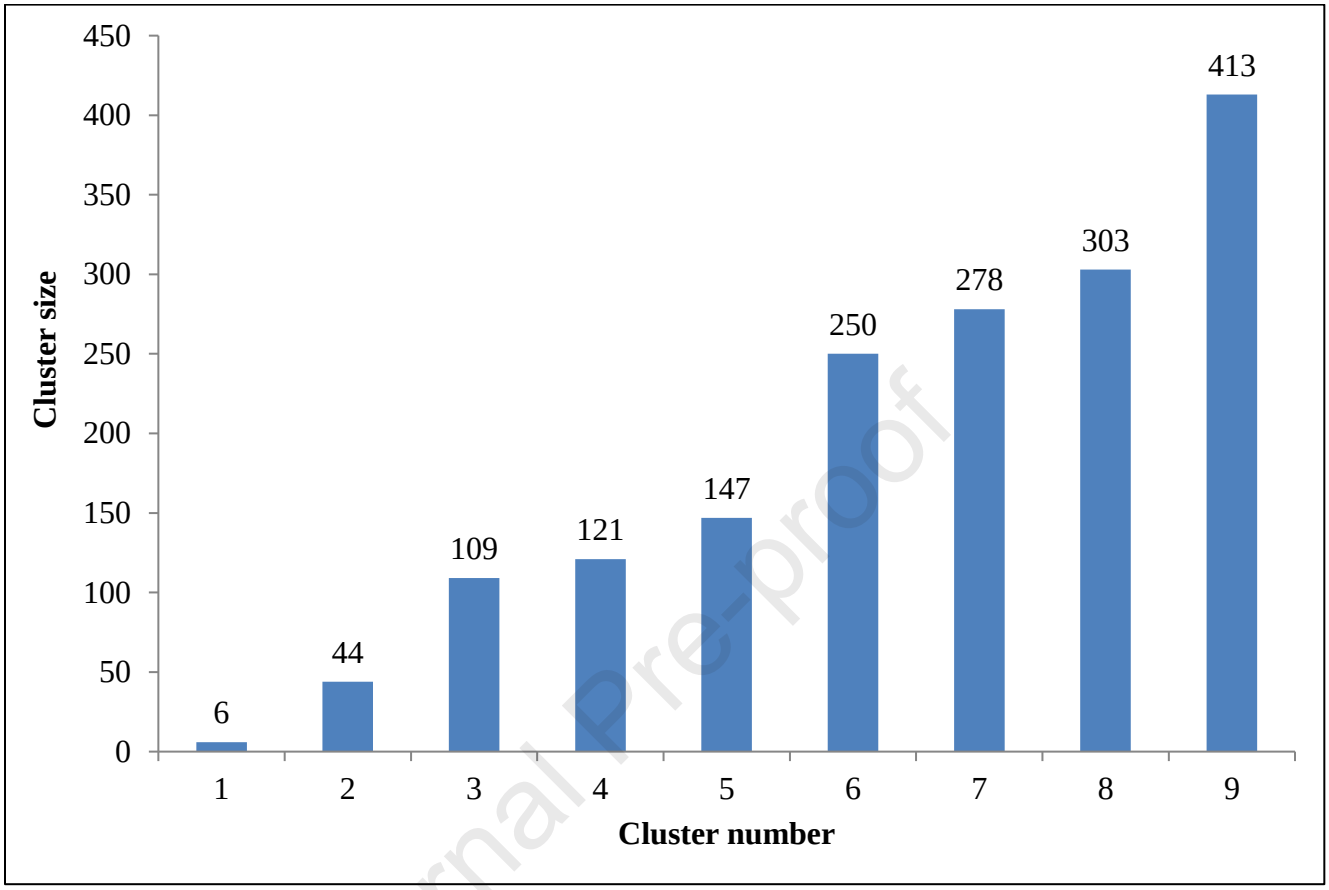


Figure 7. Distribution of approved drugs across nine clusters obtained from *k*-means clustering algorithm.

Journal Pre-proof

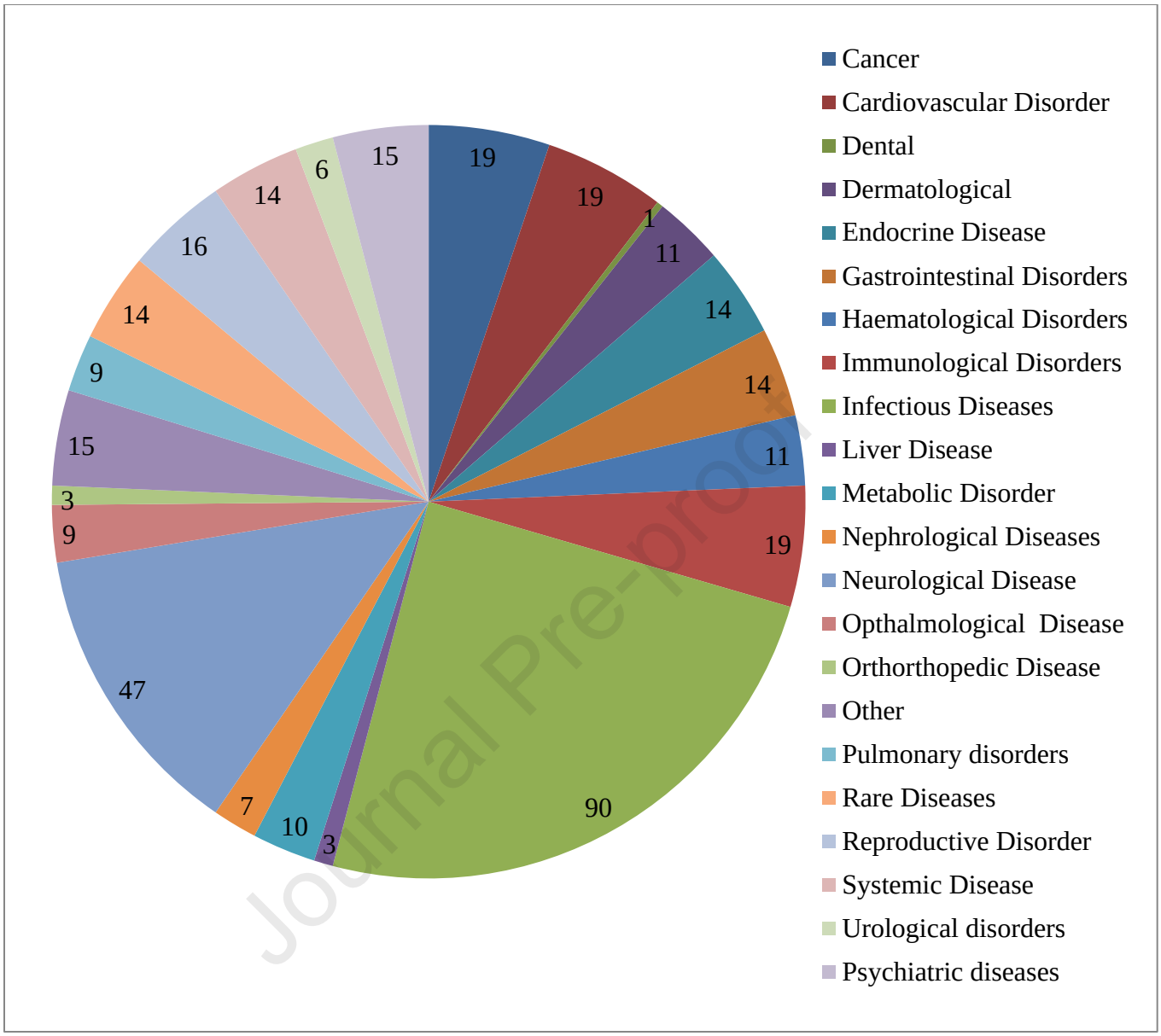
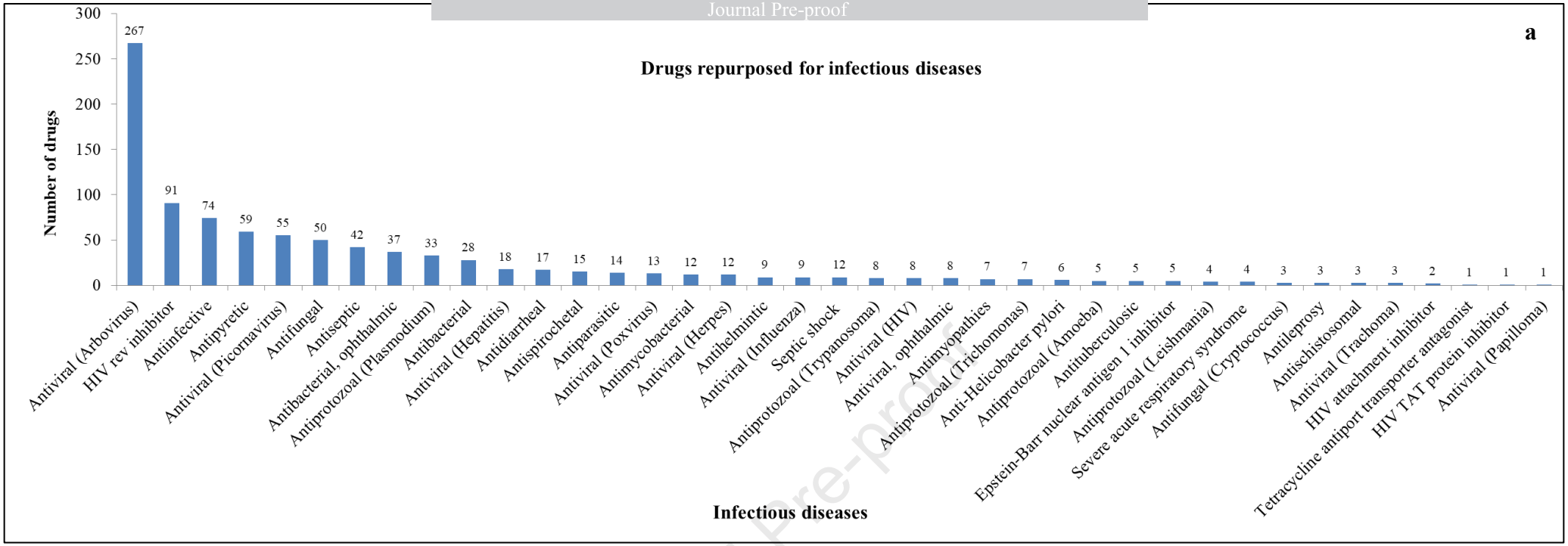


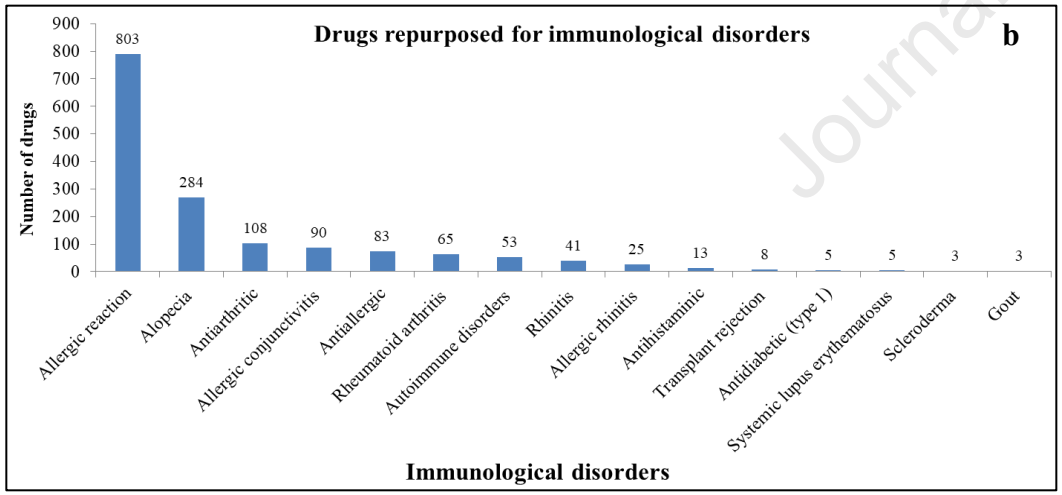
Figure 8. 366 diseases have been distributed into 22 major categories.

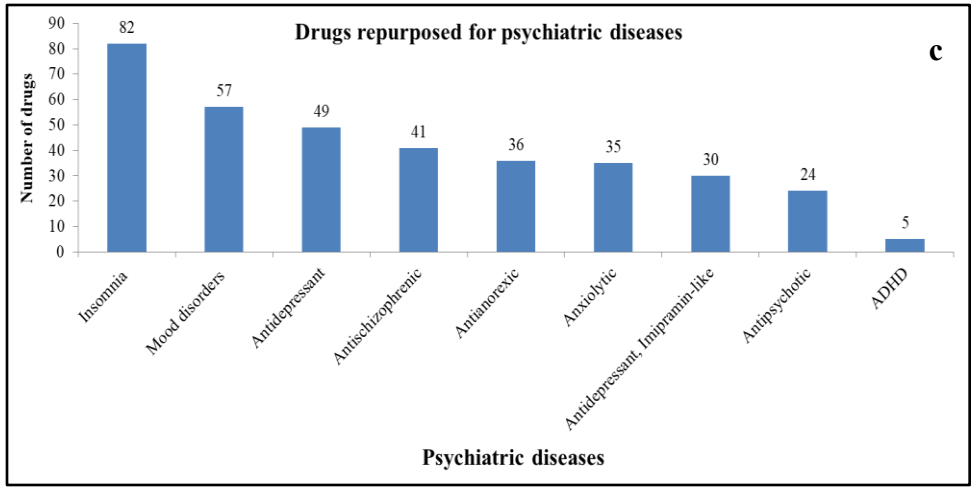
Journal Pre-proof

Journal Pre-proof



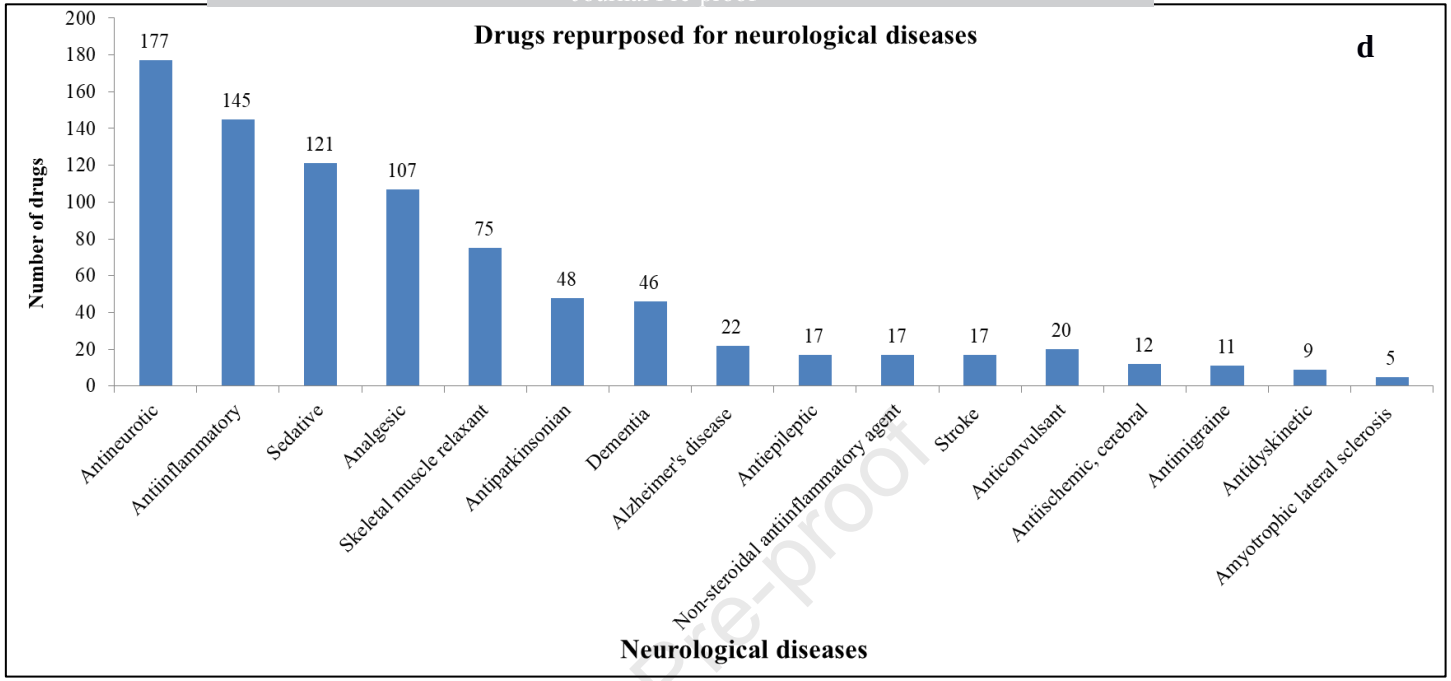
1

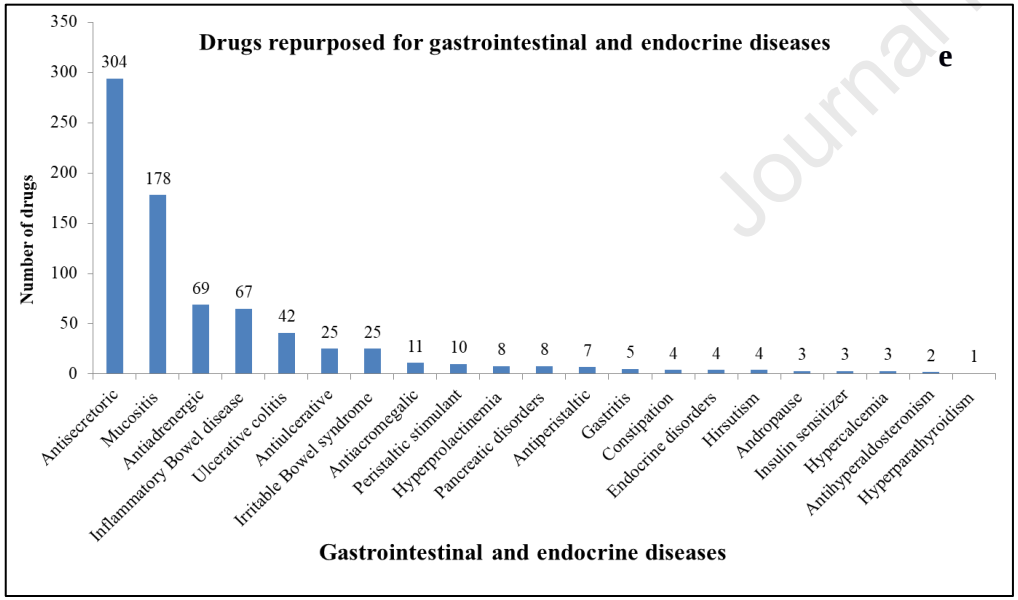


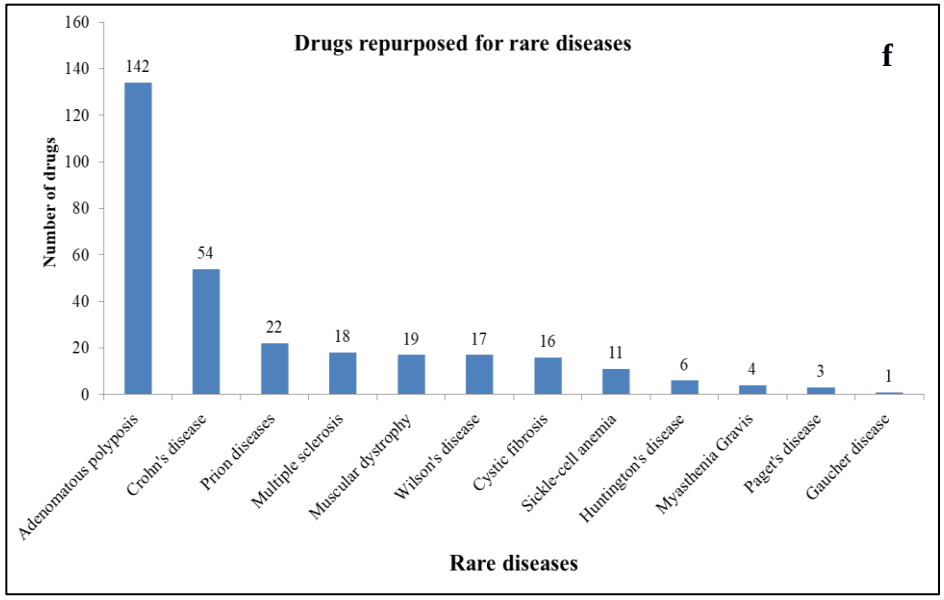


Journal Pre-proof

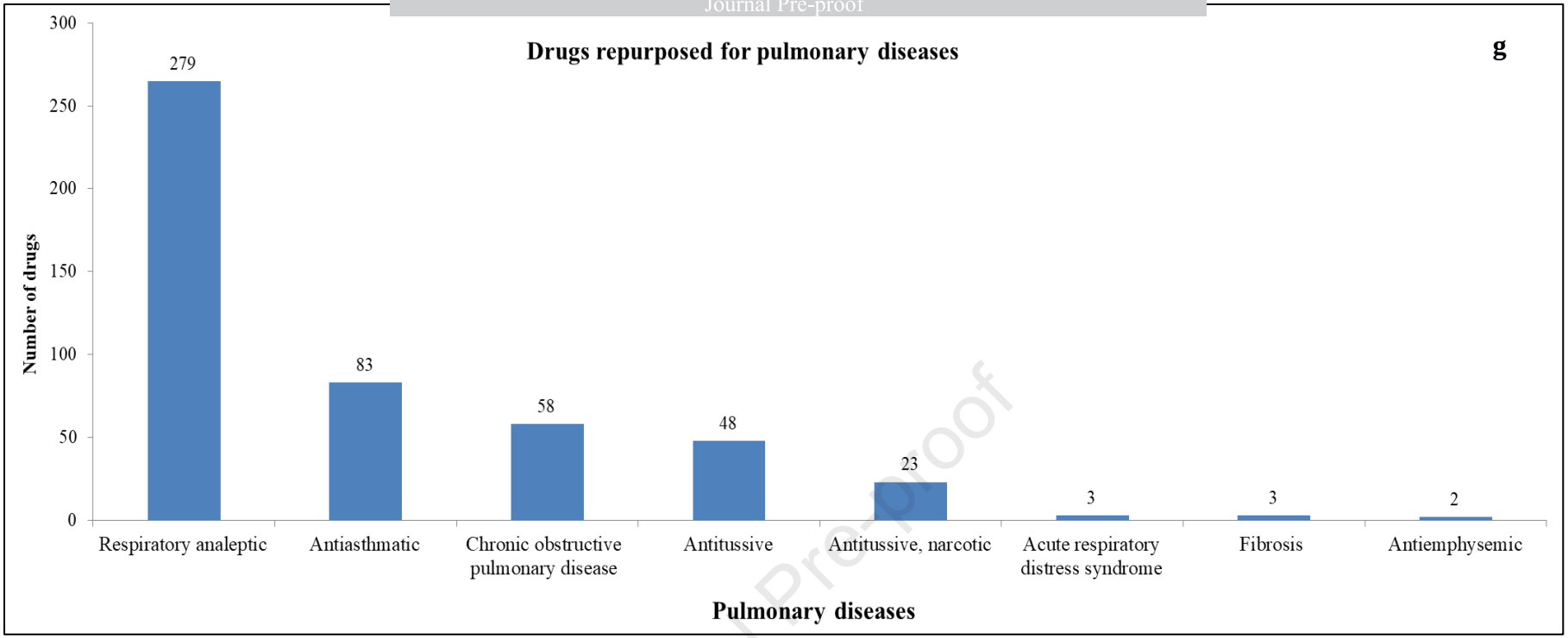
1

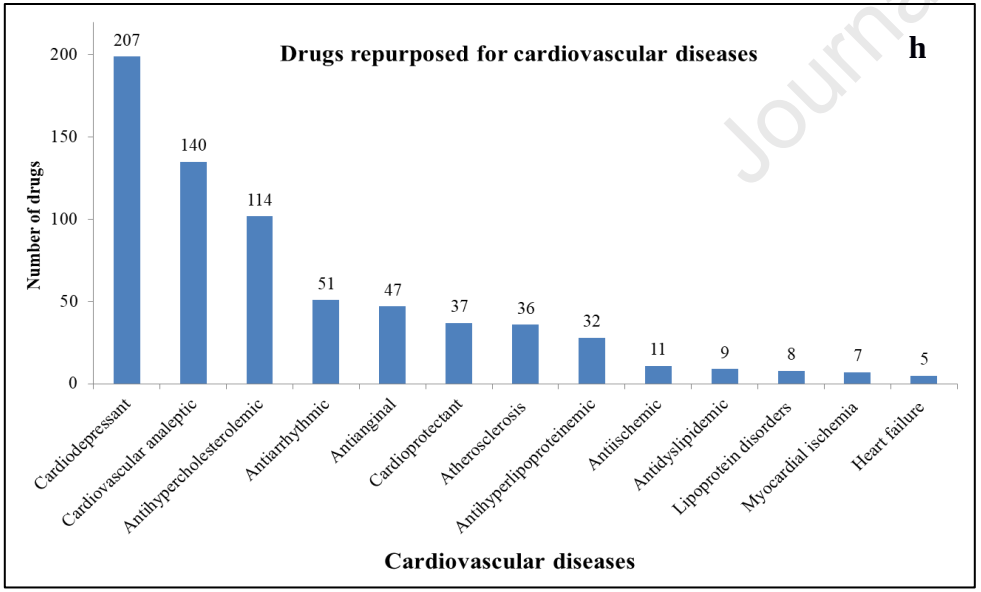


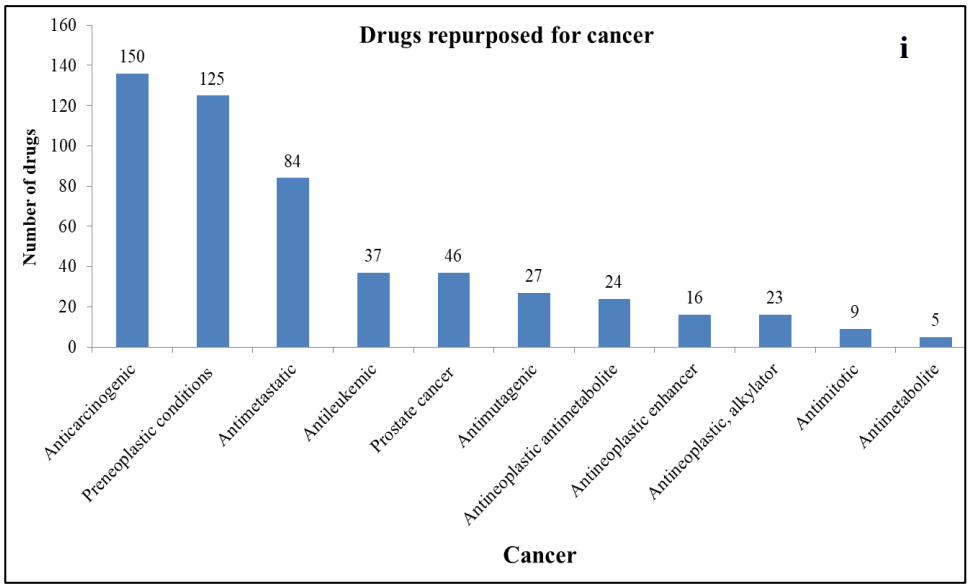




Journal Pre-proof

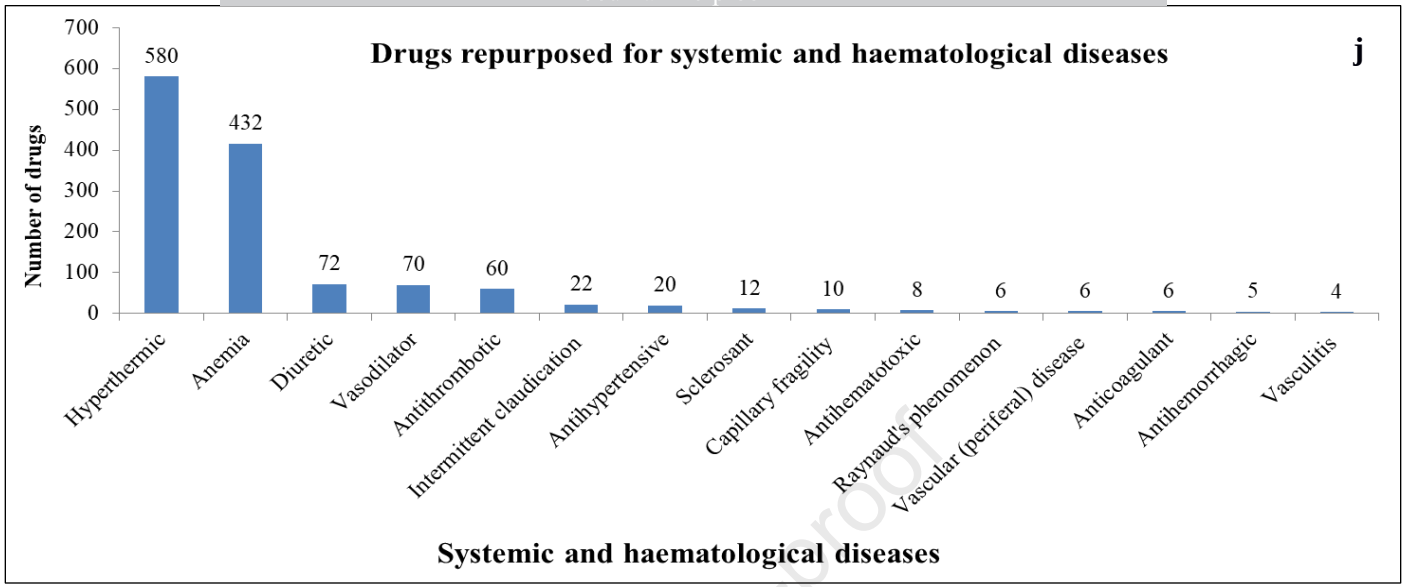


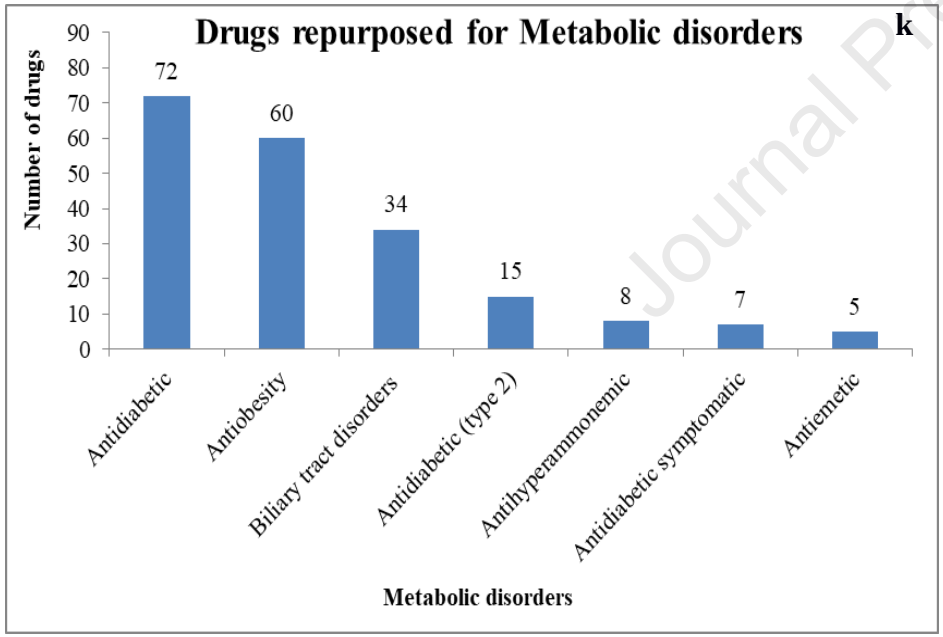




1

Journal Pre-proof





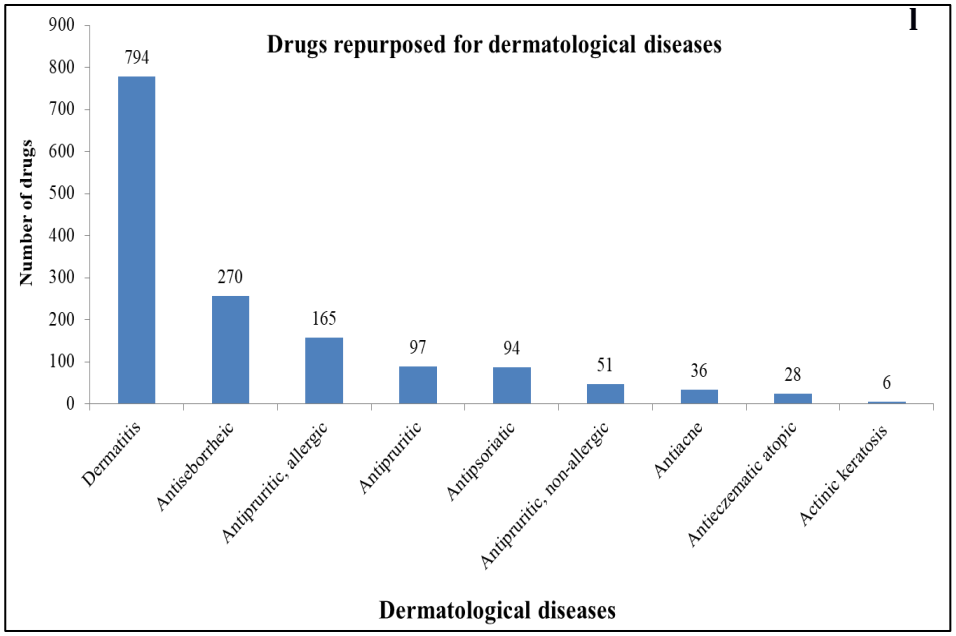


Figure 9 (a-l). Plots showing the distribution of repurposable drugs for 12 selected disease categories.

Journal Pre-proof

Molecular Descriptor Analysis of Approved Drugs Using Unsupervised Learning for Drug Repurposing

Highlights

- A dataset of 1671 approved drugs is analyzed using a combined data-driven approach and Structure-Activity Relationships (SAR) predictions for drug repurposing.
- we identified 66 drugs from the nine clusters which are structurally similar but have different therapeutic uses and can therefore be repurposed for one or more native indications of other drugs of the same cluster.
- we identified 1423 drugs that can be repurposed for 366 new indications against several diseases.

Journal Pre-proof

Journal Pre-proof

Conflict of interest:

The authors declare no conflicts of interest.

Journal Pre-proof