A Novel Ensemble Based Recommendation Approach using Network Based Analysis for Identification of Effective Drugs for Tuberculosis

Rishin Haldar (rishinh@gmail.com)
Vellore Institute of Technology: VIT University
Swathi Jamjala Narayanan
Vellore Institute of Technology: VIT University

Research Article

Keywords: Drug resistant Tuberculosis, Mtb, Drug discovery, Ensemble ranking, Pharmacokinetic Properties, Network based recommendation

Posted Date: July 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-680480/v1

Version of Record: A version of this preprint was published at Mathematical Biosciences and Engineering on January 1st, 2021. See the published version at https://doi.org/10.3934/mbe.2022040.

A Novel Ensemble Based Recommendation Approach using Network Based Analysis for Identification of Effective Drugs for Tuberculosis

Rishin Haldar[1,*], Swathi Jamjala Narayanan[2*]

[1]rishinh@gmail.com, [2]jnswathi@vit.ac.in,
School of Computer Science and Engineering,
Vellore Institute of Engineering, Vellore - 632014, Tamil Nadu, India

[*]Corresponding Author

Abstract

Tuberculosis (TB) is a fatal infectious disease which affected millions of people worldwide for many decades and now with mutating drug resistant strains, it poses bigger challenges in treatment of the patients. Computational techniques might play a crucial role in rapidly developing new or modified anti-tuberculosis drugs which can tackle these mutating strains of TB. This research work applied a computational approach to generate a unique recommendation list of possible TB drugs as an alternate to a popular drug, EMB, by first securing an initial list of drugs from a popular online database, PubChem, and thereafter applying an ensemble of ranking mechanisms. As a novelty, both the pharmacokinetic properties and some network based attributes of the chemical structure of the drugs are considered for generating separate recommendation lists. The work also provides customized modifications on a popular and traditional ensemble ranking technique to cater to the specific dataset and requirements. The final recommendation list provides established chemical structures along with their ranks, which could be used as alternatives to EMB. It is believed that the incorporation of both pharmacokinetic and network based properties in the ensemble ranking process added to the effectiveness and relevance of the final recommendation.

Keywords:
Drug resistant Tuberculosis, Mtb, Drug discovery, Ensemble ranking, Pharmacokinetic Properties, Network based recommendation

1. Introduction

Tuberculosis (TB), caused by the bacteria Mycobacterium tuberculosis (Mtb), is a major global health hazard, where the human mortality rate goes above 1 million per year. Surveys have revealed the existence of Mtb for more than four decades [1], and more importantly, it has developed various degrees of resistance to multiple anti tuberculosis drugs along this journey. Studies have also shown that [2]

the pathways that led to these drug resistant Mtb strains were very complex and indicated that the bacteria acquired a step by step process, starting from a low degree mutation to a near complete drug resistance. The role of bacterial factors like persistence, compensatory evolution, fitness, hyper mutation on the evolution of drug resistance were also explored [3].

There are various resources which documented the drug resistant properties of Mtb by various types of databases [4]. For example, the TBDRMD database [5] reported polymorphisms at different codon regions of certain genes of H37Rv, a specific strain of Mtb, in response to specific anti tuberculosis drugs. The VFDB database [6] provided virulence factors and structural features and functions of various bacteria including the different strains of Mtb. One important thing which came out from these studies [4] was the urgent need to rapidly discover new drugs which could tackle the mutating drug resistant strains of Mtb.

PubChem, an open chemistry database [7], provided an extensive resource on chemical structures, chemical and physical properties of a chemical compound (drug) as well as a comprehensive list of compounds (drugs) which were similar in nature. It was encouraging to know that a large number of growth inhibitors for Mtb were submitted [8] on PubChem after an elaborate process of high throughput screening and further evaluation based on potency, reproducibility and cytotoxicity. Thus, for this paper, PubChem was chosen to be the resource which could give an initial list of drugs which had some kind of similarity with an existing anti tuberculosis drug.

The basic functional effectiveness of any drug depends heavily on the molecular properties that are crucial to its pharmacokinetics, including absorption, distribution, metabolism and excretion (ADME). The well-established Lipinski's rule of five [9] defined some boundaries on the chemical properties of the drugs which, when satisfied, made the drugs more functionally acceptable for human intake. This important aspect was emphasized in our proposed methodology, both during data pre-processing as well as during the actual recommendation process.

There were many instances found in the literature where machine learning techniques were used to suggest new drugs, using available datasets related to drug resistant Mtb. Reviews [4] have pointed out cheminformatcs tools that were used on several datasets related to drug resistant Mtb. For example, a cheminformatics data fusion approach [10] was followed by validation on datasets having cytotoxicity

and tuberculosis data by Support Vector Machine and Bayesian models. Subsequently, Bayesian machine learning models were applied [11] on large datasets for rapid screening and prioritizing of compounds for anti-tubercular activities before in vitro testing. As both high throughput screening techniques and chemical datasets expanded [12], the needs to integrate and link the different datasets as well as consolidate the techniques by secure mechanisms were proposed. It has been reported that [13], popular machine learning techniques in cheminformatics can be useful in making the drug discovery process more efficient. This was due to the faster selection of filtered compounds for testing, when compared to the traditional in vitro screening for toxicity of compounds. This inference was also validated [14] when structure-activity relationships of a couple of Mtb enzymes were analysed both by in vitro screening as well as Naïve Bayesian model with smoothening. In another review [15], the relevance of the advancement of tools, including machine learning techniques, were highlighted with regard to development of anti-tuberculosis drugs. It also mentioned how recent advances in the field of genomics, high throughput screening have played a significant role in coming up with a variety of drugs at a more efficient rate.

Literature also revealed interesting approaches, other than throughput screening and cheminformatics tools, which were applied for drug discovery against drug resistant Mtb. Lipophilicity [16], along with structural peculiarities and energy depletion, in some cases, were reported to be more relevant than Lipinsky's rule of five, in the specific context of drug resistant TB. Reports [17] have summarised newer anti-tuberculosis drugs in different stages of clinical trials, along with their various methods of action. In another study [18[, the efficacy of Ethambutol (EMB), a popular anti-tuberculosis drug, was evaluated on how the drug affected both the cell wall integrity and the metabolism of the Mtb.

The review [15], mentioned before, even though acknowledged the effectiveness of compounds like EMB and INH as part of a multi-drug regimen against TB, it also underlined the critical need to come up with new anti-tuberculosis drugs, and suggested using repurposed drugs by exploring properties like genomics and crystallography. Reports [19] have also suggested that existing drugs meant for Parkinson's disease could be repositioned to treat drug resistant TB, by modelling protein-ligand interaction networks from looking at similarities in ligand binding sites and protein ligand docking properties. In the context of drug repurposing in general, reviews [20] classified them through two

axes, drug based or disease based. The review also highlighted that computational methods of repurposing benefitted both these axes. Studies [21] have also pointed out three actual repurposed drugs which were extensively used to treat drug resistant Mtb, when traditional combinations of anti-tuberculosis drugs were not effective. One study [22] looked at repurposing in general, by computational workflows, linking data availability with current trends in technology along with the associated algorithms. On similar lines, drug repositioning [23] was highlighted for associations between drug to target and target to disease, which consequently helped in predicting drug to disease mappings. In an experimental study [24], Primaquine derivatives were shown to be effective as both anti-malarial and anti-tuberculosis agents. In one more recent study [25], a traditional anti-tuberculosis drug, AMC, and a repurposed drug, Diosmin, were used together to treat TB.

There were some studies which looked at drug discovery, in general, from a graph oriented, network based approach. PROMISCOUS [26] was a database that had an extensive list of drugs annotated with protein-protein interactions (PPI) and drug-protein interactions. This study highlighted that structural similarities among drugs, connected with PPI, has the potential in the field of multi-pharmacology, and drug repositioning. Network analysis approaches [27] were studied in the context of drug discovery, and similarities were suggested in both social networks and biological networks. Incomplete network motifs of bi-cliques [28] were utilised in drug-target-disease networks to predict new drugs for one or more diseases. In one study, molecular interaction networks [29] were explored for lead identification in drug discovery pipeline. It also pointed out how this can help in exploring the emergence of drug resistance and drug repositioning. Reports [30] had also suggested the need to connect genome-based biological networks with anti-tuberculosis drug discovery to come up with potentially rational drugs. Surveys [31] had shown that network based computational approaches focussed on molecular interactions can address both drug repositioning, as well as drug combination. In case of finding drugs for Mtb, both of these (repositioning and combination) were very crucial for effective treatment of TB. A proteomic structural approach [32] was adopted for creation of clusters of pocket-similarity networks, or pocketomes, which consequently helped in finding sets of binding sites within Mtb. In a recent study [33], a drug-disease proximity measure was proposed by looking into the network neighbourhoods of

both disease genes and drugs. This study revealed that effectiveness of most of the drugs was limited to small sub networks of disease genes.

In summary, the entire literature study repeatedly highlighted the importance of coming up with discovery of relevant and effective anti-tuberculosis drugs in a rapid manner. Majority of these studies acknowledged the importance of the pharmacokinetic / ADME properties while discovering the new drugs, while quite a few of them looked at the problem from the structural network point of view. With these issues in context, this work is aimed at providing a recommendation list of possible anti-tuberculosis drugs which were similar to a popular and existing tuberculosis drug. The proposed method incorporated an ensemble approach, where both pharmacokinetic properties as well as network properties of the chemical structures of the recommended drugs were given special attention. For this ensemble recommendation approach, literature showed that Borda Count [34] was a very well established method and was used both in social science, as well as in machine learning domains. Studies [35] also showed that Borda Count can be effectively tailored for specific scenarios. In this context, the uniqueness of this effort was in the customisation of this popular ensemble ranking method, Borda Count, as well as utilising the network properties of the recommended drugs for the second time to fine tune the final recommendation list, and all these specific customisations were made to address the drug resistant property of the Mtb.

This section provided a basic introduction to the needs and a summary of techniques practised in discovering drugs against drug resistant Mtb. Section 2 details the methodology of this study along with our proposed ensemble recommendation system. Section 3 provides the experimental details for a popular anti-tuberculosis drug, and documents the results at each step of the process. Section 4 discusses and analyses the significance of our results, and finally, Section 5 highlights the conclusion and future possibilities.

2. Methodology

The process of consolidating the recommendation list of proposed TB drugs from a given effective TB drug is shown by the block diagram in Figure 1. The process involved data pre-processing at the beginning, followed by the generation of three different recommendation lists (Uniform weighted,

Pharmacokinetic weighted and Network weighted). These were then fed into an ensemble recommendation system resulting in a consolidated recommendation. The final step involved refinement of the consolidated ranking by utilising the Network weighted list of values.

[Suggested insertion of Figure 1]

2.1 Data Pre-Processing

The online PubChem database provided a list of drugs having similar characteristics for a TB drug given as a query. The dataset consisted of the list of drugs along with several attributes for each of these drugs. Since the objective of the research work dealt with providing recommendations by utilising different computational measures, a quantifiable approach was the basic requirement. In this context, data pre-processing was done on two fronts. The first filtering was done on numeric/non-numeric nature of the attributes, and the subsequent filtering was done on the pharmacokinetic values of the attributes.

2.1.1 Numeric attributes

From the initial dataset, at first, all non-numeric attributes were removed. This paved the way for performing numerical computations.

2.1.2 Pharmacokinetic filtering

The well-established pharmacokinetic properties of chemical drugs, provided by Lipinski's rule of five, Ghose Filter and Verber's rule, were applied to the relevant attribute values and those drugs whose attributes did not satisfy any of these rules, were removed. The combined rules are listed in Table 1.

[Suggested insertion of Table 1]

2.2 Ensemble of Ranked Lists

Three different ranked lists were generated from the processed dataset. The objective was to explore multiple ways of finding *k nearest* neighbours / drugs for the given query drug, where the entries in the ranked recommendation lists were in ascending order of dissimilarity. This meant that the topmost drug on a particular list would be the closest to the queried drug, as per the chosen measure of evaluation, and consequently the bottommost (k th) drug on the list would be the farthest. Three separate ranked recommendation lists were generated using three different evaluation measures.

2.2.1 Uniform Weighted Ranking

The first, and completely unbiased, ranking involved finding out the individual distances between each of the drugs in the processed dataset to the queried drug, and consequently storing the k *nearest* drugs. Mathematically, this process was similar to finding the Euclidean distance between vectors, and storing the k closest vectors, in sorted order, with respect to a single vector, represented by the queried drug. Each drug was represented by a vector of m numeric attributes. The formula for finding out the Euclidean dissimilarity between each of the suggested drugs and the queried drug is given in Equation 1.

$$\text{UED}(qd, rd_i)=\sqrt[2]{(qd_1 - rd_{i1})^2 + (qd_2 - rd_{i2})^2 \ldots + (qd_9 - rd_{i9})^2} \quad (1)$$

UED(qd, rd_i) stands for the uniform Euclidean distance between the vectors of the queried drug, qd,, and rd_i , the i^{th} suggested drug. Each of these vectors had nine numeric entries and qd_1 to qd_9 represents the numeric entries of the queried drug and subsequently rd_{i1} to rd_{i9} represents the numeric entries of the rd_i drug.

The key characteristic of this evaluation measure was that each of the attributes was given equal weightage when *distance / dissimilarity* were calculated for each of the suggested drugs from the queried drug.

2.2.2 Pharmacokinetic weighted Ranking

The second ranking used the same processed dataset to generate k *nearest* drugs and used the same Euclidean distance measure to find out the dissimilarity between each of the suggested drugs to the queried drugs. However, this method gave more weightage to a subset of the attributes while calculating the overall dissimilarity. These more important attributes (four in number) were the ones mentioned in Lipinsky's rule of five. Equation 2 shows the formula for calculating the dissimilarity, by incorporating the added weightage given to these more important attributes.

$$\begin{aligned}&\text{WED}(qd, rd_i)\\ &= Normalize\Big(\sqrt[2]{\left(\tfrac{2}{9}\left(qd_{p1} - rd_{pi1}\right)\right)^2 + \left(\tfrac{2}{9}(qd_{p2} - rd_{pi2})\right)^2 \ldots + \left(\tfrac{2}{9}\left(qd_{p4} - rd_{pi4}\right)\right)^2} \; +\\ &\qquad \sqrt[2]{\left(\tfrac{1}{9}(qd_1 - rd_{i1})\right)^2 + \left(\tfrac{1}{9}(qd_2 - rd_{i2})\right)^2 \ldots + \left(\tfrac{1}{9}(qd_5 - rd_{i5})\right)^2}\Big) \quad (2)\end{aligned}$$

WED(qd, rd_i) stands for the weighted Euclidean distance between the vectors of the queried drug, qd,, and rd_i , the i[th] suggested drug. Each of these vectors had nine numeric entries, out of which four of them (qd_{p1} , rd_{pi1} to qd_{p4}, rd_{pi}) were given double the weightage, compared to the remaining five numeric attributes represented by qd_1, $rd_{i1}$to qd_5, rd_{i5}. The total sum was then normalized for uniformity.

This weighted ranking approach gave more importance to the findings reported by researchers who had studied the pharmacokinetic properties of chemical drugs

2.2.3 Network Weighted Ranking

This paper introduces a new set of numeric attributes, which represents the network based characteristics of the chemical drugs. These network/graph based properties not only gave more insights into the chemical structures, by means of how the atoms were situated with respect to each other and how they were connected to each other, they also indicated the importance of an atom or a connection between two atoms to act as a bridge to connect to the remaining atoms of the network structure. This information, we believe, can be related with molecular bonding properties, which consequently determines the drug's ability to handle harmful bacterial antigens. The chosen network based attributes were i) average degree, ii) average closeness, iii) average node betweenness and iv) average edge betweenness. Equation 3 to Equation 6 shows the formula to calculate these values for a specific chemical drug structure.

$$\text{average\_degree} = (2 * E)/N \quad (3)$$

$$\text{average\_closeness} = \left(\sum_{i=1}^{N} \frac{1}{\sum_{y=1, y\neq x_i}^{N} d(y, x_i)}\right)/N \quad (4)$$

$$\text{average node betweenness} = \left(\sum_{s\neq v\neq t} \frac{\sigma_{st}^{(v)}}{\sigma_{st}}\right)/N \quad (5)$$

$$\text{average edge betweenness} = \left(\sum_{s\neq e\neq t} \frac{\lambda_{st}^{(e)}}{\lambda_{st}}\right)/N \quad (6)$$

E stands for the total number of edges, and N stands for the total number of nodes in a network. In the case of average closeness, $d(y, x_i)$ signifies the shortest distance between *nodes* y and x_i. For average node betweenness, σ_{st} represents the total number of shortest paths between the *nodes* s *and* t, while $\sigma_{st}^{(v)}$ shows the ones among those which pass through the *node* y. Similarly, for average edge

betweenness, λ_{st} generates the total number of shortest paths between the *nodes* s *and* t, while $\lambda_{st}^{(e)}$ gives the ones which pass through the *edge* e .

This network based ranked recommendation list stored *k nearest neighbours / drugs* with respect to the queried drug, where the input was only the ids of the suggested drugs, and the four attributes (i to iv). All these four numeric values were calculated from the graphical representations of the chemical drugs. The *k nearest neighbours /drugs*, in sorted order, with respect to the queried drug, with this customized dataset, were generated in the same manner as 3.2.1 using Equation 1 to calculate the dissimilarity.

2.3 Consolidated Ranking

The three different *k*-ranked recommendation lists were fed as input to the consolidated ranking system to generate a combined and consistent *k*-ranked recommendation list of drugs for a queried drug. The popular Borda ranking technique was initially chosen for this task, not only for its simple and effective way of ranking multiple ranked lists, but also for its unbiased way of giving weightage to a candidate even if it did not feature in the top *k* ranks in any of the ranked lists.

This paper proposes a customized version of Borda ranking to cater to the specific properties of the dataset. The ranking mechanism is explained in Table 2 as a demonstration. In the traditional Borda ranking system with three (A, B, C) evaluation measures, if C1, C2 and C3 were the chosen candidates, then the cells where C4 and C5 are placed (grey coloured cells), in the 3rd ranked position of evaluation measure B and C respectively, would be kept blank. This would have resulted in each of C1, C2 and C3 securing 6 points. However, the proposed modification looked at the absolute difference of ranking when lower ranked candidates were considered. In the case of Table 2, rank 4 and 5 (which were both lower than rank 3) resulted in Borda values of 0 and -1 respectively. Even though the proposed modification looked at lower ranks, the final result still looked at the top 3 ranks. The rankings generated by traditional Borda system and the rankings generated by our proposed modifications are highlighted in Table 2.

[Suggested insertion of Table 2]

2.4 Refinement of Recommendation

The consolidated ranked recommendation list provided *k nearest neighbours / drugs* with respect to the queried drugs by incorporating three kinds of evaluation measures. This paper specifically looked at one Tuberculosis (TB) drug in the query and consequently generated the *k* nearest recommended drugs. The queried TB drug, although being quite effective, was not perfect, as limited polymorphisms were reported by certain genes of a specific Mtb strain. This indicated that the effectiveness of the drug might reduce drastically in the future once the Mtb strain gradually becomes drug resistant to that specific drug. Therefore, an effective way to recommend other drugs could be to look for similar drugs, i.e., drugs that are high up in the ranked recommendation list, but not those drugs which are extremely similar, or identical, in terms of the attribute values. In this specific context, this paper proposes a unique refinement on the consolidated recommendation, by revisiting the network based recommendation list, and removing those topmost recommended drugs from the consolidated ranked recommendation list which had identical network structure based attribute values when compared with the queried drug. This meant that drugs which had almost identical chemical structure might not be an effective recommended drug in the long run, while top recommended drugs which are similar, but did not have identical network structures were viable suggestions.

The consolidated ranked *k nearest* neighbour recommendation list was thus refined by removing those drug entries which had identical attribute values with respect to the queried drug in the network based ranked recommendation list.

2.5 Algorithm

The proposed methodology is detailed by the algorithm given below.

Input: A dataset "**ID**" with "**n**" number of tuples and "**c**" number of columns. Each tuple represents a drug along with its properties. The dataset was provided by PubChem in response to a queried drug

Output: A table "**FR**" having "**fn**" number of tuples and 2 columns, the first column being the unique numeric (cid) code of the drug and the second column representing the consolidated rank of the drug, 1 being the highest rank

Begin

Step 1: **ID'[n x num] ← ID[n x c]** // **ID[n x c]** is reduced to **ID'[n x num]**, where **num** represents the numeric attributes of the drug. For our experiment, **num**=10, consisting of the cid and nine numeric attributes.

Step 2: D[n' x num] ← ID'[n x num] // **ID'[n x num]** is reduced to **D[n' x num],** after tuples were removed because these tuples (drugs) did not satisfy the pharmacokinetic properties, as specified in **Table 1**.

Step 3a: A[tn,3] ← D[n' x num] //**A[tn,3]** is acquired from **D[n' x num],** by calculating uniform weighted dissimilarity using **Equation(1)**. The first column in **A** corresponds to the cid of the recommended drug, the second column gave the uniform dissimilarity score and the third column gave the dissimilarity rank. **tn** represented the top number of drugs that was asked for.

Step 3b: B[tn,3] ← D[n' x num] //**B[tn,3]** is acquired from **D[n' x num],** by calculating pharmacokinetic weighted dissimilarity using **Equation(2)**. The first column in **B** corresponds to the cid of the recommended drug, the second column gave the weighted dissimilarity score and the third column gave the dissimilarity rank. **tn** represented the top number of drugs asked for.

Step 3c: C[tn,3] ← D[n' x 4] //**C[tn,3]** is acquired from **D[n' x 4],** by calculating four Network based dissimilarity features for each of the tuples. The four features were calculated using **Equation (3-6)**. The final dissimilarity of each of these four attributed vectors from the queried drug, 14052 was calculated using **Equation 1**. The first column in **C** corresponds to the cid of the recommended drug, the second column gave the Network based dissimilarity score and the third column gave the dissimilarity rank. **tn** represented the top number of drugs that was asked for.

Step 4: F[tn x 3]←Ensemble ranking (**A[tn,2],B[tn,2],C[tn,2])** by customised Borda //**F** corresponds to a table which holds the **i) cid**, **ii) customised Borda Value** and **iii)customised Borda rank** for **tn** number of tuples (drugs). The inputs were the datasets **A, B** and **C**, with **tn** entries and only two columns, the **i) cid** and the **ii) dissimilarity rank**.

Step 5: FR[fn x 2] ← F[tn x 3] // **FR** represents the final recommendation table with **fn** number of entries and 2 columns. The columns being **i) cid** and **ii) final recommendation rank**. The value **fn** (**<= tn**) corresponds to the number of drugs which are still there after filtering is done on the table **F** by entries in table **C**. If any entry in the **second** column of **C** was **zero**, then that corresponding **cid** tuple entry is removed from the **FR** table.

End

3. Experimental Results

The online PubChem database was used to gather the initial dataset for this effort. The dataset provided a list of chemical structures (and their details) that was similar to a queried chemical structure. The queried chemical structure was EMB (Ethambutol), an effective TB drug used for analysing drug resistant properties of H37Rv strain of Mycobacterium Tuberculosis, and this was gathered from VFDB and TBDRMD datasets. In response to the queried TB drug, EMB, PubChem generated a dataset having a dimension of 352 x 20, signifying 352 similar drugs/chemical structures, each having 20 attributes. The 20 attributes consisted of nine numeric attributes, namely i) molecular weight, ii) polar area, iii) complexity, iv) log p, v) heavy count, vi) hbond donor, vii) hbond acceptor, viii) rotatable bonds and ix) annotation hit count. There were 13 non-numeric attributes like the cid, molecular formula, name of

component, synonym of component, inchikey, meshheading etc. For our calculations, only the cid and the nine numeric attributes were kept and the remaining attributes were pruned as the first part of data pre-processing, which reduced the dataset to be 352 x 10. Thereafter, pharmacokinetic filtering was applied as the second part of data pre-processing, which involved with the values in six (i, ii, iv, vi, vii and vii) of the numeric attributes. The filtering was done as per rules described in Table 1, and the final pre-processed dataset, D, came out with a dimension of 236 x 10.

The cleaned dataset D was first utilised to generate the uniform weighted ranked recommendation list. For the experiments conducted, we chose the top 15 (k=15) chemical structures / drugs which were similar to the queried drug, EMB, having a cid of 14052. The uniform weighted ranked recommendation list is shown in Table 3. This was calculated using Equation 1, where k-nearest neighbours/drugs of 14052 were listed with the dissimilarity values and the corresponding ranks, the 1st rank being the one with the least dissimilarity.

[Suggested insertion of Table 3]

The dataset D was again used to generate the pharmacokinetic weighted ranked recommendation list, where the top k(=15) nearest neighbours/drugs of 14052 were listed (in Table 4) with the dissimilarity values and the corresponding ranks, the 1st rank being the one with the least dissimilarity. The dissimilarity was calculated using Equation 2, by a Euclidean distance measure, where more weightage was given to four of the attributes (mw, xlogp, hbond donor count, hbond acceptor count).

[Suggested insertion of Table 4]

The dataset D had a list of drugs (cids) which were similar to EMB (14052) as per PubChem query. PubChem also provided 2-dimensional chemical structures of these cids. This paper, as a novelty, looked into four network based properties (average degree, average closeness, average node betweenness and average edge betweenness) of each of these chemical structures and found out how each of them were dissimilar from the chemical structural properties of popular TB drug EMB. The dissimilarity was calculated by the Euclidean distance measure and consequently the network weighted ranked recommendation list was generated for the k (=15) nearest neighbours/drugs as shown in Table 5.

[Suggested insertion of Table 5]

After generating the three different ranked recommendation lists (Table 3, 4 and 5), a consolidated ranked list was prepared by using a modified Borda ranking technique. The unique modification to the traditional Borda technique is explained in Section 3.3. For our experiment, we chose the top 15 drugs/chemical structures for the consolidated list, thus rank 16, 17 and 18 had a value of 0, -1 and -2 respectively. The 'Consolidated Rank' column of Table 6 shows the result of this modified Borda ranking.

[Suggested insertion of Table 6]

The final list of suggested drugs with their corresponding ranks is listed in the 'Refined Recommendation" column of Table 6. According to our proposed recommendation mechanism, the chemical compound Myambutol, having cid 3279, was top ranked among the suggested chemical structures, which could be studied as an alternative to the queried drug, EMB, with cid 14052. The network based properties of both 14052 and 3279 are shown in Figure 2 and Figure 3 respectively. The circles represent atoms and the edges represent the connections between the atoms. The Nitrogen, Hydrogen and Oxygen atoms are represented as N, H and O respectively, followed by a serial number, while the other atoms are represented by a serial number only. The size of a circle is proportional to the value for that atom for a particular attribute, signifying that a bigger value of the specific attribute resulted in a bigger sized circle.

[Suggested insertion of Figure 2]

[Suggested insertion of Figure 3]

4. Discussion

It is evident from Table 6 that three chemical compounds, namely 162045, 470071 and 470072, were omitted from the final recommendation list ('Refined Recommendation' column). This was due to the fact that these three chemical compounds had no difference with EMB, and consequently held the 1st recommendation rank jointly, based on the chosen network based properties, as shown in Table 5. The motivation for their omission in the Refined Recommendation was to give importance to similar drugs/chemical structures, but not to those which had almost identical chemical structures. The reason was inferred from the fact that although EMB was an effective drug against the H37Rv strain of Mtb due to the very low number of polymorphisms, the complex drug resistance properties of Mtb [1][2]

would eventually make an existing drug, like EMB, to be ineffective in due course of time. Hence, finding an alternative drug with almost identical chemical structure as the current one would also be ineffective. However, a recommended drug which had minor changes in the chemical structure, when compared to an existing effective drug, could be effective in dealing with a mutated strain of Mtb. This drug recommendation, needless to say, should also adhere to the pharmacokinetic limitations or boundaries.

The literature revealed that several promising studies were conducted[27][29][31][33] which focussed on the chemical / drug from the structural network point of view only and consequently came up with new drugs or existing drugs which could be repurposed. On the other hand, the recommended drug had to satisfy basic pharmacokinetic properties for effective utilization by the human body. In this context, this study attempted to provide a balanced approach by giving due weightage to both these aspects by generating two additional recommendation lists, one with emphasis on network centric similarities and the other with emphasis on pharmacokinetic related similarities. Consequently, the customised ensemble ranking was aimed at providing an elegant way to come up with a consolidated ranking which took into account all these aspects.

The final list of suggested drugs (from the last column in Table 6) had 15 entries, even though Table 3, Table 4 and table 5 showed 18 entries. These 3 additional entries were utilised in our modified Borda ranking technique. These additional entries were also helpful because after the final refinement of recommendation, where three chemical structures were omitted, it was still able to suggest 15 chemical structures at the end.

5. Conclusion and Future Work

The effort was aimed at generating a list of possible TB drugs by using an ensemble of ranking mechanisms in which one of the ranking mechanisms included a novel way of looking at the chemical structures of the drugs and evaluating them based on network based attributes, and another ranking mechanism stressed on pharmacokinetic compatibility. The effort also proposed unique modifications on a popular ensemble ranking technique to suit our specific requirement. The results dealt with

providing established chemical structures along with their ranks, which could possibly be used as alternatives to EMB, a popular drug used for H37Rv strain of TB. It is believed that the incorporation of the pharmacokinetic properties as well as network based properties in the ensemble ranking mechanism would make the proposed recommendations logical, practical and effective.

The proposed methodology can be applied not only for TB drugs, but to any other drugs, provided they are recorded in the online PubChem database. From the pharmacokinetic point of view, more studies can be done to verify that the recommendation does not produce an unwanted inhibitory or toxic effect on the biological system. On the other hand, while evaluating network based properties of the chemical structures, four popular properties were looked at. One can also look at more network properties as well as graph isomorphism properties to find out the level of similarity/dissimilarity among chemical structures of drugs.

Authorship contribution statement:

Rishin Haldar: Conceptualization, Writing- original draft, Methodology, Software. **Swathi Jamjala Narayanan:** Supervision, Validation, reviewing & editing.

Funding statement:

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest:

The authors declare that they have no conflicts of interest to report regarding the research work.

Data availability statement:

Data sharing is not applicable to this article as no new data were created or analysed in this study.

Acknowledgment:

The authors take this opportunity to thank the management of VIT for providing the facilities and encouragement to carry out this work.

References

[1] Eldholm V, Monteserin J, Rieux A, Lopez B, Sobkowiak B, Ritacco V, Balloux F. Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. Nature communications. 2015 May 11;6(1):1-9.

[2] Fonseca JD, Knight GM, McHugh TD. The complex evolution of antibiotic resistance in Mycobacterium tuberculosis. International journal of infectious diseases. 2015 Mar 1;32:94-100.

[3] Müller B, Borrell S, Rose G, Gagneux S. The heterogeneous evolution of multidrug-resistant Mycobacterium tuberculosis. Trends in Genetics. 2013 Mar 1;29(3):160-9.

[4] Ekins S, Freundlich JS, Choi I, Sarker M, Talcott C. Computational databases, pathway and cheminformatics tools for tuberculosis drug discovery. Trends in microbiology. 2011 Feb 1;19(2):65-74.

[5] Sandgren A, Strong M, Muthukrishnan P, Weiner BK, Church GM, Murray MB. Tuberculosis drug resistance mutation database. PLoS Med. 2009 Feb 10;6(2):e1000002. .

[6] Chen L, Xiong Z, Sun L, Yang J, Jin Q. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic acids research. 2012 Jan 1;40(D1):D641-5.

[7] Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L. PubChem 2019 update: improved access to chemical data. Nucleic acids research. 2019 Jan 8;47(D1):D1102-9.

[8] Goldman RC. Target discovery for new antitubercular drugs using a large dataset of growth inhibitors from PubChem. Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders). 2020 Jun 1;20(3):352-66.

[9] Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews. 1997 Jan 15;23(1-3):3-25.

[10] Ekins S, Freundlich JS, Reynolds RC. Fusing dual-event data sets for Mycobacterium tuberculosis machine learning models and their evaluation. Journal of chemical information and modeling. 2013 Nov 25;53(11):3054-63.

[11] Ekins S, Casey AC, Roberts D, Parish T, Bunin BA. Bayesian models for screening and TB Mobile for target inference with Mycobacterium tuberculosis. Tuberculosis. 2014 Mar 1;94(2):162-9.

[12] Ekins S, Clark AM, Swamidass SJ, Litterman N, Williams AJ. Bigger data, collaborative tools and the future of predictive drug discovery. Journal of computer-aided molecular design. 2014 Oct;28(10):997-1008.

[13] Ekins S, Clark A, Perryman A, Freundlich J, Korotcov A, Tkachenko V. Accessible machine learning approaches for toxicology. Computational toxicology: risk assessment for chemicals. 2018 Feb 13:1-29.

[14] Djaout K, Singh V, Boum Y, Katawera V, Becker HF, Bush NG, Hearnshaw SJ, Pritchard JE, Bourbon P, Madrid PB, Maxwell A. Predictive modeling targets thymidylate synthase ThyX in Mycobacterium tuberculosis. Scientific reports. 2016 Jun 10;6(1):1-11.

[15] Chetty S, Ramesh M, Singh-Pillay A, Soliman ME. Recent advancements in the development of anti-tuberculosis drugs. Bioorganic & medicinal chemistry letters. 2017 Feb 1;27(3):370-86.

[16] Machado D, Girardini M, Viveiros M, Pieroni M. Challenging the drug-likeness dogma for new drug discovery in tuberculosis. Frontiers in microbiology. 2018 Jul 3;9:1367.

[17] AlMatar M, AlMandeal H, Var I, Kayar B, Köksal F. New drugs for the treatment of Mycobacterium tuberculosis infection. Biomedicine & Pharmacotherapy. 2017 Jul 1;91:546-58.

[18] Ghiraldi-Lopes LD, Campanerut-Sá PA, Evaristo GP, Meneguello JE, Fiorini A, Baldin VP, de Souza EM, de Lima Scodro RB, Siqueira VL, Cardoso RF. New insights on Ethambutol Targets in Mycobacterium tuberculosis. Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders). 2019 Mar 1;19(1):73-80.

[19] Kinnings SL, Liu N, Buchmeier N, Tonge PJ, Xie L, Bourne PE. Drug discovery using chemical systems biology: repositioning the safe medicine Comtan to treat multi-drug and extensively drug resistant tuberculosis. PLoS Comput Biol. 2009 Jul 3;5(7):e1000423.

[20] Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. Briefings in bioinformatics. 2011 Jul 1;12(4):303-11.

[21] Maitra A, Bates S, Kolvekar T, Devarajan PV, Guzman JD, Bhakta S. Repurposing—a ray of hope in tackling extensively drug resistance in tuberculosis. International Journal of Infectious Diseases. 2015 Mar 1;32:50-55.

[22] Vanhaelen Q, Mamoshina P, Aliper AM, Artemov A, Lezhnina K, Ozerov I, Labat I, Zhavoronkov A. Design of efficient computational workflows for in silico drug repurposing. Drug Discovery Today. 2017 Feb 1;22(2):210-22.

[23] March-Vila E, Pinzi L, Sturm N, Tinivella A, Engkvist O, Chen H, Rastelli G. On the integration of in silico drug design methods for drug repurposing. Frontiers in pharmacology. 2017 May 23;8:298.

[24] Pavić K, Perković I, Pospíšilová Š, Machado M, Fontinha D, Prudêncio M, Jampilek J, Coffey A, Endersen L, Rimac H, Zorc B. Primaquine hybrids as promising antimycobacterial and antimalarial agents. European journal of medicinal chemistry. 2018 Jan 1;143:769-79.

[25] Pushkaran AC, Vinod V, Vanuopadath M, Nair SS, Nair SV, Vasudevan AK, Biswas R, Mohan CG. Combination of repurposed drug diosmin with amoxicillin-clavulanic acid causes synergistic inhibition of mycobacterial growth. Scientific reports. 2019 May 1;9(1):1-4.

[26] Von Eichborn J, Murgueitio MS, Dunkel M, Koerner S, Bourne PE, Preissner R. PROMISCUOUS: a database for network-based drug-repositioning. Nucleic acids research. 2010 Nov 10;39(suppl_1):D1060-6.

[27] Hasan S, Bonde BK, Buchan NS, Hall MD. Network analysis has diverse roles in drug discovery. Drug discovery today. 2012 Aug 1;17(15-16):869-74.

[28] Daminelli S, Haupt VJ, Reimann M, Schroeder M. Drug repositioning through incomplete bi-cliques in an integrated drug–target–disease network. Integrative Biology. 2012 Jun 25;4(7):778-88.

[29] Chandra N, Padiadpu J. Network approaches to drug discovery. Expert opinion on drug discovery. 2013 Jan 1;8(1):7-20.

[30] Chung BK, Dick T, Lee DY. In silico analyses for the discovery of tuberculosis drug targets. Journal of Antimicrobial chemotherapy. 2013 Dec 1;68(12):2701-9.

[31] Wu Z, Wang Y, Chen L. Network-based drug repositioning. Molecular BioSystems. 2013;9(6):1268-81.

[32] Anand P, Chandra N. Characterizing the pocketome of Mycobacterium tuberculosis and application in rationalizing polypharmacological target selection. Scientific reports. 2014 Sep 15;4(1):1-7.

[33] Guney E, Menche J, Vidal M, Barábasi AL. Network-based in silico drug efficacy screening. Nature communications. 2016 Feb 1;7(1):1-3.

[34] Emerson P. The original Borda count and partial voting. Social Choice and Welfare. 2013 Feb;40(2):353-8.

[35] Fraenkel, J., & Grofman, B. (2014). The Borda Count and its real-world alternatives: Comparing scoring rules in Nauru and Slovenia. Australian Journal of Political Science, 49(2), 186-205.

Table 1: Filters applied on the attributes of the list of drugs generated from PubChem

Attribute Constraints	Filter Rule
hbond donor count <=5	Lipinski's Rule
hbond receptor count <=10	Lipinski's Rule
180<= Molecular weight <=480	Lipinski's Rule and Ghose Filter (combined)
-0.4 <= logP <5	Lipinski's Rule and Ghose Filter (combined)
20<=Total number of atoms<=70	Ghose Filter
Polar surface area <=140	Verber's Rule
Number of rotatable bonds <=10	Verber's Rule

Table 2: An illustrative example of our customized Borda ranking. C1 to C5 are the candidates and A, B, C are the three evaluation measures, and the objective is to get the top three candidates. The upper table shows the ranks of the 5 candidates for each of the evaluation measures. The bottom left table shows the ranks of C1, C2, C3 by the traditional Borda ranking, while the bottom right table shows the same by our customized Borda ranking

Evaluation Measures	Rank of Candidates				
	1	2	3	4	5
A	C2	C3	C1	C5	C4
B	C1	C2	C4	C3	C5
C	C3	C1	C5	C2	C4

Candidates	Borda Score	Borda Rank
C1	6	Equal
C2	6	
C3	6	

Candidates	Modified Borda Score	Modified Borda Rank
C1	6	1
C2	5	2
C3	5	2
C4	-1	
C5	0	

Note: The grey coloured cells represent locations where traditional Borda ranking would have kept blank, and would have given C3 and C2 1 point each, respectively.

Table 3: Recommendation list for the cid 14052, based on uniformly weighted properties, where all nine numeric attributes were given equal weightage in calculating the dissimilarity.

Pub Chem Sl No.	**cid**	mw	Polar area	complexity	xlogp	heavycnt	hbonddonor	hbondacc	rotbonds	annothitcnt	**Dissimilarity**	**Dis Rank**
	14052	204.31	64.5	109	-0.1	14	4	4	9	15		
1	**3279**	204.31	64.5	109	0	14	4	4	9	15	**0.1**	**1**
2	**162045**	204.31	64.5	123	0.4	14	4	4	9	0	**20.5243758**	**6**
3	**465436**	204.31	64.5	109	0	14	4	4	9	1	**14.0003571**	**4**
4	**469919**	188.31	35.5	105	0	13	2	3	8	1	**36.2768521**	**13**
5	**469924**	202.34	35.5	129	0	14	2	3	8	1	**38.0380191**	**14**
6	**469926**	218.34	64.5	120	0	15	4	4	10	1	**22.7123513**	**7**
7	**469930**	218.34	64.5	145	0	15	4	4	9	3	**40.4703706**	**15**
8	**469973**	219.32	81.8	147	0	15	4	5	9	1	**46.5467518**	**16**
9	**469988**	204.31	64.5	128	0	14	4	4	7	2	**23.1086564**	**8**
10	**469989**	232.36	64.5	149	0	16	4	4	9	1	**50.8607167**	**18**
11	**469990**	204.31	64.5	109	0	14	4	4	9	1	**14.0003571**	**4**
12	**469991**	232.36	64.5	145	0	16	4	4	9	3	**47.2314779**	**17**
13	**469996**	204.31	44.7	123	0	14	2	4	9	0	**28.5840865**	**9**
14	**470016**	192.3	45.3	96.5	0	13	4	4	7	0	**29.9856649**	**11**
15	**470071**	204.31	64.5	109	0	14	4	4	9	6	**9.00055554**	**2**
16	**470072**	204.31	64.5	109	0	14	4	4	9	6	**9.00055554**	**2**
17	**470075**	188.31	44.3	109	0	13	3	3	8	1	**29.3947274**	**10**
18	**470076**	188.31	44.3	121	0	13	3	3	7	1	**31.7970124**	**12**

Table 4: Recommendation list for the cid 14052 based on differentially weighted properties, where four attributes(mw, xlogp, hbond donor count and hbond acceptor count), which attributed to pharmacokinetic properties were given more weight than the others

Pubchem Sl No.	**cid**	**mw**	Polar area	complexity	**xlogp**	heavycnt	**hbonddonor**	**hbondacc**	rotbonds	annothitcnt	**Weighted Dissimilarity**	**Weighted Dis Rank**
	14052	204.31	64.5	109	-0.1	14	4	4	9	15		
1	**3279**	204.31	64.5	109	0	14	4	4	9	15	**0.015385**	**1**
2	**162045**	204.31	64.5	123	0.4	14	4	4	9	0	**1.580203**	**6**
3	**465436**	204.31	64.5	109	0	14	4	4	9	1	**1.077033**	**4**
4	**469919**	188.31	35.5	105	0	13	2	3	8	1	**3.524252**	**14**
5	**469924**	202.34	35.5	129	0	14	2	3	8	1	**2.952848**	**11**
6	**469926**	218.34	64.5	120	0	15	4	4	10	1	**2.558665**	**9**
7	**469930**	218.34	64.5	145	0	15	4	4	9	3	**3.631228**	**15**
8	**469973**	219.32	81.8	147	0	15	4	5	9	1	**4.103347**	**16**
9	**469988**	204.31	64.5	128	0	14	4	4	7	2	**1.777639**	**7**
10	**469989**	232.36	64.5	149	0	16	4	4	9	1	**5.410515**	**18**
11	**469990**	204.31	64.5	109	0	14	4	4	9	1	**1.077033**	**4**
12	**469991**	232.36	64.5	145	0	16	4	4	9	3	**5.212213**	**17**
13	**469996**	204.31	44.7	123	0	14	2	4	9	0	**2.214904**	**8**
14	**470016**	192.3	45.3	96.5	0	13	4	4	7	0	**2.807313**	**10**
15	**470071**	204.31	64.5	109	0	14	4	4	9	6	**0.692479**	**2**
16	**470072**	204.31	64.5	109	0	14	4	4	9	6	**0.692479**	**2**
17	**470075**	188.31	44.3	109	0	13	3	3	8	1	**3.113323**	**12**
18	**470076**	188.31	44.3	121	0	13	3	3	7	1	**3.250016**	**13**

Table 5: Recommendation list for the cid 14052 based on four network based properties

Pub Rank	**cid**	Avg degree	Avg closeness	Avg Nbetween	Avg Ebetween	**N/w Weighted Dissimilarity**	**N/w Weighted Rank**
	14052	1.833	0.030437	11.667	18.7272		
1	**3279**	1.8571	0.025008	14.2857	22.3846	**4.49830799**	**9**
2	**162045**	1.833	0.030437	11.667	18.7272	**0**	**1**
3	**465436**	1.8461	0.02755192	12.923	20.5	**2.1726794**	**6**
4	**469919**	1.7778	0.0511514	6.222	11.5	**9.04897348**	**13**
5	**469924**	1.8182	0.0358856	9.5454	16	**3.45529378**	**7**
6	**469926**	1.8571	0.025008	14.2857	22.3846	**4.49830799**	**9**
7	**469930**	1.875	0.0194469	19.1875	28.4667	**12.3051886**	**15**
8	**469973**	1.875	0.020659	17.6875	28.8667	**11.7922746**	**14**
9	**469988**	1.8889	0.016112	23.7778	34.1765	**19.6304783**	**17**
10	**469989**	1.8889	0.016112	23.7778	34.1765	**19.6304783**	**17**
11	**469990**	1.8571	0.025008	14.2857	22.3846	**4.49830799**	**9**
12	**469991**	1.8889	0.0167056	22.5556	32.8823	**17.8586612**	**16**
13	**469996**	1.8182	0.037331	9	15.4	**4.26420161**	**8**
14	**470016**	1.7333	0.011509	8.6	15.2307	**4.6521285**	**12**
15	**470071**	1.833	0.030437	11.667	18.7272	**0**	**1**
16	**470072**	1.833	0.030437	11.667	18.7272	**0**	**1**
17	**470075**	1.846	0.028601	12.1538	19.6667	**1.05820927**	**4**
18	**470076**	1.8333	0.031711	10.9167	17.9091	**1.1100628**	**5**

Table 6: Final recommendation list for the TB drug EMB (cid 14052). "Consolidated Rank" column ranked the drugs as per the customized Borda ranking, and "Refined Recommendation" provided the final recommendation ranks after revisiting Network Based dissimilarities.

Sl. No.	cid	Uniform D Rank	Weighted D Rank	N/w D Rank	Modified Borda Value	**Consolidated Rank**	**Refined Recommendation**
1	3279	1	1	9	37	**3**	**1**
2	162045	6	6	1	35	**4**	
3	465436	4	4	6	34	**5**	**2**
4	469919	13	14	13	8	**14**	**11**
5	469924	14	11	7	16	**11**	**8**
6	469926	7	9	9	23	**7**	**4**
7	469930	15	15	15	3	**15**	**12**
8	469973	16	16	14	2	**16**	**13**
9	469988	8	7	17	16	**12**	**9**
10	469989	18	18	17	-5	**18**	**15**
11	469990	4	4	9	31	**6**	**3**
12	469991	17	17	16	-2	**17**	**14**
13	469996	9	8	8	23	**8**	**5**
14	470016	11	10	12	15	**13**	**10**
15	470071	2	2	1	43	**1**	
16	470072	2	2	1	43	**2**	
17	470075	10	12	4	22	**9**	**6**
18	470076	12	13	5	18	**10**	**7**

Note: The grey coloured entries showed the ranks of the cids which were below the rank of 15. Our proposed modification to the Borda ranking incorporated these ranks also. The violet entries signified drugs which had identical network based attribute properties to the queried drug, EMB with cid 14092.

Figure 1: Block Diagram of proposed methodology

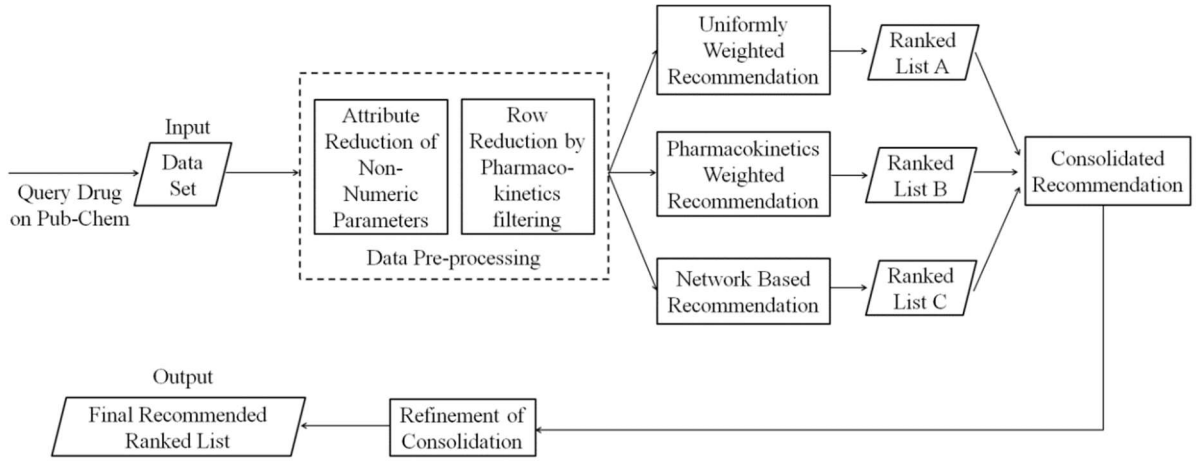

Figure 2: Graphical representation of the network attributes for the queried drug EMB (cid 14052)

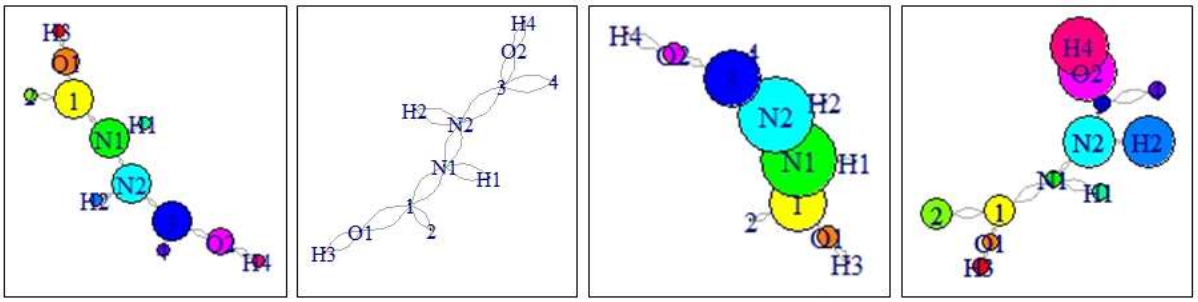

Figure 3: Graphical representation of the network attributes for the top ranked drug, Myambutol (cid 3279)