**A peer-reviewed version of this preprint was published in PeerJ on 17 July 2014.**



Jamal S, Scaria V, Open Source Drug Discovery Consortium. 2014. Data-mining of potential antitubercular activities from molecular ingredients of traditional Chinese medicines. PeerJ 2:e476
https://doi.org/10.7717/peerj.476

# Data-mining of potential antitubercular activities from molecular ingredients of Traditional Chinese Medicines

**Background** Traditional Chinese medicine encompasses a well established alternate system of medicine based on a broad range of herbal formulations and is practiced extensively in the region for the treatment of a wide variety of diseases. In recent years, several reports describe in depth studies of the molecular ingredients of Traditional Chinese Medicines on the biological activities including anti-bacterial activities. The availability of a well-curated dataset of molecular ingredients of Traditional Chinese Medicines and accurate in-silico cheminformatics models for data mining for antitubercular agents and computational filters to prioritize molecules has prompted us to search for potential hits from these datasets.

**Results** We used a consensus approach to predict molecules with potential antitubercular activities from a large dataset of molecular ingredients of Traditional Chinese Medicines available in the public domain. We further prioritized 160 molecules based on five computational filters (SMARTSfilter) so as to avoid potentially undesirable molecules. We further examined the molecules for permeability across Mycobacterial cell wall and for potential activities against non-replicating and drug tolerant Mycobacteria. Additional in-depth literature surveys for the reported antitubercular activities of the molecular ingredients and their sources were considered for drawing support to prioritization.

**Conclusions** Our analysis suggests that datasets of molecular ingredients of Traditional Chinese Medicines offer a new opportunity to mine for potential biological activities. In this report, we suggest a proof-of-concept methodology to prioritize molecules for further experimental assays using a variety of computational tools. We also additionally suggest that a subset of prioritized molecules could be used for evaluation for tuberculosis due to their additional effect against non-replicating tuberculosis as well as the additional hepato-protection offered by the source of these ingredients.

*PeerJ PrePrints* | http://dx.doi.org/10.7287/peerj.preprints.276v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014

# Data-mining of potential antitubercular activities from molecular ingredients of Traditional Chinese Medicines

Salma Jamal, Open Source Drug Discovery Consortium, Vinod Scaria[,$]

[1] CSIR Open Source Drug Discovery Unit, Anusandhan Bhavan, 2 Rafi Marg, Delhi 110001
[2] GN Ramachandran Knowledge Center for Genome Informatics, CSIR Institute of Genomics and Integrative Biology (CSIR-IGIB), Mall Road, Delhi 110007, India

[$] Address for correspondence: vinods@igib.in



## Abstract

*PeerJ PrePrints* | http://dx.doi.org/10.7287/peerj.preprints.276v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014

## Background

Traditional Chinese medicine encompasses a well established alternate system of medicine based on a broad range of herbal formulations and is practiced extensively in the region for the treatment of a wide variety of diseases. In recent years, several reports describe in depth studies of the molecular ingredients of Traditional Chinese Medicines on the biological activities including anti-bacterial activities. The availability of a well-curated dataset of molecular ingredients of Traditional Chinese Medicines and accurate in-silico cheminformatics models for data mining for antitubercular agents and computational filters to prioritize molecules has prompted us to search for potential hits from these datasets.

## Results

We used a consensus approach to predict molecules with potential antitubercular activities from a large dataset of molecular ingredients of Traditional Chinese Medicines available in the public domain. We further prioritized 160 molecules based on five computational filters (SMARTSfilter) so as to avoid potentially undesirable molecules. We further examined the molecules for permeability across Mycobacterial cell wall and for potential activities against non-replicating and drug tolerant Mycobacteria. Additional in-depth literature surveys for the reported antitubercular activities of the molecular ingredients and their sources were considered for drawing support to prioritization.

## Conclusions

Our analysis suggests that datasets of molecular ingredients of Traditional Chinese Medicines offer a new opportunity to mine for potential biological activities. In this report, we suggest a proof-of-concept methodology to prioritize molecules for further experimental assays using a variety of computational tools. We also additionally suggest that a subset of prioritized molecules could be used for evaluation for tuberculosis due to their additional effect against non-replicating tuberculosis as well as the additional hepato-protection offered by the source of these ingredients.



## Keywords:

Tuberculosis, Traditional Chinese Medicine, Cheminformatics, Virtual Screening, Data-mining

# Introduction

Traditional Medicine still forms the mainstay of healthcare in many parts of the world. Traditional Chinese Medicine (TCM) is one of the well developed and established systems of traditional medicine, and largely followed in some parts of Eastern Asia where it forms one of the major alternative medicinal practices [1]. TCM as a system of medicine was, founded almost 2000 years ago and is dependent on the concepts of five elements and guided by the Chinese philosophy of Ying and Yang [2, 3]. Recently, efforts have been underway to investigate the practice of TCM using molecular approaches. This has led to the identification and molecular characterization of ingredients used in Traditional Chinese Medicines [4, 5]. These efforts have led to the systematic curation of the molecular structures and the biological activities of ingredients of Traditional Chinese Medicines [6-9]. In addition, molecular basis of the action and mechanisms of modulation [10, 11], immunomodulatory and antimicrobial activities of Traditional Chinese Medicines have also been actively pursued [12, 13].

Tuberculosis is considered one of the major tropical diseases, caused by intracellular pathogen *Mycobacterium tuberculosis*. According to the World Health Organization (WHO) Global Tuberculosis Report 2012, Tuberculosis causes over 1.4 million deaths annually worldwide and a major cause of morbidity and mortality especially in the developing countries in Asia and Africa [14]. The paucity of new drugs for the treatment of Tuberculosis along with the rampant and unprecedented rise of drug-resistant strains made it imperative to discover potential new drugs for tuberculosis [15]. The conventional process of drug discovery involves screening of large molecular libraries of molecules for biological activities, and it is a tedious, expensive and time-consuming process [16]. Data mining approaches based on cheminformatics modeling has been extensively used to prioritize molecules from large chemical datasets for specific biological activities. Such in-silico prioritization of molecules has been suggested to accelerate drug discovery by drastically reducing the time and cost-factor in conventional drug discovery processes [17-20].

Cheminformatics and data mining approaches have been used to mine biological activities from molecular data sets of ingredients in traditional Chinese Medicines [21, 22].The availability of large molecular databases with systematically curated molecular data, sources and activities of ingredients of Traditional Chinese Medicines offer a new opportunity to use advanced data-mining tools to mine for potential activities, especially for pathogens causing neglected tropical diseases [6-9]. Previously we used high-throughput bioassay data sets to create highly accurate data-mining classifiers based on machine learning of molecular properties including antimicrobial activities for a number of neglected tropical diseases including Tuberculosis, and Malaria [23-25].



*PeerJ PrePrints* | http://dx.doi.org/10.7287/peerj.preprints.276v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014

In the present report, we used one of the largest and well characterized compilation of molecular ingredients in traditional Chinese Medicine and applied a host of previously generated cheminformatics models aimed at identifying potential hits with antitubercular activity against Tuberculosis. We additionally employed methodologies for filtering out potential molecules using a series of in-silico filters. Our analysis revealed a total of 19 hits for antitubercular activity from the dataset. In-depth literature survey suggests 4 of these molecules are derived from plant products known to be used against tuberculosis, suggesting that the computational approach can be immensely useful in identifying and characterizing molecular activities. To the best of our knowledge, this is the first and most comprehensive data-mining and cheminformatic analysis of potential antitubercular agents from traditional Chinese medicine ingredients.

# Materials and Methods

## Data Sets

Molecular Data Sets of ingredients of Traditional Chinese Medicines were retrieved from Traditional Chinese Medicines Integrated Database (TCMID) [27]. TCMID constitutes one of the most comprehensive online resources for ingredients used in TCM. The database hosts information on over 25, 210 pure molecules retrieved from literature and other data resources.

## Computational models for antitubercular activity

The computational predictive models used in our analysis were based on the following two confirmatory screens conducted to identify novel inhibitors of *Mycobacterium tuberculosis* H37Rv, previously published by our group [23, 24]. The computational models used are available online at http://vinodscaria.rnabiology.org/2C4C/models.

Briefly these models were based on two bioassays deposited in PubChem and carrying IDs AID 1332 and AID 449762. Both the assays were based on microdilution Alamar Blue assays. The former used 7H12 broth while the latter used 7H9 media. A total of 1,120 and 327, 669 compounds were screened in the respective assays. The models were generated using a machine learning approach as described in Periwal et al and Periwal et al [23, 24]. For the AID 1332 assay model was generated based on the Random forest classification algorithm and was evaluated using a variety of statistical measures which include accuracy, Balanced Classification Rate (BCR) and Area under Curve (AUC). Balanced Classification Rate is an average of sensitivity and specificity which introduces a balance in the classification rate. The model had an accuracy of 82.57%, BCR value of 82.2% and AUC value of 0.87. The AID 449762 assay model was generated based on SMO (Sequential Minimization Optimization) algorithm and was found to be 80.52 % accurate, with BCR value of 66.30% and AUC as 0.75.

In addition, we created an additional model to predict the molecules active against non-replicating drug tolerant *Mycobacterium tuberculosis*. The assay was deposited in PubChem with identifier AID 488890. A total of 3, 24, 437 compounds were screened for the activity. The model was generated using Random forest classification algorithm as described in the previous papers [23-26] and had an accuracy of 76%, BCR value 85.2% and AUC 0.66.

## Molecular Descriptors

Molecular descriptors for each of the molecules were computed using PowerMV [28], popular cheminformatics software widely used to compute molecular descriptors. A total of 179 molecular descriptors were computed for each molecule. Out of the total 179 molecular descriptors, a few descriptors were pruned using bespoke scripts written in Perl depending on whether they were used in creating the respective models. We pruned a total of 29 and 25 descriptors corresponding to AID 1332 and AID 449762 respectively, while 25 were pruned for the AID 488890 model.

## Formats and Format conversion

The molecules were downloaded in mol2 format and converted to SDF (Structural Data Format) format using Openbabel [29]. The molecular descriptors were converted to ARFF format compatible with Machine learning toolkit Weka [30]. We used custom scripts written in Perl for the format conversions. A complete list of scripts is also available at Crowd Computing for Cheminformatics (2C4C) repository at URL: http://vinodscaria.rnabiology.org/2C4C/models.

## SMARTS filters

The SMARTS filter is employed to remove the molecules with fragments leading to toxicity or unwanted reactivity. We used a set of SMARTS filters for the consensus candidate anti-tubercular molecules. The online server SMARTSfilter (http://pasilla.health.unm.edu/tomcat/biocomp/smartsfilter) web application was used for all comparisons. The web application was used to filter out molecules, which match to any of the five undesirable SMARTS catalogs.

## *Mycobacterium tuberculosis* permeability prediction

The small molecules could not be effective unless they are able to penetrate the cell wall. Recent computational tool, MycPermCheck [31], to predict permeability of small molecules across *Mycobacterium tuberculosis* was employed to filter the subset of potential active molecules.

## Data Mining

We used Weka, a popular and freely available Data Mining Software toolkit. Predictions were performed for the dataset across the two models corresponding to assays AID 1332 and AID 449762 independently. Further, molecules predicted active in both the

datasets were collated and analyzed for additional properties including activity against non-replicating drug tolerant *Mycobacterium tuberculosis* and potential to permeate the *Mycobacterium tuberculosis* cell wall. Additional filters which discount molecules with toxic fingerprints were removed using SMARTS filters. The summary of the entire workflow of prioritization is depicted as a Schema (Figure 1).

# Results

### Summary of Datasets and Molecules

A total of 25,210 ingredients were downloaded from Traditional Chinese Medicines Integrated Database (TCMID). We could retrieve molecular information for only 12,018 of the ingredients in the form of SMILE notations and the rest were not considered for further analysis. The molecules considered along with their SMILES are detailed in Supplementary Table 1. A total of 179 descriptors were calculated using PowerMV as described above. The descriptors were further pruned for each of the models as described in the Materials and Methods section using custom scripts in Perl. This corresponds to 150 and 154 descriptors respectively for models AID 1332 and AID 449762 and 154 for AID 488890. The models, descriptors and scripts for formatting the files are available at the Crowd Computing for Cheminformatics Model Repository [http://vinodscaria.rnabiology.org/2C4C/models].

### Prediction of potential anti-tubercular hits

The 12, 018 molecules obtained from TCMID were analyzed for the antitubercular activity using the computational predictive models as described above. The AID 1332 and AID 449762 models predicted 2, 363 compounds and 5, 864 compounds respectively as potentially active anti-tubercular. Of these molecules, a total of 1,472 molecules were predicted potential actives by both the models based on molecular descriptors and were considered for further analysis (Supplementary Table 2).

Briefly we used a popular approach for filtering molecules with undesirable properties. These included briefly using SMARTS filters. Molecules which passed the filtering step were further evaluated for their effect against drug-tolerant and slow growing Mycobacterium. Molecules were further evaluated for their potential permeability with respect to the Mycobacterial cell wall.

### SMARTS filter for filtering undesirable structures

We used a set of five SMARTS filters to remove the molecules matching to any of these filters. Such substructure based filtering approach has been extensively used to prioritize molecules by filtering unwanted or potential false positives in cheminformatics screens [32]. The SMARTS filters included 5 independent approaches namely Glaxo, PAINS, Oprea, Blake and ALARM-NMR used in tandem. Pan Assay Interference Compounds (PAINS) describes a set of substructures known to be promiscuous and have issues in high throughput assays [33], while the Glaxo filter describes unsuitable

hits or unsuitable natural products [34]. ALARM NMR assay to detect reactive molecules by nuclear magnetic resonance (ALARM-NMR) set filters for molecules which are reactive false positives in high-throughput assays by oxidizing or alkylating a protein target [35]. The Glaxo, Oprea and Blake filters were based on specific fitness properties. The Glaxo method involves classification of the molecules into different chemical categories based on the presence of acids, bases, electrophiles and nucleophiles in the molecule. Prior to the categorization the molecules are filtered for non-drug like properties and to remove inappropriate functional groups (unsuitable leads and unsuitable natural products) [34].

Out of a total of 1472 molecules, 160 molecules passed all the filters. A total of 63.1% (929) molecules failed the ALARM NMR filter, while 49.9% (734) failed to pass Oprea filter. Similarly 49% (722) failed to pass the PAINS filter. The detailed schema showing the number of molecules failed by each filter is depicted in Figure 3. A similar comparison of the complete set of 12, 018 TCMID compounds revealed that only 1,539 compounds passed all the filters. We observed that most of the molecules did not pass through ALARM NMR (60.7%, 7, 295) molecules followed by Oprea filter (52.4%, 6,303) molecules and 5,799, 48.3% molecules could not pass through PAINS filter.

**Molecules potentially active against non-replicating drug tolerant *Mycobacterium tuberculosis*.**

A total of 160 compounds filtered through SMARTSfilter were tested using a computational predictive model for potential activity against non-replicative *Mycobacterium tuberculosis.* The model predicted 19 compounds as active to act as potential inhibitors of non-replicating drug tolerant *Mycobacterium tuberculosis.* The detailed description about 19 compounds is given in Table 1. The table also shows the permeability probability of the molecules to pass through Mtb cell wall.

***Mycobacterium tuberculosis* permeability prediction**

We employed the MycPermCheck a recently published methodology to predict molecular permeability to Mycobacterial cell wall to estimate the potential permeability of the prioritized molecules. All the 160 molecules which passed the five SMARTSfilters were further evaluated for their ability to penetrate Mtb cell wall. Analysis revealed 9 molecules with highest probability (>0.98) to permeate *Mycobacterium* cell wall barrier (Supplementary Table 1).

**Literature search suggests evidence of the sources and molecules used with antitubercular properties**

We further searched for the role of the plant sources of the molecules in regard to their use or known information on antibacterial or anti tubercular activities. We found several molecules herbs to have antitubercular effects. These are *Petasites japonicus* [36],

*Piper trichostachyon* [37], *Solanum torvum* [38], *Fritillaria przewalskii* [39], *Hernandia sonora* 40] and *Phyllanthus urinari* [12]. In addition, many of the herbs have been shown to have hepatoprotective activities, which include *Annona reticulata* [41, 42], *Annona squamosa* [41, 42], and *Camellia sinensis* [43]. This offers a new opportunity for new drug development considering that most of the established first-line drugs used in the treatment of tuberculosis are hepatotoxic [44, 45]. We also found the molecules, Hinokiol [46], Totarol [47], Murrayafoline a [48] and 2-hexenyl benzoate [49] have been known to show antitubercular effects.

## Discussion and Conclusions

Traditional Chinese Medicine (TCM) has been a major alternative medicine practice, widely followed in many parts of China and Southeast Asia [1]. Enormous efforts in the recent years have been invested in the systematic identification and characterization of the molecular activities of the ingredients and scientific validation of their effects [10, 11]. The availability of well curated databases of ingredients of Traditional Chinese Medicines has opened up new avenues for molecular screening as well as in-silico studies, including target-based docking [6-9]. In depth screens of Chinese Medicine derived compounds have been performed for a variety of pathophysiologies, including cancer [50], inflammatory diseases [51, 52], cardiovascular diseases [53] and infections [54] etc, just to name a few. These databases are being extensively used for therapeutic development [55].

Our group has earlier used a machine learning based approach on publicly available high-throughput screen datasets to create highly accurate models for predicting specific molecular activities against pathogens causing Tuberculosis [23, 24] and Malaria [25]. Such accurate in-silico models offer a new opportunity to prioritize large molecular databases in silico, significantly reducing the failures, cost and effort. The availability of a well-curated database of molecular ingredients of traditional Chinese Medicines offer a new opportunity to mine potential active anti-tubercular agents and prioritize them for screening and in-depth functional assays.

In the present study, we have used two computational models based on high throughput assays on *Mycobacterium tuberculosis*. In addition to the predictive models, we used a filter based approach to filter out potential false positives/toxic molecules. Our analysis revealed a total of 1,472 molecules predicted active by both the models, of which 160 molecules passed all the five filters. These molecules were further evaluated for their permeability to mycobacterial cell wall and potential additional activity on drug-tolerant and non-replicating Mycobacterium tuberculosis. We also further show evidence from literature that these molecules or their sources have been used in the treatment of therapeutics. This study is not without caveats; the primary one being that the consensus approach used in the present study could be over-stringent so as to miss out

on potential antitubercular hits from the screening approach. The second, being that the findings would require re-screening and in-depth functional analysis. Nevertheless we show from independent evidence that molecular ingredients or sources of the prioritized molecules have been extensively used as antibacterial or specifically in the treatment of tuberculosis. In the present study we show a proof-of-concept that data-mining approaches using accurate cheminformatics models could possibly be used to mine large datasets and prioritize molecules for antitubercular screening.

Our analysis suggests that molecular ingredients of Traditional Chinese Medicines offer an attractive starting point to mine for potential antitubercular agents. Chinese Medicines alone [56] or in combination [57] with western medicine have been explored for the treatment of tuberculosis. Potential use of Chinese Medicines in combination with the standard antitubercular drugs could be an attractive alternative that could be explored in much detail. There is ample evidence in published literature that some of the ingredients of the short-listed antitubercular molecules have additional hepatoprotective action, which could be effectively used in the background of hepatotoxicity induced by the first line of drugs. We also suggest that 19 of the prioritized molecules have additional activity against drug-tolerant and non-replicating *Mycobacterium tuberculosis* suggesting that they could be potentially developed into leads for Multidrug resistant and latent tuberculosis.. We hope that this report would accelerate in in-depth analysis and discovery of anti-tubercular agents from molecular ingredients of Traditional Chinese Medicines.

## Competing interests

The authors declare that they have no competing interests.

## Authors' contributions

SJ under the supervision of VS carried out the analysis and reviewed the results. OSDDC supported the work through regular discussions and funding. Both authors wrote, reviewed and approved the final manuscript.

## Acknowledgements

The authors thank Ms Vinita Periwal for sharing the models for antitubercular activities and for discussions. Authors also thank Dr S Ramachandran and Dr Sridhar Sivasubbu for reviewing the manuscript and suggestions. The help and support from the National Knowledge Network (NKN) and the CDAC-Garuda grid for the connectivity and access to the compute facility is acknowledged. This work was funded by the Council of Scientific and Industrial Research (CSIR), India through Open Source Drug Discovery Project (HCP001).

# References


1. Ooi GL: **Chinese medicine in Malaysia and Singapore: the business of healing.** *Am J Chin Med.* 1993, **21(3-4):**197-212.
2. Qiu J: **Traditional medicine—a culture in the balance.** *Nature* 2007, **448:**126–128.
3. Normile D: **Asian medicine. The new face of traditional Chinese medicine.** *Science* 2003, **299:** 188–190.
4. Wang MW, Hao X, Chen K: **Biological screening of natural products and drug innovation in China.** *Philos. Trans. Roy. Soc. Lond. B Biol. Sci.* 2007, **362:**1093–1105.
5. Sucher NJ: **Insights from molecular investigations of traditional Chinese herbal stroke medicines: implications for neuroprotective epilepsy therapy.** *Epilepsy Behav.* 2006, **8(2):**350-62.
6. Chen X, Zhou H, Liu YB, Wang JF, Li H, Ung CY, Han LY, Cao ZW, Chen YZ**:** **Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation.** *Br. J. Pharmacol.* 2006, **149:**1092–1103.
7. Fang YC, Huang HC, Chen HH, Juan HF: **TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining.** *BMC Complement. Altern. Med.* 2008, **8:**58.
8. Chen CYC: **TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico.** *Plos One* 2011, **6:**e15939.
9. Zhou J, Xie G, Yan X: *(eds) Encyclopedia of Traditional Chinese Medicines – Molecular Structures, Pharmacological Activities, Natural Sources and Applications.* Springer: New York; 2011.
10. Li XM, Brown L: **Efficacy and mechanisms of action of traditional Chinese medicines for treating asthma and allergy.** *Journal of Allergy and Clinical Immunology* 2009, **123(2):**297–306.
11. Wen Z, Wang Z, Wang S, Ravula R, Yang L, et al.: **Discovery of Molecular Mechanisms of Traditional Chinese Medicinal Formula Si-Wu-Tang Using Gene Expression Microarray and Connectivity Map.** *PLoS ONE* 2011, **6(3):**e18278.

12. Nair RR, Abraham RS: **Integrating the Science of Pharmacology and Bio Informatics Phyllanthus "The wonder plant".** *Advanced Biotech January* 2008.
13. Nader LA, de Mattos AA, Picon PD, Bassanesi SL, De Mattos AZ, Pineiro Rodriguez M: **Hepatotoxicity due to rifampicin, isoniazid and pyrazinamide in pateints with tuberculosis: is anti-HCV a risk factor?** *Ann Hepatol.* 2010, **9(1):**70-74.
14. World Health Organization 2012. http://www.who.int/tb/publications/global_report/gtbr12_main.pdf
15. Shah NS, Wright A, Bai GH, Barrera L, Boulahbal F, et al.: **Worldwide emergence of extensively drug-resistant tuberculosis.** *Emerg Infect Dis.* 2007, **13(3):**380-7.
16. DiMasi, JA, Hansen, RW, Grabowski, HG: **The price of innovation: new estimates of drug development costs.** *Journal of Health Economics* 2003, **22:**151–185.
17. Vert JP, Jacob L: **Machine learning for in silico virtual screening and chemical genomics: new strategies.** *Comb Chem High Throughput Screen* 2008, **11:**677-685.
18. Melville JL, Burke EK, Hirst JD: **Machine Learning in Virtual Screening.** *Comb Chem High Throughput Screen* 2009, **12:**332-343.
19. Schierz AC: **Virtual screening of bioassay data.** *J Cheminform* 2009, **1:**21.
20. Vasanthanathan P, Taboureau O, Oostenbrink C, Vermeulen NP, Olsen L, Jorgensen FS: **Classification of cytochrome P450 1A2 inhibitors and non-inhibitors by machine learning techniques.** *Drug Metab Dispos* 2009, **37:**658-664.
21. Li XJ, Kong DX, Zhang HY: **Chemoinformatics approaches for traditional Chinese medicine research and case application in anticancer drug discovery.** *Curr Drug Discov Technol.* 2010, **7(1):**22-31.
22. Zhang K, Li Y, Zhang ZR, Guan WH, Pu YC: **Chemoinformatics study on antibacterial activity of traditional Chinese medicine compounds.** *Zhongguo Zhong Yao Za Zhi.* 2013, **38(5):**777-80.
23. Periwal V, Rajappan JK, Jaleel AU, Scaria V: **Predictive models for anti-tubercular molecules using machine learning on high-throughput biological screening datasets.** *BMC Res Notes* 2011, **4:**504.
24. Periwal V, Kishtapuram S, Scaria V: **Computational models for in-vitro antitubercular activity of molecules based on high-throughput chemical biology screening datasets.** *BMC Pharmacol* 2012, **12:**1.

25. Jamal S, Periwal V, Scaria V: **Predictive modeling of anti-malarial molecules inhibiting apicoplast formation.** *BMC Bioinformatics* 2013, **14:**55.
26. Jamal S, Periwal V, Scaria V: **Computational analysis and predictive modeling of small molecule modulators of microRNA.** *Journal of Cheminformatics* 2012, **4:**16.
27. Ruichao Xue, Zhao Fang, Meixia Zhang, Zhenghui Yi, Chengping Wen, Tieliu Shi: **TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis.** *Nucleic Acids Research* 2013, **41:**D1089–D1095.
28. Liu K, Feng J, Young SS: **PowerMV: a software environment for molecular viewing, descriptor generation, data analysis and hit evaluation.** *J Chem Inf Model* 2005, **45:**515-522.
29. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open Babel: An open chemical toolbox.** *J Cheminform* 2011, **7(3):**33.
30. Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P: **Weka -Experiences with a Java Open-Source Project.** *Journal of Machine Learning Research* 2010, 2533–2541.
31. Merget B, Zilian D, Müller T, Sotriffer CA: **MycPermCheck: The Mycobacterium tuberculosis permeability prediction tool for small molecules.** *Bioinformatics* 2013, **29(1):** 62-68.
32. Singla D, Tewari R, Kumar A, Raghava GP: **Open Source Drug Discovery Consortium. Designing of inhibitors against drug tolerant Mycobacterium tuberculosis (H37Rv).** *Chem Cent J.* 2013, **7(1):** 49.
33. Baell JB, Holloway GA: **New Substructure Filters for Removal of Pan Assay Interference Compounds (PAINS) from Screening Libraries and for their Exclusion in Bioassays.** *J. Med. Chem* 2010, **53:**2719-2740.
34. Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N: **Strategic pooling of compounds for high-throughput screening.** *J Chem Inf Comput Sci.* 2009, **39(5):**897-902.
35. Huth JR, Mendoza R, Olejniczak ET, Johnson RW, et al.: **ALARM NMR: A Rapid and Robust Experimental Method to Detect Reactive False Positives in Biochemical Screens.** *J. Am. Chem. Soc.* 2005, **127:**217-224.
36. John EF: **A Barefoot Doctors Manual: The American Translation of the Official Chinese Paramedical Manual.** 1999.

37. Wagner H, Wolff P: **New Natural Products and Plant Drugs with Pharmacological, Biological or Therapeutical Activity.** Springer-Varlag Berlin; 1977:p212.
38. Silva KN, Silva RC, Coelho VPM, Agra M: **A pharmacobotanical study of vegetative organs of Solanum torvum.** *Brazilian Journal of Pharmacognosy* 2011, **21(4):**568-574.
39. Chang HM, Paul PH: *Pharmacology and Applications of Chinese Materia Medica*. 2001.
40. Udino L, Abaul J, Bourgeois P, Gorrichon L, Duran H, Zedde C: **Lignans from the Seeds of Hernandia sonora.** *Planta Med.* 1999, **65(3):**279-81.
41. Thattakudian Sheik Uduman MS, Sundarapandian R, Muthumanikkam A, et al.: **Protective effect of methanolic extract of Annona squamosa Linn in isoniazid rifampicin induced hepatotoxicity in rats.** *Pak J Pharm Sci.* 2011, **24(2):**129-34.
42. Mohamed Saleem TS, Christina AJM, Chidambaranathan N, Ravi V: **Hepatoprotective activity of Annona squamosa Linn. on experimental animal model.** *Int J Pharm Pharm Sci.* 2008, **1(3):**1-7.
43. Issabeagloo E, Taghizadieh M: **Hepatomodulatory Action of Camellia sinensis Aqueous Extract against Isoniazid-Rifampicin Combination Induced Oxidative Stress in Rat.** *Advances in Bioresearch* 2012, **3:**18-27.
44. Yew WW, Leung CC: **Antituberculosis drugs and hepatotoxicity.** *Respirology* 2006, **11(6):**699-707.
45. Liu Q, Garner P, Wang Y, Huang B, Smith H: **Drugs and herbs given to prevent hepatotoxicity of tuberculosis therapy: systematic review of ingredients and evaluation studies.** *BMC Public Health* 2008, **21(8):**365.
46. Chen JJ, Wu HM, Peng CF, Chen IS, Chu SD: **seco-Abietane diterpenoids, a phenylethanoid derivative, and antitubercular constituents from Callicarpa pilosissima.** *J Nat Prod.* 2009, **72(2):**223-8.
47. Jaiswal R, Beuria TK, Mohan R, Mahajan SK, Panda D: **Totarol inhibits bacterial cytokinesis by perturbing the assembly dynamics of FtsZ.** *Biochemistry* 2007, **46(14):**4211-20.
48. Choi TA, Czerwonka R, Fröhner W, Krahl MP, Reddy KR, Franzblau SG, Knölker HJ: **Synthesis and activity of carbazole derivatives against Mycobacterium tuberculosis.** *ChemMedChem* 2006, **1(8):**812-5.

49. He, Yantao, Li-fan ZENG, Zhong-Yin Zhang: **Tyrosine phosphatase inhibitors and uses thereof to modulate the activity of enzymes involved in the pathology of mycobacterium tuberculosis.** WIPO Patent Application PCT/US2012/035039.
50. Hu Y, Wang S, Wu X, Zhang J, Chen R, Chen M, Wang Y: **Chinese herbal medicine-derived compounds for cancer therapy: A focus on hepatocellular carcinoma.** *J Ethnopharmacol.* 2013**, S0378-8741(13)**00531-X.
51. Han C, Guo J: **Antibacterial and anti-inflammatory activity of traditional Chinese herb pairs, Angelica sinensis and Sophora flavescens.** *Inflammation* 2012, **35(3):**913-9.
52. Su SY, Hsieh CL: **Anti-inflammatory effects of Chinese medicinal herbs on cerebral ischemia.** *Chin Med.* 2011, **6:**26.
53. Wang L, Qiu XM, Hao Q, Li DJ: **Anti-inflammatory effects of a Chinese herbal medicine in atherosclerosis via estrogen receptor β mediating nitric oxide production and NF-κB suppression in endothelial cells.** *Cell Death Dis.* 2013, **4:**e551.
54. Jiang L, Deng L, Wu T: **Chinese medicinal herbs for influenza.** *Cochrane Database Syst Rev.* 2013, **3:**CD004559.
55. Cheng HM, Li CC, Chen CY, Lo HY, Cheng WY, Lee CH, Yang SZ, Wu SL, Hsiang CY, Ho TY: **Application of bioactivity database of Chinese herbal medicine on the therapeutic prediction, drug development, and safety evaluation.** *J Ethnopharmacol* 2010, **132(2):**429-37.
56. Lu J, Ye S, Qin R, Deng Y, Li CP: **Effect of Chinese herbal medicine extracts on cell-mediated immunity in a rat model of tuberculosis induced by multiple drug-resistant bacilli.** *Mol Med Rep.* 2013, **8(1):**227-32.
57. Li SZ: **Combined traditional Chinese medicine and western medicine in the treatment of 356 cases of scrofula of the neck and axilla.** *Zhong Xi Yi Jie He Za Zhi* 1984, **4(2):**90-2.

# Tables and Figures

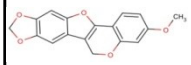
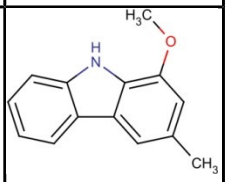
# Figures

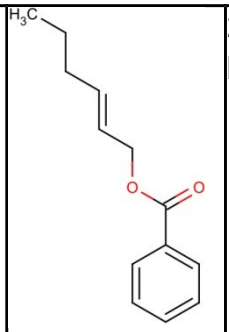
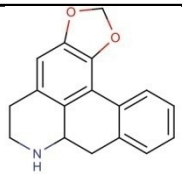
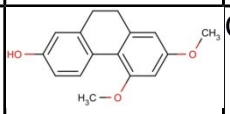
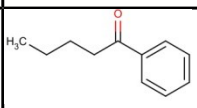
**Figure 1:** Summary of the data-mining and prioritization approach involving prediction of actives, consensus building and filtering for permeability and undesirable substructures.

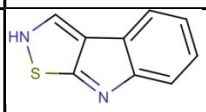
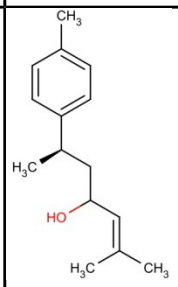
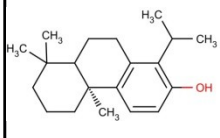
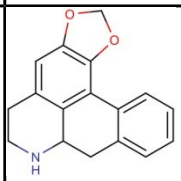
**Figure 2:** Venn diagram showing active molecules filtered by any of the five SMARTS filters.

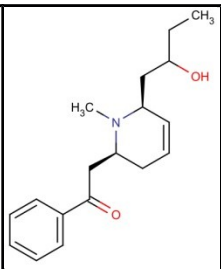
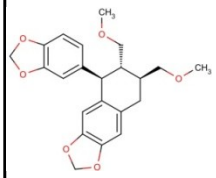
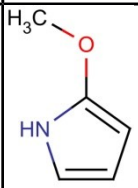
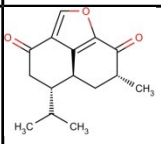
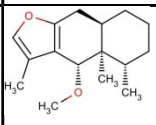
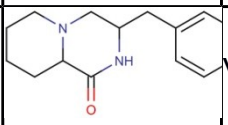
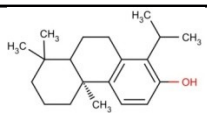
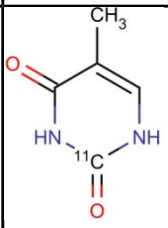
# Tables

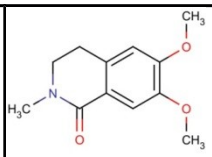
**Table 1 shows the 19 compounds predicted as active against non replicating antibiotic tolerant *Mycobacterium tuberculosis.***

| Compound No. | Compound structure | Name | English Name | Latin Name | Permeability probability | Sources with antitubercular activities |
|---|---|---|---|---|---|---|
| 1. |  | Flemichapparin b | Climbing Jewelvine | Derris scandens | 0.993 | |
| 2. |  | Murrayafoline a | Taiwan Common Jasminorange, Indian Common Jasminorange, Euchretaleaf Common Jasminorange, Narrowfruit Glycosmis Root | Murraya crenulata, Murraya koenigii, Murraya euchrestifolia, Glycosmis stenocarpa | 0.98 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| 3. |  | 2-hexenyl benzoate | Common Tea, Szechwan Tangshen | Camellia sinensis , Codonopsis tangshen | 0.855 | |
| 4. |  | Anonaine | Hindu Lotus Large Rhizome, Bullockshear t Custardappl e, Custard Apple, Chinaberry-tree Flower, Uncinate Tailgrape | Nelumbo nucifera, Annona reticulata, Annona squamosa, Melia azedarach, Artabotrys uncinatus, | 0.52 | |
| 5. |  | Orchinol | Frog Orchid, European Gymnadenia , Liriop Equivalent plant: Liriope spicata var prolifera | Coeloglossu m viride [Syn. Coeloglossu m viride var. bracteatum], Gymnadenia albida, Ophiopogon japonicus | 0.407 | |
| 6. |  | 1-phenyl-1-pentanone | Chuanxiong rhizome, Szechuan lovage root, Chuanxiong (Wallich Ligusticum) Equivalent plant: | Radix chuanxiong Rhizoma Chuanxiong, Ligusticum chuanxiong | 0.338 | |

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | Cnidium officinale | | | |
| 7. |  | Brassilexin | India Mustard | Brassica juncea | 0.295 | |
| 8. |  | Bisacumol | Zedoary Turmeric Equivalent plant: Curcuma kwangsiensis, Common Turmeric Equivalent plant: Curcuma aromatica | Curcuma zedoaria, Curcuma longa | 0.104 | |
| 9. |  | Totarol | Longleaf Podocarpus Leaf Equivalent plant: Podocarpus macrophyllus var maki, Water Nightshade | Podocarpus macrophyllus, Solanum torvum | 0.037 | Solanum torvum [Agra et al.,2011] |
| 10. |  | Cyclostachine a | Hairspike Pepper | Piper trichostachyon | 0.029 | Piper trichostachyon [Wolff et al., 1977] |

| | | | | | | |
|---|---|---|---|---|---|---|
| 11. |  | Isolobinine | Indian Tobacco, Chinese Lobelia | Lobelia inflata, Lobelia chinensis | 0.018 | |
| 12. |  | Urinatetralin | Common Leafflower | Phyllanthus urinaria | 0.012 | Phyllanthus urinaria [Abraham and Nair, 2008] |
| 13. |  | 2-methoxy-1h-pyrrole | | | 0.004 | |
| 14. |  | Gmelofuran | Medicinal Breynia Leaf | Breynia officinalis | 0.00 | |
| 15. |  | Petasalbin methyl ether | Japanese Butterbur | Petasites japonicus | 0.00 | Petasites japonicus [Fogarty, 1990] |
| 16. |  | Verruculotoxin | | | 0.00 | |
| 17. |  | Hinokiol | Yellowish Rabdosia | Isodon flavidus | 0.00 | |
| 18. |  | Thymine | Przewalsk Fritillary, Anhui Fritillary, Ussuri Fritillary | Fritillaria przewalskii, Fritillaria anhuiensis, Fritillaria ussuriensis | 0.00 | Fritillaria przewalskii [Chang and Paul, 2001] |

| | | | | | | |
|---|---|---|---|---|---|---|
| 19. |  | n-methylcorydaldine | Fendler's Meadowrue, Bracteate Poppy, Asiatic Moonseed Root, Lotusleaftung | Thalictrum fendleri, Papaver bracteatum, Menispermum dauricum, Hernandia sonora | 0.00 | Hernandia sonora [Bourgeois et al., 1999] |



462 ## Supplementary Data

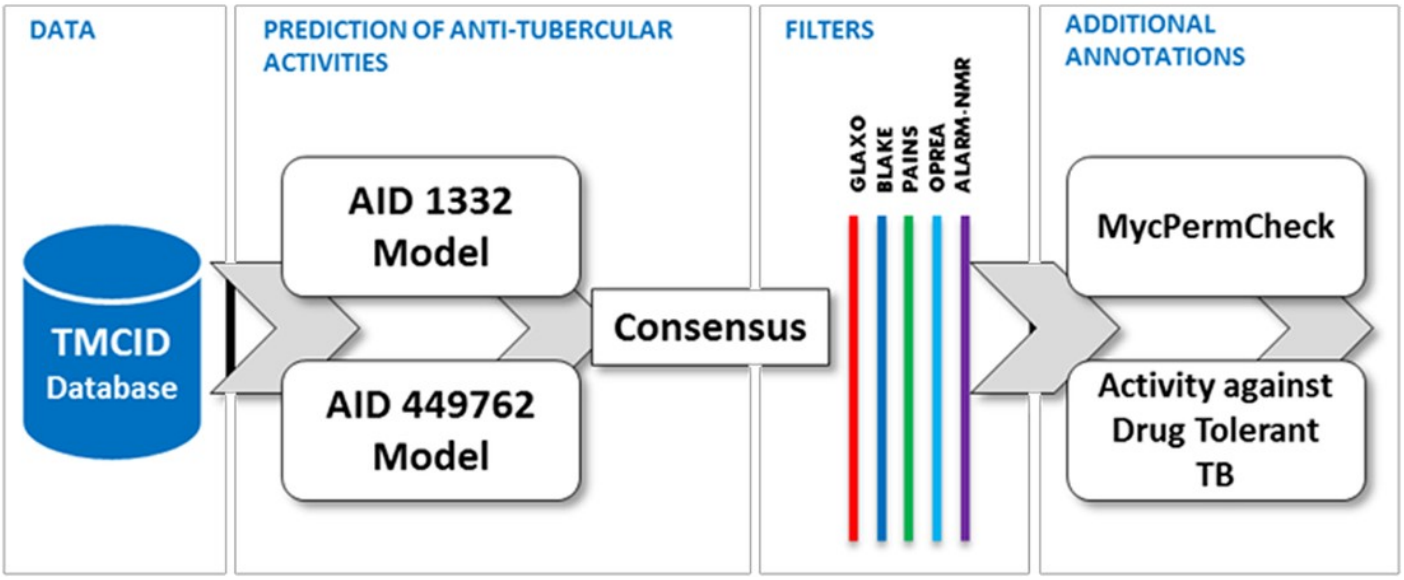
463 **Supplementary Table 1** shows the Chinese molecules used in the present study with
464 their smiles.
465 **Supplementary Table 2** shows the molecules predicted to have anti tubercular activity
466 by our models.
467 **Supplementary Table 3** shows the 9 molecules which could penetrate the
468 *Mycobacterium tuberculosis* cell wall.

469

# Figure 1

Figure 1

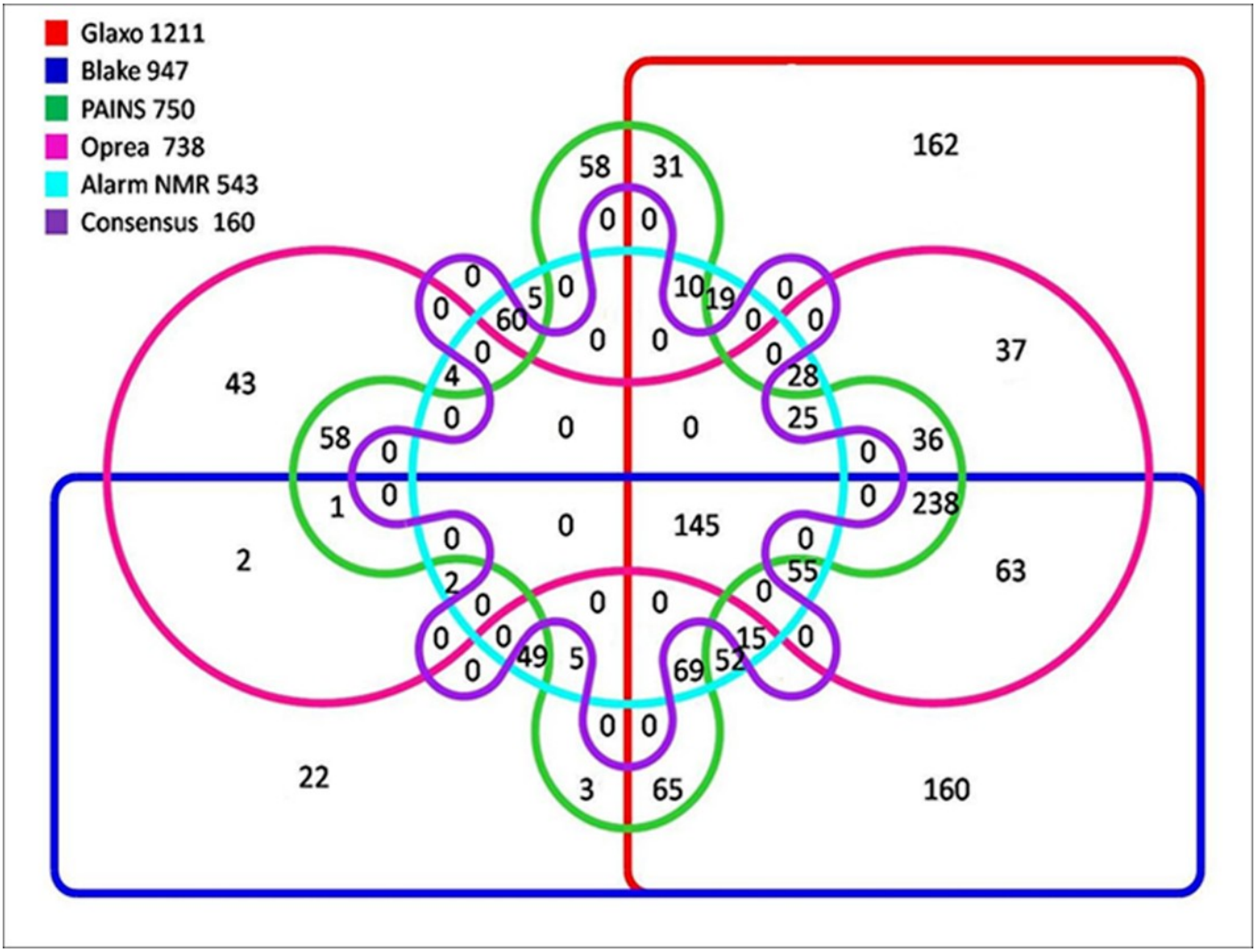
Summary of the data-mining and prioritization approach involving prediction of actives, consensus building and filtering for permeability and undesirable substructures.



# Figure 2

Figure 2

Venn diagram showing active molecules filtered by any of the five SMARTS filters.






*PeerJ PrePrints* | http://dx.doi.org/10.7287/peerj.preprints.276v1 | CC-BY 4.0 Open Access | received: 7 Mar 2014, published: 7 Mar 2014