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

View the peer-reviewed version (peerj.com/articles/476), which is the preferred citable publication unless you specifically need to cite this preprint.

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.

1   **Data-mining of potential antitubercular activities from**
2   **molecular ingredients of Traditional Chinese**
3   **Medicines**

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

5   [1] CSIR Open Source Drug Discovery Unit, Anusandhan Bhavan, 2 Rafi Marg, Delhi
6   110001
7   [2] GN Ramachandran Knowledge Center for Genome Informatics, CSIR Institute of
8   Genomics and Integrative Biology (CSIR-IGIB), Mall Road, Delhi 110007, India

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

10 **Abstract**

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].

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

Materials and Methods

Data Sets

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

Computational models for antitubercular activity

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

101   Briefly these models were based on two bioassays deposited in PubChem and carrying
102   IDs AID 1332 and AID 449762. Both the assays were based on microdilution Alamar
103   Blue assays. The former used 7H12 broth while the latter used 7H9 media. A total of
104   1,120 and 327, 669 compounds were screened in the respective assays. The models
105   were generated using a machine learning approach as described in Periwal et al and
106   Periwal et al [23, 24]. For the AID 1332 assay model was generated based on the
107   Random forest classification algorithm and was evaluated using a variety of statistical
108   measures which include accuracy, Balanced Classification Rate (BCR) and Area under
109   Curve (AUC). Balanced Classification Rate is an average of sensitivity and specificity
110   which introduces a balance in the classification rate. The model had an accuracy of
111   82.57%, BCR value of 82.2% and AUC value of 0.87. The AID 449762 assay model was
112   generated based on SMO (Sequential Minimization Optimization) algorithm and was
113   found to be 80.52 % accurate, with BCR value of 66.30% and AUC as 0.75.

114 In addition, we created an additional model to predict the molecules active against non-
115 replicating drug tolerant *Mycobacterium tuberculosis*. The assay was deposited in
116 PubChem with identifier AID 488890. A total of 3, 24, 437 compounds were screened for
117 the activity. The model was generated using Random forest classification algorithm as
118 described in the previous papers [23-26] and had an accuracy of 76%, BCR value
119 85.2% and AUC 0.66.

Molecular Descriptors

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

Formats and Format conversion

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

SMARTS filters

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

Mycobacterium tuberculosis permeability prediction

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

Data Mining

148 We used Weka, a popular and freely available Data Mining Software toolkit. Predictions
149 were performed for the dataset across the two models corresponding to assays AID
150 1332 and AID 449762 independently. Further, molecules predicted active in both the

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

Results

Summary of Datasets and Molecules
158 A total of 25,210 ingredients were downloaded from Traditional Chinese Medicines
159 Integrated Database (TCMID). We could retrieve molecular information for only 12,018
160 of the ingredients in the form of SMILE notations and the rest were not considered for
161 further analysis. The molecules considered along with their SMILES are detailed in
162 Supplementary Table 1. A total of 179 descriptors were calculated using PowerMV as
163 described above. The descriptors were further pruned for each of the models as
164 described in the Materials and Methods section using custom scripts in Perl. This
165 corresponds to 150 and 154 descriptors respectively for models AID 1332 and AID
166 449762 and 154 for AID 488890. The models, descriptors and scripts for formatting the
167 files are available at the Crowd Computing for Cheminformatics Model Repository
168 [http://vinodscaria.rnabiology.org/2C4C/models].

Prediction of potential anti-tubercular hits
170 The 12, 018 molecules obtained from TCMID were analyzed for the antitubercular
171 activity using the computational predictive models as described above. The AID 1332
172 and AID 449762 models predicted 2, 363 compounds and 5, 864 compounds
173 respectively as potentially active anti-tubercular. Of these molecules, a total of 1,472
174 molecules were predicted potential actives by both the models based on molecular
175 descriptors and were considered for further analysis (Supplementary Table 2).
176 Briefly we used a popular approach for filtering molecules with undesirable properties.
177 These included briefly using SMARTS filters. Molecules which passed the filtering step
178 were further evaluated for their effect against drug-tolerant and slow growing
179 Mycobacterium. Molecules were further evaluated for their potential permeability with
180 respect to the Mycobacterial cell wall.

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

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

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

206 **Molecules potentially active against non-replicating drug tolerant _Mycobacterium_**
207 **_tuberculosis_.**
208 A total of 160 compounds filtered through SMARTSfilter were tested using a
209 computational predictive model for potential activity against non-replicative
210 _Mycobacterium tuberculosis._ The model predicted 19 compounds as active to act as
211 potential inhibitors of non-replicating drug tolerant _Mycobacterium tuberculosis._ The
212 detailed description about 19 compounds is given in Table 1. The table also shows the
213 permeability probability of the molecules to pass through Mtb cell wall.

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

221 **Literature search suggests evidence of the sources and molecules used with**
222 **antitubercular properties**
223 We further searched for the role of the plant sources of the molecules in regard to their
224 use or known information on antibacterial or anti tubercular activities. We found several
225 molecules herbs to have antitubercular effects. These are _Petasites japonicus_ [36],

226 *Piper trichostachyon* [37], *Solanum torvum* [38], *Fritillaria przewalskii* [39], *Hernandia*
227 *sonora* 40] and *Phyllanthus urinari* [12]. In addition, many of the herbs have been shown
228 to have hepatoprotective activities, which include *Annona reticulata* [41, 42], *Annona*
229 *squamosa* [41, 42], and *Camellia sinensis* [43]. This offers a new opportunity for new
230 drug development considering that most of the established first-line drugs used in the
231 treatment of tuberculosis are hepatotoxic [44, 45]. We also found the molecules, Hinokiol
232 [46], Totarol [47], Murrayafoline a [48] and 2-hexenyl benzoate [49] have been known to
233 show antitubercular effects.

234 Discussion and Conclusions

235 Traditional Chinese Medicine (TCM) has been a major alternative medicine practice,
236 widely followed in many parts of China and Southeast Asia [1]. Enormous efforts in the
237 recent years have been invested in the systematic identification and characterization of
238 the molecular activities of the ingredients and scientific validation of their effects [10, 11].
239 The availability of well curated databases of ingredients of Traditional Chinese
240 Medicines has opened up new avenues for molecular screening as well as in-silico
241 studies, including target-based docking [6-9]. In depth screens of Chinese Medicine
242 derived compounds have been performed for a variety of pathophysiologies, including
243 cancer [50], inflammatory diseases [51, 52], cardiovascular diseases [53] and infections
244 [54] etc, just to name a few. These databases are being extensively used for therapeutic
245 development [55].

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

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

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

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

Competing interests
286 The authors declare that they have no competing interests.

Authors' contributions

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

References

299

300 1. Ooi GL: **Chinese medicine in Malaysia and Singapore: the business of healing.**
301   *Am J Chin Med.* 1993, **21(3-4):**197-212.

302 2. Qiu J: **Traditional medicine—a culture in the balance.** *Nature* 2007, **448:**126–128.

303 3. Normile D: **Asian medicine. The new face of traditional Chinese medicine.**
304   *Science* 2003, **299:** 188–190.

305 4. Wang MW, Hao X, Chen K: **Biological screening of natural products and drug**
306   **innovation in China.** *Philos. Trans. Roy. Soc. Lond. B Biol. Sci.* 2007, **362:**1093–
307   1105.

308 5. Sucher NJ: **Insights from molecular investigations of traditional Chinese herbal**
309   **stroke medicines: implications for neuroprotective epilepsy therapy.** *Epilepsy*
310   *Behav.* 2006, **8(2):**350-62.

311 6. Chen X, Zhou H, Liu YB, Wang JF, Li H, Ung CY, Han LY, Cao ZW, Chen YZ**:**
312   **Database of traditional Chinese medicine and its application to studies of**
313   **mechanism and to prescription validation.** *Br. J. Pharmacol.* 2006, **149:**1092–
314   1103.

315 7. Fang YC, Huang HC, Chen HH, Juan HF: **TCMGeneDIT: a database for**
316   **associated traditional Chinese medicine, gene and disease information using**
317   **text mining.** *BMC Complement. Altern. Med.* 2008, **8:**58.

318 8. Chen CYC: **TCM Database@Taiwan: the world's largest traditional Chinese**
319   **medicine database for drug screening in silico.** *Plos One* 2011, **6:**e15939.

320 9. Zhou J, Xie G, Yan X: *(eds) Encyclopedia of Traditional Chinese Medicines –*
321   *Molecular Structures, Pharmacological Activities, Natural Sources and Applications.*
322   Springer: New York; 2011.

323 10. Li XM, Brown L: **Efficacy and mechanisms of action of traditional Chinese**
324   **medicines for treating asthma and allergy.** *Journal of Allergy and Clinical*
325   *Immunology* 2009, **123(2):**297–306.

326 11. Wen Z, Wang Z, Wang S, Ravula R, Yang L, et al.: **Discovery of Molecular**
327   **Mechanisms of Traditional Chinese Medicinal Formula Si-Wu-Tang Using Gene**
328   **Expression Microarray and Connectivity Map.** *PLoS ONE* 2011, **6(3):**e18278.

329 12. Nair RR, Abraham RS: **Integrating the Science of Pharmacology and Bio**
330 **Informatics Phyllanthus "The wonder plant".** *Advanced Biotech January* 2008.
331 13. Nader LA, de Mattos AA, Picon PD, Bassanesi SL, De Mattos AZ, Pineiro Rodriguez
332 M: **Hepatotoxicity due to rifampicin, isoniazid and pyrazinamide in pateints**
333 **with tuberculosis: is anti-HCV a risk factor?** *Ann Hepatol.* 2010, **9(1):**70-74.
334 14.   World   Health   Organization   2012.
335 http://www.who.int/tb/publications/global_report/gtbr12_main.pdf
336 15. Shah NS, Wright A, Bai GH, Barrera L, Boulahbal F, et al.: **Worldwide emergence**
337 **of extensively drug-resistant tuberculosis.** *Emerg Infect Dis.* 2007, **13(3):**380-7.
338 16. DiMasi, JA, Hansen, RW, Grabowski, HG: **The price of innovation: new estimates**
339 **of drug development costs.** *Journal of Health Economics* 2003, **22:**151–185.
340 17. Vert JP, Jacob L: **Machine learning for in silico virtual screening and chemical**
341 **genomics: new strategies.** *Comb Chem High Throughput Screen* 2008, **11:**677-
342 685.
343 18. Melville JL, Burke EK, Hirst JD: **Machine Learning in Virtual Screening.** *Comb*
344 *Chem High Throughput Screen* 2009, **12:**332-343.
345 19. Schierz AC: **Virtual screening of bioassay data.** *J Cheminform* 2009, **1:**21.
346 20. Vasanthanathan P, Taboureau O, Oostenbrink C, Vermeulen NP, Olsen L, Jorgensen
347 FS: **Classification of cytochrome P450 1A2 inhibitors and non-inhibitors by**
348 **machine learning techniques.** *Drug Metab Dispos* 2009, **37:**658-664.
349 21. Li XJ, Kong DX, Zhang HY: **Chemoinformatics approaches for traditional**
350 **Chinese medicine research and case application in anticancer drug discovery.**
351 *Curr Drug Discov Technol.* 2010, **7(1):**22-31.
352 22. Zhang K, Li Y, Zhang ZR, Guan WH, Pu YC: **Chemoinformatics study on**
353 **antibacterial activity of traditional Chinese medicine compounds.** *Zhongguo*
354 *Zhong Yao Za Zhi.* 2013, **38(5):**777-80.
355 23. Periwal V, Rajappan JK, Jaleel AU, Scaria V: **Predictive models for anti-tubercular**
356 **molecules using machine learning on high-throughput biological screening**
357 **datasets.** *BMC Res Notes* 2011, **4:**504.
358 24. Periwal V, Kishtapuram S, Scaria V: **Computational models for in-vitro**
359 **antitubercular activity of molecules based on high-throughput chemical**
360 **biology screening datasets.** *BMC Pharmacol* 2012, **12:**1.

361 25. Jamal S, Periwal V, Scaria V: **Predictive modeling of anti-malarial molecules**
362   **inhibiting apicoplast formation.** *BMC Bioinformatics* 2013, **14:**55.

363 26. Jamal S, Periwal V, Scaria V: **Computational analysis and predictive modeling of**
364   **small molecule modulators of microRNA.** *Journal of Cheminformatics* 2012, **4:**16.

365 27. Ruichao Xue, Zhao Fang, Meixia Zhang, Zhenghui Yi, Chengping Wen, Tieliu Shi:
366   **TCMID: traditional Chinese medicine integrative database for herb molecular**
367   **mechanism analysis.** *Nucleic Acids Research* 2013, **41:**D1089–D1095.

368 28. Liu K, Feng J, Young SS: **PowerMV: a software environment for molecular**
369   **viewing, descriptor generation, data analysis and hit evaluation.** *J Chem Inf*
370   *Model* 2005, **45:**515-522.

371 29. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR: **Open**
372   **Babel: An open chemical toolbox.** *J Cheminform* 2011, **7(3):**33.

373 30. Bouckaert RR, Frank E, Hall MA, Holmes G, Pfahringer B, Reutemann P: **Weka**
374   **-Experiences with a Java Open-Source Project.** *Journal of Machine Learning*
375   *Research* 2010, 2533–2541.

376 31. Merget B, Zilian D, Müller T, Sotriffer CA: **MycPermCheck: The Mycobacterium**
377   **tuberculosis permeability prediction tool for small molecules.** *Bioinformatics*
378   2013, **29(1):** 62-68.

379 32. Singla D, Tewari R, Kumar A, Raghava GP: **Open Source Drug Discovery**
380   **Consortium. Designing of inhibitors against drug tolerant Mycobacterium**
381   **tuberculosis (H37Rv).** *Chem Cent J.* 2013, **7(1):** 49.

382 33. Baell JB, Holloway GA: **New Substructure Filters for Removal of Pan Assay**
383   **Interference Compounds (PAINS) from Screening Libraries and for their**
384   **Exclusion in Bioassays.** *J. Med. Chem* 2010, **53:**2719-2740.

385 34. Hann M, Hudson B, Lewell X, Lifely R, Miller L, Ramsden N: **Strategic pooling of**
386   **compounds for high-throughput screening.** *J Chem Inf Comput Sci.* 2009,
387   **39(5):**897-902.

388 35. Huth JR, Mendoza R, Olejniczak ET, Johnson RW, et al.: **ALARM NMR: A Rapid**
389   **and Robust Experimental Method to Detect Reactive False Positives in**
390   **Biochemical Screens.** *J. Am. Chem. Soc.* 2005, **127:**217-224.

391 36. John EF: **A Barefoot Doctors Manual: The American Translation of the Official**
392   **Chinese Paramedical Manual.** 1999.

393 37. Wagner H, Wolff P: **New Natural Products and Plant Drugs with**
394 **Pharmacological, Biological or Therapeutical Activity***.* Springer-Varlag Berlin;
395 1977:p212.

396 38. Silva KN, Silva RC, Coelho VPM, Agra M: **A pharmacobotanical study of**
397 **vegetative organs of Solanum torvum.** *Brazilian Journal of Pharmacognosy* 2011,
398 **21(4):**568-574.

399 39. Chang HM, Paul PH: *Pharmacology and Applications of Chinese Materia Medica*.
400 2001.

401 40. Udino L, Abaul J, Bourgeois P, Gorrichon L, Duran H, Zedde C: **Lignans from the**
402 **Seeds of Hernandia sonora.** *Planta Med.* 1999, **65(3):**279-81.

403 41. Thattakudian Sheik Uduman MS, Sundarapandian R, Muthumanikkam A, et al.:
404 **Protective effect of methanolic extract of Annona squamosa Linn in isoniazid**
405 **rifampicin induced hepatotoxicity in rats.** *Pak J Pharm Sci.* 2011, **24(2):**129-34.

406 42. Mohamed Saleem TS, Christina AJM, Chidambaranathan N, Ravi V:
407 **Hepatoprotective activity of Annona squamosa Linn. on experimental animal**
408 **model***.* *Int J Pharm Pharm Sci.* 2008, **1(3):**1-7.

409 43. Issabeagloo E, Taghizadieh M: **Hepatomodulatory Action of Camellia sinensis**
410 **Aqueous Extract against Isoniazid-Rifampicin Combination Induced Oxidative**
411 **Stress in Rat.** *Advances in Bioresearch* 2012, **3:**18-27.

412 44. Yew WW, Leung CC: **Antituberculosis drugs and hepatotoxicity.** *Respirology*
413 2006, **11(6):**699-707.

414 45. Liu Q, Garner P, Wang Y, Huang B, Smith H: **Drugs and herbs given to prevent**
415 **hepatotoxicity of tuberculosis therapy: systematic review of ingredients and**
416 **evaluation studies.** *BMC Public Health* 2008, **21(8):**365.

417 46. Chen JJ, Wu HM, Peng CF, Chen IS, Chu SD: **seco-Abietane diterpenoids, a**
418 **phenylethanoid derivative, and antitubercular constituents from Callicarpa**
419 **pilosissima.** *J Nat Prod.* 2009, **72(2):**223-8.

420 47. Jaiswal R, Beuria TK, Mohan R, Mahajan SK, Panda D: **Totarol inhibits bacterial**
421 **cytokinesis by perturbing the assembly dynamics of FtsZ.** *Biochemistry* 2007,
422 **46(14):**4211-20.

423 48. Choi TA, Czerwonka R, Fröhner W, Krahl MP, Reddy KR, Franzblau SG, Knölker HJ:
424 **Synthesis and activity of carbazole derivatives against Mycobacterium**
425 **tuberculosis.** *ChemMedChem* 2006, **1(8):**812-5.

426   49. He, Yantao, Li-fan ZENG, Zhong-Yin Zhang: **Tyrosine phosphatase inhibitors and**
427   **uses thereof to modulate the activity of enzymes involved in the pathology of**
428   **mycobacterium tuberculosis.** WIPO Patent Application PCT/US2012/035039.

429   50. Hu Y, Wang S, Wu X, Zhang J, Chen R, Chen M, Wang Y: **Chinese herbal**
430   **medicine-derived compounds for cancer therapy: A focus on hepatocellular**
431   **carcinoma.** *J Ethnopharmacol.* 2013**, S0378-8741(13)**00531-X.

432   51. Han C, Guo J: **Antibacterial and anti-inflammatory activity of traditional Chinese**
433   **herb pairs, Angelica sinensis and Sophora flavescens.** *Inflammation* 2012,
434   **35(3):**913-9.

435   52. Su SY, Hsieh CL: **Anti-inflammatory effects of Chinese medicinal herbs on**
436   **cerebral ischemia.** *Chin Med.* 2011, **6:**26.

437   53. Wang L, Qiu XM, Hao Q, Li DJ: **Anti-inflammatory effects of a Chinese herbal**
438   **medicine in atherosclerosis via estrogen receptor β mediating nitric oxide**
439   **production and NF-κB suppression in endothelial cells.** *Cell Death Dis.* 2013,
440   **4:**e551.

441   54. Jiang L, Deng L, Wu T: **Chinese medicinal herbs for influenza.** *Cochrane*
442   *Database Syst Rev.* 2013, **3:**CD004559.

443   55. Cheng HM, Li CC, Chen CY, Lo HY, Cheng WY, Lee CH, Yang SZ, Wu SL, Hsiang
444   CY, Ho TY: **Application of bioactivity database of Chinese herbal medicine on**
445   **the therapeutic prediction, drug development, and safety evaluation.** *J*
446   *Ethnopharmacol* 2010, **132(2):**429-37.

447   56. Lu J, Ye S, Qin R, Deng Y, Li CP: **Effect of Chinese herbal medicine extracts on**
448   **cell-mediated immunity in a rat model of tuberculosis induced by multiple**
449   **drug-resistant bacilli.** *Mol Med Rep.* 2013, **8(1):**227-32.

450   57. Li SZ: **Combined traditional Chinese medicine and western medicine in the**
451   **treatment of 356 cases of scrofula of the neck and axilla.** *Zhong Xi Yi Jie He Za*
452   *Zhi* 1984, **4(2):**90-2.
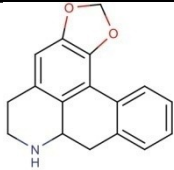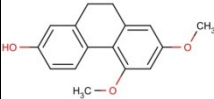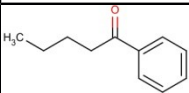
453   ## Tables and Figures
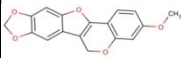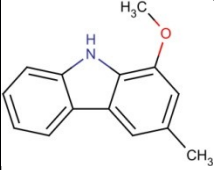
454   ## Figures
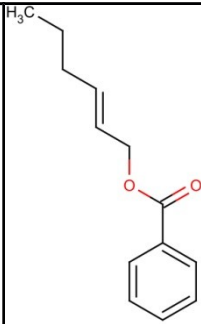
Figure 1: Summary of the data-mining and prioritization approach involving prediction of actives, consensus building and filtering for permeability and undesirable substructures.
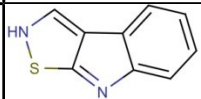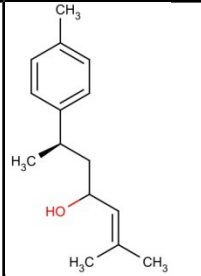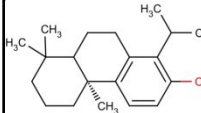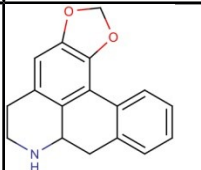
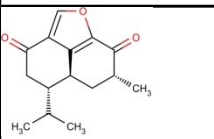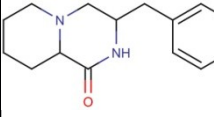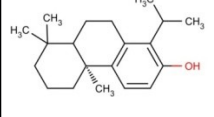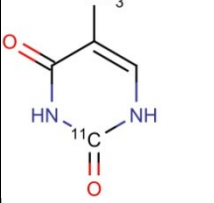**Figure 2:** Venn diagram showing active molecules filtered by any of the five SMARTS filters.
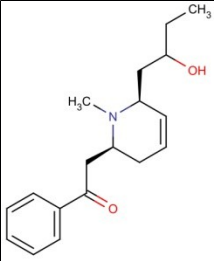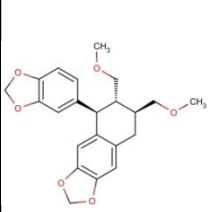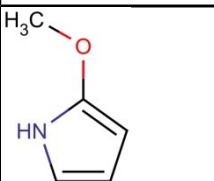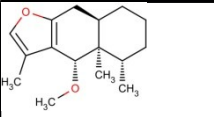
Tables
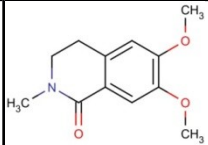
Table 1 shows the 19 compounds predicted as active against non replicating antibiotic tolerant *Mycobacterium tuberculosis*.

Comp ound No.	Compound structure	Name	English Name	Latin Name	Permeabili ty probability	Sources with antitubercul ar activities
1.		F lemichapparin b	Climbing Jewelvine	Derris scandens	0.993	
2.		Murrayafoline a	Taiwan Common Jasminorang e, Indian Common Jasminorang e, Euchretaleaf Common Jasminorang e, 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.		Cyclostachinea	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]

Supplementary Data

Supplementary Table 1 shows the Chinese molecules used in the present study with their smiles.

Supplementary Table 2 shows the molecules predicted to have anti tubercular activity by our models.

Supplementary Table 3 shows the 9 molecules which could penetrate the *Mycobacterium tuberculosis* cell wall.

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.
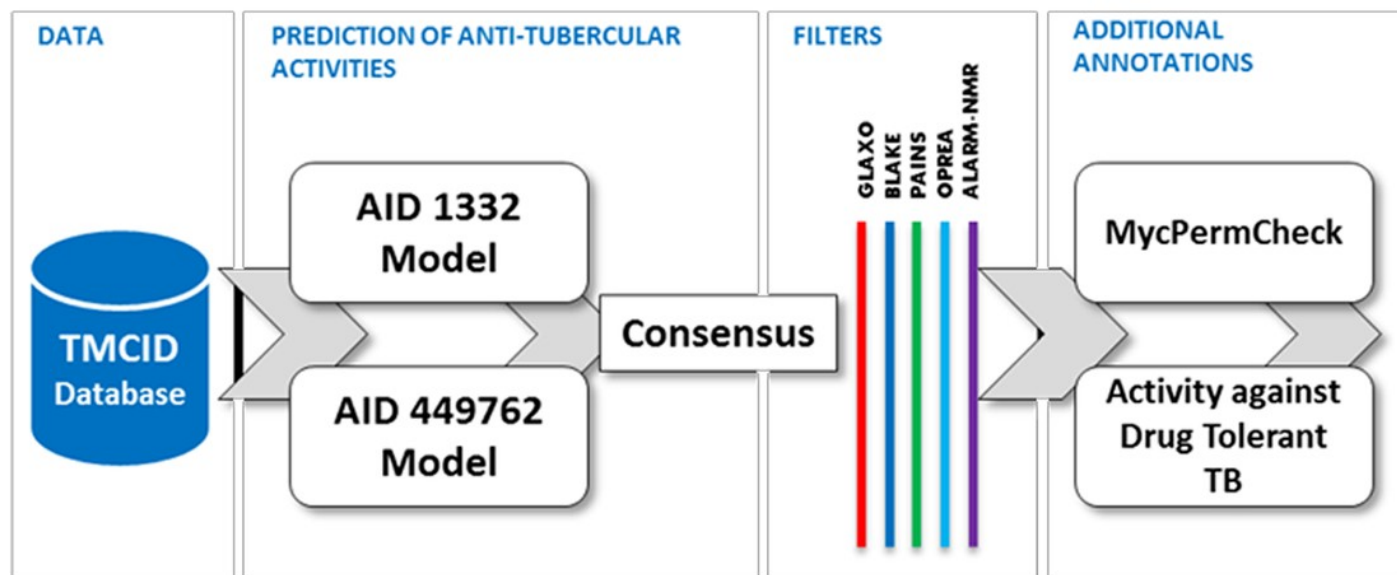
Figure 2

Figure 2

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