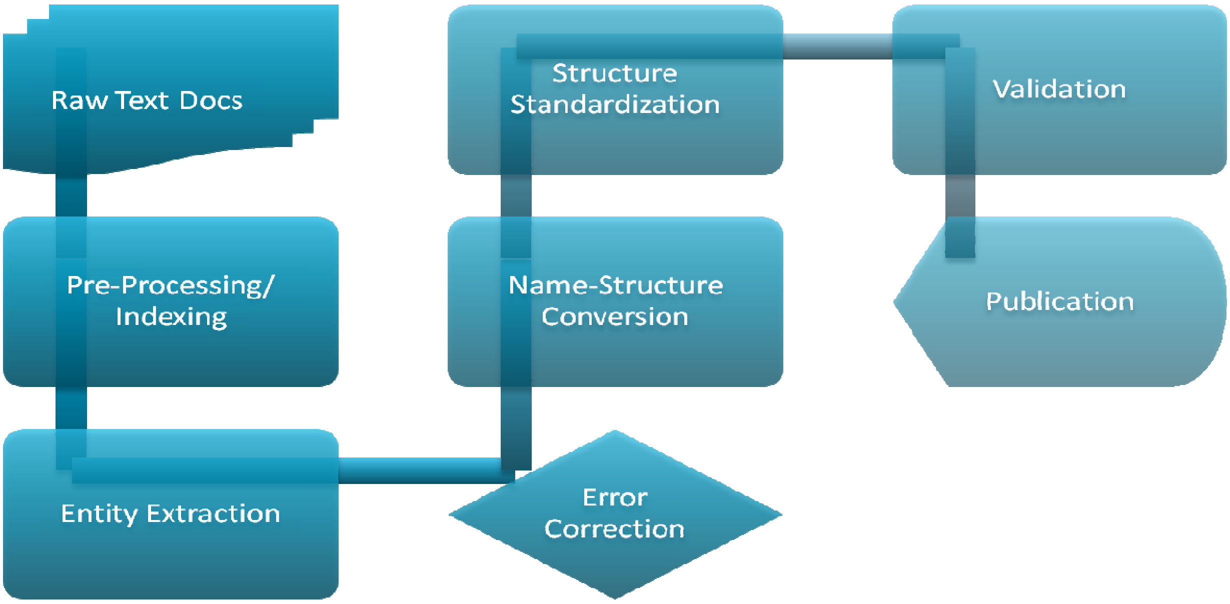
Supplemental Figure 1.

The process for the data generation workflow in SureChem for extracting compounds from patents. Tuberculosis small molecule patent data was obtained from SureChem (www.surechem.org, Macmillan Publishers Limited) which uses a database of 11 million unique chemical structures automatically extracted and annotated from world full-text patents and MEDLINE journal abstracts. Chemical structure data in SureChem is generated using proprietary machine-learning based entity extraction technology which identifies occurrences of chemical names in text.

These names then run through a series of correction heuristics and passed through numerous third-party name-to-structure conversion tools, which generate 2D chemical structures from the names SureChem has extracted. Names that fail to convert to structures are then run through another set of correction heuristics to increase the rate of name/structure conversion (Supplemental Figure 1).

The set of structures were selected as follows. Patents were identified using keyword queries related to tuberculosis, “ACLM: mycobacterium tuberculosis AND all:(inhibitors OR inhibition)” where 'ACLM' is the claims section of the patents and 'all' covers every field of the patent. Patents containing less than 60 unique chemicals were filtered out to eliminate those not relevant to drug discovery or small molecules. Chemicals were filtered out if they occurred more than 1000 times across the patent database and were then screened out to reduce the number of irrelevant compounds. Further work to remove fragments of chemical names and radicals was subsequently performed.

The resulting data set initially comprised 56,751 structures from a total of US, European and World Intellectual Property Organization (WIPO) 855 patent applications and granted patents. From those molecules, compounds like solvents and other low molecular weight regents as well as fragments of molecules were removed to clean up this list further to leave 20,775 molecules annotated with physicochemical calculated properties and links out directly to the corresponding US-PTO, EPO, and WO patents.



Supplemental Methods File 1

Accelrys SMARTS filters

The following SMARTS were used with default settings: Sulfonyl halide (0-1), Primary alkyl halide (0-1), Epoxide or aziridine (0-1), Sulfonate ester (0-1), Phosphonate ester (0-1), Long aliphatic chain (0-1), Peroxide (0-1), 1-2 Dicarbonyl (0-1), Acid halide (0-1), Non-Hydrogen atoms (2-35), Carbons (1-30), N-O-S (0-9),Sulfonyl halides (0-0), Acid halides (0-0), Alkyl halides (0-0), Acid anhydrides (0-0), Isocyanates or Isothiocyanates (0-0), Thiocyanates (0-0), Carbodiimides (0-0), Sulfonates (0-0), Acylhydrazides (0-0), Isonitriles (0-1), Imines (0-0), Acrylonitriles (0-0), Propenals (0-0), Macrocycles (0-0), Phosphorus 3 (0-0), Hexanes (0-0), 5 rotatable bonds (0-0), Aliphatic alcohols (0-3), Perchlorates (0-0), Fluorines (0-7), Cl-Br-I (0-3), P halides (0-0), Cyanohydrines (0-0), Sulfate esters (0-0), Pentafluorophenyl esters (0-0), Paranitrophenyl esters (0-0), HOBt esters (0-0), Lawessons reagents (0-0), Phosphoramides (0-0), Aromatic azides (0-0), Quaternary C-Cl-I-P-S (0-0), Beta carbonyl quaternary N (0-0), Acyl cyanides (0-0), Sulfonyl cyanides (0-0), Thioepoxides (0-0), Benzylic quaternary N (0-0), Di or Triphosphates (0-0), Aminooxy-oxo (0-0), Nitros (0-1), N-halides (0-0), Aldehyde (0-1), Cyano (0-1), Acid halides (0-0), Carbazides (0-0), Sulfate esters (0-0), Sulfonates (0-0), Acid anhydrides (0-0), Peroxides (0-0), Pentafluorophenyl esters (0-0), Paranitrophenyl esters (0-0), Esters of HOBT (0-0), Isocyanates and Isothiocyanates (0-0), Triflates (0-0), Lawesson reagent and derivatives (0-0), Phosphoramides (0-0), Aromatic azides (0-0), Beta carbonyl quaternary Nitrogen (0-0), Acylhydrazide (0-0), Quaternary C or C1 or I or P or S (0-0), Phosphoranes (0-0), Nitroso (0-0), P or S Halides

(0-0), Carbodiimide (0-0), Isonitrile (0-0), Triacyloximes (0-0), Cyanohydrins (0-0), Acyl cyanides (0-0), Sulfonyl cyanides (0-0), Cyanophosphonates (0-0), Azocyanamides (0-0), Azoalkanals (0-0), Aliphatic methylene chains 7 or more long (0-0), Compounds with 4 or more acidic groups (0-0), Crown ethers (0-0), Disulfides (0-0), Thiols (0-0), Epoxides or Thioepoxides or Aziridines (0-0), 2-4-5 trihydroxyphenyl (0-0), 2-3-4 trihydroxyphenyl (0-0), Hydrazothiourea (0-0), Thiocyanate (0-0), Benzylic quaternary Nitrogen (0-0), Thioesters (0-0), Cyanamides (0-0), Four numbered Lactones (0-0), Di and Triphosphates (0-0), Betalactams (0-0), Quinones (0-0), Polyenes (0-0), Saponin derivatives (0-0), Cytochalasin derivatives (0-0), Cycloheximide derivatives (0-0), Monensin derivatives (0-0), Cyanidin derivatives (0-0) and Squalestatin derivatives (0-0).