Materials and Methods

Screening database. We used the ZINC8 predefined subset of drug-like\textsuperscript{52} molecules downloaded on 14\textsuperscript{th} April 2009\textsuperscript{53}. This consists of 8,784,580 commercially available compounds with associated vendor information. Each molecule was provided with a single protonation state calculated at pH=7 and a single low-energy 3D conformation. Further details on the protocols used to build this database can be found in the original publication\textsuperscript{20}.

Shape similarity screen. Ultrafast Shape Recognition (USR) was used for the shape similarity screen (see ref.\textsuperscript{22} for a detailed description of USR, ref.\textsuperscript{54} for a critical assessment of small-molecule shape similarity techniques, ref.\textsuperscript{55,22} for retrospective validations and, most importantly, refs.\textsuperscript{23,39} for prospective validations of USR). Here we performed three USR runs, each using a different co-crystallised DHQase ligand as the search template (CA2 from PDB code 2BT4, RP4 from 2CJF and GAJ from 2C4W), which led to 2963, 918 and 498 similarly shaped molecules, respectively (4,379 molecules or unique ZINC codes, with no molecules in common among the three sets of USR hits). USR performed these three database searches, i.e. a combined 26.4 million shape similarity calculations with similarity threshold set to 0.9, in just 90 seconds using a single processing core of a laptop computer (including both reading descriptors from file and screening time). InstantJchem software\textsuperscript{56} was used to retrieve the 3D conformers corresponding to USR hits from the screening database.

Toxicity screen. The \textit{in silico} toxicity model aimed at discarding those molecules predicted to be both mutagenic and carcinogenic. In order to maximise specificity with respect to both endpoints simultaneously, we built a consensus model using Derek for Windows (version 11.0)\textsuperscript{57,58} and ToxTree (version 1.51)\textsuperscript{59}, with compounds
only identified as mutagenic/carcinogenic if predicted by both methods. Derek for Windows (DfW) was interpreted as predicting mutagenicity or carcinogenicity when in vitro mutagenicity in *Salmonella typhimurium* or carcinogenicity in humans was reported as “certain”, “probable” or “plausible”. ToxTree (TT) was interpreted as predicting mutagenicity when a genotoxic structural alert was “fired”, with carcinogenicity also predicted in these cases and when its inbuilt QSAR for the carcinogenicity of aromatic amines in rodents returned a positive prediction, with negative predictions being ignored.

**Docking.** The resulting 3908 molecules were docked into the five PDB structures of type II dehydroquinase 2BT4, 2C4W, 2CJF, 1GU1 and 1H0R using GOLD version 4.1.1. The X-ray structures have a resolution of 1.55 to 2.1 Å and are for *S. coelicolor* DHQase (PDB codes 1GU1, 2BT4 and 2CJF), *H. pylori* DHQase (2C4W) and *M. tuberculosis* DHQase (1H0R). In each structure, the five water molecules closest to the catalytic residue Tyr28 were kept as active water molecules, with the remaining structural waters being removed (the active water molecules were set to 'toggle', meaning that GOLD decides whether these are bound or displaced along the docking run). The centre of the binding site was defined as the centroid between the side chain O of Tyr28 and Nδ of His106, with search space boundaries (10 Å from this centre) including all the active site residues (Pro15, Asn16, Arg23, Tyr28, Asn79, Ala82, Glu104, His106 and Arg113). The residues were assigned their default protonation state as in aqueous solution, with all the histidines represented by the Nε single-protonated tautomer except for His106 which was protonated on both side chain nitrogens. The side chains were kept rigid during the run, with a maximum of 10 different poses being generated for each docked molecule. Early termination was scheduled in case the top three poses reached an RMSD of less than 1.5 Å. The genetic
algorithm was set to automatic with a search efficiency of 100%. The docking poses were generated with ChemScore. All other docking options were in default settings. Pose generation took approximately four days per target using a single processing core of a 3.40GHz 2GB RAM workstation. As expected, this process resulted in a similar number of docking poses against each target: 21,788 (1GU1), 28,191 (1H0R), 21,771 (2BT4), 24,814 (2C4W) and 20,976 (2CJF). In addition to ChemScore values arising from the pose generation process, each of these five sets of docking poses were re-scored with GoldScore and ASP as well as an early and less predictive version of RF-Score that differs with the published version in that uses a 6Å cutoff and just six out of the nine atom types (all but those involving Cl, Br and I).

**Virtual screening protocols.** We investigated four different protocols. Protocol 1 is a consensus scoring strategy that considered the three sets of docking poses containing the three largest co-crystallised ligands (2BT4, 2CJF and 2C4W sets) and each set was sorted with the average rank of the pose according to ChemScore, GoldScore and ASP. By contrast, Protocol 2 used RF-Score alone to rank these three sets of docking poses. We restricted ourselves to three docking sets because we wanted to investigate the rank of each co-crystallised ligand with respect to the corresponding docking poses for both protocols. Consequently, the other two sets of docking poses were not considered because the co-crystallised ligand in 1GU1 and 1H0R (FA1) was deemed too small to be likely to have a competitive potency and thus rank high. Protocol 3 considered all five sets of docking sets (2BT4, 2C4W, 2CJF, 1GU1 and 1H0R). For each set, poses were sorted according to the average rank from ChemScore, GoldScore and ASP ranked lists. Thereafter, we identified 61 molecules that ranked in the top 500 molecules against all five targets. This strategy was chosen to increase the likelihood of obtaining compounds inhibiting all three versions of the enzyme and
thus have a broad spectrum of activity. Lastly, Protocol 4 simply consisted of considering those molecules similarly shaped to the RP4 crystal pose, so as to evaluate how protocols using more chemical information perform with respect to ligand-based shape similarity only.

**Purchasing compounds for testing.** The compounds for experimental laboratory testing were selected through three different processes. First, we extracted the vendor information for the top 100 compounds in each target for both protocol 1 and 2. Since each vendor catalogue only comprises a part of the ZINC database, we selected the three vendors with the best coverage of our hits (Ambinter, ChemDiv and Enamine). That is, we did not select every vendor to reduce shipping costs as a way to make the most of our modest budget (£5,000). Once ordered, some of these compounds turned out to be unavailable. As a result of these practical constraints, only 136 of these top ranked molecules were finally purchased for testing (69 from protocol 1 and 67 from protocol 2). Second, 24 of the resulting 61 structures resulting from protocol 3 overlapped with these 136 compounds identified in the first selection process. An analogous purchase process was followed which resulted in the acquisition of eight of these 37 non-overlapping compounds from protocol 3. Third, we ended up with five top ranking molecules from protocol 4 to be used as a baseline for performance. The overall process resulted in 148 candidates for biological testing (there are two compounds in common between protocols 1 and 2, and another compound in common between protocols 1 and 3).

**Synthesis of substrate 3-Dehydroquinate.** Potassium 3-dehydroquinate was synthesised from (-)-quinic acid using the method of Le Sann et al.\textsuperscript{63} Calibration of aqueous 3-dehydroquinate solutions was determined from the absorbance difference
at 234 nm resulting from the total conversion of an aliquot of 3-dehydroquinate to 3-dehydroshikimate by 1 µL of *S. coelicolor* DHQase (5.1 mg mL$^{-1}$) using the kinetic assay below.

**Kinetic assay for DHQases.** Enzyme stocks were diluted to 120 nM (*S. coelicolor* DHQase) and 1800 nM (*M. tuberculosis* type II DHQase) with 50 mM Tris.HCl buffer pH 7.0. The enzymes were assayed by monitoring the increase in absorbance at 234 nm due to the enone-carboxylate chromophore of 3-dehydroshikimate ($\varepsilon = 1.2 \times 10^4$ M$^{-1}$ cm$^{-1}$). The assays were performed at 25 °C in a 96-well plate, reading 12 wells per run, each well every 9s for 3 minutes. The assay was initiated by addition of enzyme (10 µL) to assay mixture to give a final volume of 200 µL. Inhibitors added as 20× DMSO stock solutions. **Inhibition assay conditions.** *M. tuberculosis* DHQase $IC_{50}$: 90 nM enzyme, 50 mM Tris.HCl pH 7.0, 50 µM potassium 3-dehydroquinate, 5% DMSO, 2.5-500 µM or 10-1000 µM inhibitor (12 concentrations). *S. coelicolor* DHQase $IC_{50}$: 6 nM enzyme, 50 mM Tris.HCl pH 7.0, 150 µM potassium 3-dehydroquinate, 5% DMSO, 2.5-500 µM or 10-1000 µM inhibitor (12 concentrations). *S. coelicolor* DHQase $K_i$: 6 nM enzyme, 50 mM Tris.HCl pH 7.0, 50-600 µM potassium 3-dehydroquinate (5 concentrations), 5% DMSO, 0-300 µM inhibitor (8 concentrations). All experiments performed in duplicate. $IC_{50}$ and $K_i$ values were determined using GraFit$^{64}$ software.

**Protein preparation.** Over-expression and purification of His6-AroQ (DHQase from *Mycobacterium tuberculosis* and *Streptomyces coelicolor*): 10 mL of an overnight culture (20 mL) of E. coli C41 (DE3) competent cells transformed with either plasmid pET-28a/aroQ (*M. tuberculosis*) or pET-15a/aroQ (*S. coelicolor*) was used to inoculate 1L of 2 x YT medium supplemented with either kanamycin 30 µg mL$^{-1}$
(pET-28a vector) or ampicillin 100 μg mL⁻¹ (pET-15a vector) and divided into two 500 mL cultures. The cultures were incubated at 37 °C, 220 rpm until the optical density at 600 nm reached 0.6. Protein over-expression was induced by addition of isopropyl-β-thiogalactopyranoside (IPTG) to a final concentration of 1 mM. The cultures were incubated at 37 °C, 220 rpm for 4.5 hours, and cells harvested by centrifugation, 11,000 x g, 15 min, 4 °C. Cell pellets stored at -20 °C. Harvested cells were suspended in 30 mL buffer A (50 mM sodium phosphate, 300 mM NaCl, 10 mM imidazole, 1 mM β-mercaptoethanol, pH8.0) containing lysozyme from chicken egg white (0.1 mg mL⁻¹), DNase I from bovine pancreas - type II (0.1 mg mL⁻¹) (both Sigma-Aldrich), and 1 protease inhibitor cocktail tablet (Roche). The suspension was homogenized by sonication on ice (2 seconds on, 6 seconds off, 5 min). The crude extract was centrifuged at 25,000 x g, 30 min, 4 °C. The resulting supernatant was filtered (0.2 μM syringe filter - Millipore) and loaded onto a 7 x 1.5cm nickel-nitrilotriacetic acid (Ni-NTA) column (Qiagen Ni-NTA superflo; Bio-Rad econo-column 15 x 1.5 cm) previously equilibrated with buffer A (50 mL). After washing with buffer A (20 mL), 20 mM imidazole in buffer A (50 mL), and 50 mM imidazole in buffer A (50 mL), the protein was eluted with 250 mM imidazole in buffer A (50 mL) collecting 1 x 4.5 mL, then 7 x 6.5 mL fractions. The elution was followed via sodium dodecyl sulphate-polyacrylamide gel electrophoresis (SDS-PAGE). The protein containing fractions were pooled and concentrated using a Vivaspin 20 (10,000 molecular weight cut-off (MWCO)) centrifugal concentrator, 4,500 rpm, 4°C, to a volume of 2.5 mL. The concentrated 3-dehydroquinases were buffer exchanged into either 100 mM Tris.HCl pH 7.8 (M. tuberculosis) or 50 mM Tris.HCl pH 7.5, 0.4 mM dithiothreitol, 50% glycerol (S. coelicolor) using a PD-10 column containing Sephadex G25 resin (GE Healthcare). Aliquots (250 μL) of the 3-dehydroquinase
from *S. coelicolor* were flash frozen in liquid nitrogen and stored at -20 °C. The 3-dehydroquinase from *M. tuberculosis* was concentrated in a Vivaspin 6 (10,000 MWCO) centrifugal concentrator, 4,500 rpm, 4 °C to ~1.6 mL, and aliquots (50 µL) flash frozen in liquid nitrogen and stored at -80 °C. Enzyme concentration was determined by measuring the absorbance at 280 nm and conversion from the calculated extinction coefficient value determined by the SwissProt - ProtParam function (0.623 equals 1 mg mL\(^{-1}\) for *M. tuberculosis* and 0.608 equals 1 mg mL\(^{-1}\) for *S. coelicolor* 3-dehydroquinase).

**IC\(50\) to calculated Ki.** In order to compare the newly discovered inhibitors with those previously known, we started by evaluating the Cheng-Prusoff equation\(^{65}\) with \(K_m = 150\) µM and substrate concentration \([S] = 150\) µM for *S. coelicolor* leading to the conversion expression \(K_i = 0.50\ IC_{50}\), which is valid for competitive inhibition\(^{66}\). To test the latter assumption, we used the measured \(K_i\) of six of the most potent inhibitors for *S. coelicolor* DHQase and performed a linear regression with the corresponding IC\(50\) values (Pearson correlation coefficient 0.82). The resulting expression (\(K_i = 0.43\ IC_{50} + 0.92\)) is fairly similar to the theoretical expression and hence supports the assumption of competitive inhibition for these new active molecules. Likewise, we also evaluated Cheng-Prusoff equation with \(K_m = 45\) µM and substrate concentration \([S] = 50\) µM for *M. tuberculosis* leading to the conversion expression \(K_i = 0.47\ IC_{50}\).

**Scaffold diversity analysis.** The diversity of DHQase inhibitors was analysed as follows. First, all the inhibitors in the ChEMBL database for *M. tuberculosis* DHQase (Target ID 20064; 14 previously known inhibitors compiled from references\(^{18,67}\)) and *S. coelicolor* DHQase (Target ID 18038; 20 previously known inhibitors compiled from references\(^{18,68}\)) were retrieved. Second, all the IC\(50\) values of our new inhibitors
were converted into K_i using the semi-empirical expression from the previous section and merged with the previously known inhibitors (this was done separately for each target). Third, the chemical structure of all these molecules was encoded using the CDK^69 1024 bit standard (path-based hashed) fingerprint as implemented in the R package rcdk^70 and all pairwise Tanimoto scores between the resulting fingerprints were calculated. Next, hierarchical agglomerative complete clustering was carried out on the resulting similarity matrix. We found that a similarity cutoff of 0.6 provided a set of clusters that correlated with the number and composition of new chemical series arising from our virtual screen. This led to several clusters of molecules with similar structures, providing a rough identification of new chemical series arising from the virtual screen. After manual refinement, a final list of new molecular scaffolds was obtained.

**Structural filters to flag promiscuous hits.** The structural filters proposed by Baell and Holloway^71 were used to flag hits that might demonstrate a broad promiscuous activity (these compounds were termed PAINS for Pan-Assay INterference compoundS). We are using the most up-to-date version we have found of these filters (Petrova et al.^72), which is based on the 18 most reliable substructures based on the information about more than 100 types of promiscuous compounds that have been proven problematic in a number of HTS campaigns. This substructure filter set is represented by the 18 SMARTS strings and is fully disclosed in the supplementary information of Petrova et al. It is important to notice that the compound filtering may also eliminate useful molecules from further consideration. Indeed, it is enough for a member in a chemical series to be found promiscuous for all other members to be also considered promiscuous. For example, a significant fraction of known drug molecules does not pass through these substructure filters. The Drugbank database
(http://www.drugbank.ca/) contains structural data for 1408 FDA-approved small molecule drugs. Analysis of the Drugbank by Petrova et al. showed that 1.78% of these FDA-approved are discarded by the PAINS substructure filter. When running these conservative filters on the 100 DHQase hits reported in this study, only three of these (ZINC12436746, ZINC00316138 and ZINC00978022) were flagged as being potentially non-specific or promiscuous compounds.

**Detergent controls.** Detergent controls performed using the methods described for Inhibition Assay Conditions using Triton X-100 as the detergent. Triton X-100 was added to stock buffer solutions prior to the sequential addition of potassium-3-dehydroquininate (aqueous solution), inhibitors (DMSO solution), and initiation of reaction by addition of enzyme such that the final concentration of detergent was 0.01% (0.01% to 2% for range finding experiments). Results of these controls support the specificity of the most potent hits as outlined in the Electronic Supplementary Materials (ESM3 - Detergent controls results.pdf).


56 InstantJchem (version 5.3.8) (Chemaxon, Budapest, 1037 Hungary) at http://www.chemaxon.com/

57 Derek for Windows (version 11.0) (Lhasa Limited, 22-23 Blenheim Terrace, Woodhouse Lane, Leeds, LS2 9HD) at https://www.lhasalimited.org/derek/


60 http://www.ebi.ac.uk/pdbe/ (last accessed on January 31, 2012)


64 Leatherbarrow RJ, GraFit Version 5.0.6. Erithacus Software Ltd, Horley, UK.

65 Cheng, Y. & Prusoff, W.H. 1973 Relationship between the inhibition constant (Kᵢ) and the concentration of inhibitor which causes 50 per cent inhibition (I₅₀) of an enzymatic reaction. Biochem Pharmacol 22, 3099-3108.


