JOURNAL OF THE ROYAL SOCIETY Interface

*J. R. Soc. Interface* (2012) **9**, 3196–3207
doi:10.1098/rsif.2012.0569
*Published online* 29 August 2012

# Hierarchical virtual screening for the discovery of new molecular scaffolds in antibacterial hit identification

**Pedro J. Ballester[1,*,†], Martina Mangold[2,†], Nigel I. Howard[2], Richard L. Marchese Robinson[2], Chris Abell[2], Jochen Blumberger[3] and John B. O. Mitchell[4]**

[1]*European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK*
[2]*Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK*
[3]*Department of Physics and Astronomy, University College London, London WC1E 6BT, UK*
[4]*Biomedical Sciences Research Complex and EaStCHEM School of Chemistry, University of St Andrews, North Haugh, St Andrews, Fife KY16 9ST, UK*

One of the initial steps of modern drug discovery is the identification of small organic molecules able to inhibit a target macromolecule of therapeutic interest. A small proportion of these hits are further developed into lead compounds, which in turn may ultimately lead to a marketed drug. A commonly used screening protocol used for this task is high-throughput screening (HTS). However, the performance of HTS against antibacterial targets has generally been unsatisfactory, with high costs and low rates of hit identification. Here, we present a novel computational methodology that is able to identify a high proportion of structurally diverse inhibitors by searching unusually large molecular databases in a time-, cost- and resource-efficient manner. This virtual screening methodology was tested prospectively on two versions of an antibacterial target (type II dehydroquinase from *Mycobacterium tuberculosis* and *Streptomyces coelicolor*), for which HTS has not provided satisfactory results and consequently practically all known inhibitors are derivatives of the same core scaffold. Overall, our protocols identified 100 new inhibitors, with calculated $K_i$ ranging from 4 to 250 μM (confirmed hit rates are 60% and 62% against each version of the target). Most importantly, over 50 new active molecular scaffolds were discovered that underscore the benefits that a wide application of prospectively validated *in silico* screening tools is likely to bring to antibacterial hit identification.

**Keywords: virtual screening; antibacterial hit identification; chemoinformatics; bioinformatics; machine learning; high-throughput screening**

## 1. INTRODUCTION

In recent years, bacterial epidemics have been fuelled by the emergence of multi-drug-resistant pathogen strains, which increasingly challenge existing treatments [1]. Despite this growing threat, many new antibiotic candidates are chemical molecules re-engineered from old drug classes discovered decades ago for which there are already underlying resistance mechanisms [1–4]. Fischbach & Walsh [5] have argued that, while making incremental improvements to existing scaffolds is a good short-term strategy for refilling the antibiotic pipeline, the discovery of new molecular scaffolds should be a priority owing to the emergence of multi-drug-resistance among pathogens and the need for a sustainable plan for combating resistance. Simmons *et al.* [6] go even further by pointing out that the discovery of drugs with novel modes of action will be vital to meet the threats created by the emergence of resistance [7]. With the deciphering of the genome sequences of major human pathogens [8], many companies vigorously pursued the identification of novel antibiotic agents from high-throughput screening (HTS) campaigns using purified enzyme targets that were validated by genomic approaches as being essential for the viability of the pathogen [6]. The expected outcome was the production of brand new inhibitor classes against these novel targets that could eventually lead to innovative drugs and hence reduced likelihood of resistance emerging rapidly. Furthermore, finding a high number of inhibitor classes was an important requirement, as multiple high-quality hits are needed to counteract the attrition in the subsequent

*Author for correspondence (pedro.ballester@ebi.ac.uk).
†These authors contributed equally to this study.

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2012.0569 or via http://rsif.royalsocietypublishing.org.

*Received* 17 July 2012
*Accepted* 3 August 2012

hit-to-lead, lead optimization and clinical stages required to generate a novel antibiotic [3].

Unfortunately, the performance of HTS against post-genomic antibacterial targets has generally been unsatisfactory. Payne *et al.* [3] critically assessed the results of what these authors describe as an unprecedented concentration of screening resource for a single therapy area. From the 70 antibacterial HTS campaigns run between 1995 and 2001 (67 target-based and three whole cell), it was found that a mere 16 HTS gave rise to hits and only five of these ultimately resulted in leads (i.e. molecules that not only inhibited the enzyme target, but also reduced the growth of the pathogen *in vitro*). On the basis of GlaxoSmithKline (GSK) metrics, the success rate from antibacterial HTS was four- to fivefold lower than for targets from other therapeutic areas available at that time. These authors concluded that this was a disappointing and financially unsustainable outcome, especially in view of the length of time devoted to these experiments and considering that the costs per HTS campaign were estimated then to be around US$1 million. Payne *et al.* [3] also argued that the difficulty of finding antibacterial hits from HTS was not unique to GSK, a view further supported by Simmons *et al.* [6], who consider that success in discovering inhibitors, using HTS of chemical libraries is rare in this area.

The poor efficacy of HTS in this area has been attributed to the limited chemical diversity of the screened collections [3] and to the assumption that antibacterial targets have intrinsically low druggability [9]. Remarkably, well-known technical difficulties such as false negatives in HTS have not been investigated as a contributor to poor efficacy despite false negatives probably occurring frequently [10], representing typically an estimated 15–26% of the total number of actives [11] and having a particularly large impact on hard targets [12]. Most importantly, the cost and slow operation of HTS imposes the selection of a relatively small set of promising compounds that in an extreme case may not contain inhibitors. Hence, while the false-positive rate of HTS will generally be much better than that from a virtual screening technique applied to the same compound library, the fact that virtual screening can quickly search a much larger portion of the chemical space means that the latter has, in principle, access to many more inhibitors and consequently may find more inhibitors in some cases. As a result of this situation, lead discovery has generally become a key bottleneck for the development of new treatments for infectious diseases [13].

A consensus is now emerging that new approaches for finding antibacterial inhibitors are required. Fischbach & Walsh [5] consider that retooled target-based strategies can play an important role in lead discovery. Hopkins *et al.* [9] find that new, more cost-effective and efficient methods of drug discovery are urgently required if we are to tackle the multiple global health challenges of emerging and neglected infectious diseases for which there is relatively little basic science investment. Recently, Simmons *et al.* [6] have made a compelling case for the use of structure-based virtual screening for antibacterial hit identification. Indeed, the application of advanced computational methods to predict molecular bioactivity has distinctive advantages such as much reduced time scales and financial costs that enable the effective exploration of extremely large molecular databases against a high number of validated drug targets.

In this study, we investigate a novel computational methodology that exploits the structures of antibacterial targets in order to identify brand new classes of inhibitors. We combine state-of-the-art docking protocols hierarchically with ultrafast shape recognition (USR) capable of quickly identifying database molecules that are similarly shaped to a known inhibitor. In this way, only the molecules that can fit the target's active site, and hence are likely to bind to the enzyme, are subsequently fed to the much more computationally demanding docking calculation. This hybrid approach permits the effective exploration of truly large and diverse molecular databases in a time- and resource-efficient manner while exploiting both ligand and protein structure data, as a way to increase the likelihood of identifying new active scaffolds that could be unaffected by existing resistance mechanisms. The methodology is tested prospectively on two versions of a post-genomic antibacterial target (type II dehydroquinase from *Mycobacterium tuberculosis* and *Streptomyces coelicolor*), which is representative of this situation in that HTS has not provided satisfactory results and consequently little scaffold diversity is currently known. Testing any resulting hit for whole-cell antibacterial activity would be necessary to determine which of these enzyme inhibitors are able to reach the intracellular target with sufficient concentration, which implies crossing the bacterial membrane and being unaffected by resistance mechanisms such as efflux pumps. Such an extensive follow-up work is however out of the scope of this study.

## 2. RESULTS AND DISCUSSION

The enzyme 3-dehydroquinate dehydratase (dehydroquinase; EC 4.2.1.10) catalyses the reversible dehydration of 3-dehydroquinate to form 3-dehydroshikimate [14–16]. Type II dehydroquinase (DHQase for short) is the third enzyme of the Shikimate pathway, which is essential for the viability of bacteria such as *M. tuberculosis*, *S. coelicolor* and *Helicobacter pylori* [17]. This pathway is present in bacteria, fungi, plants and apicomplexan parasites, but not in mammals, and hence represents an ideal target for the development of antibacterial agents, as these agents would be expected to have a spectrum of antibacterial activity restricted to those human pathogens expressing DHQase such as *M. tuberculosis* and *H. pylori*. An HTS of some 150 000 compounds against *H. pylori* DHQase was used as a starting point to identify novel inhibitors [18]. While approximately 100 molecules with more than 50 per cent inhibition of *H. pylori* DHQase enzyme activity at a concentration of 20 $\mu g\ ml^{-1}$ were identified in the primary screening, only one confirmed inhibitor against *H. pylori* DHQase was reported (the ligand named GAJ in figure 1, which inhibited this enzyme with $K_i = 20\ \mu M$). This molecule also showed micromolar activity against

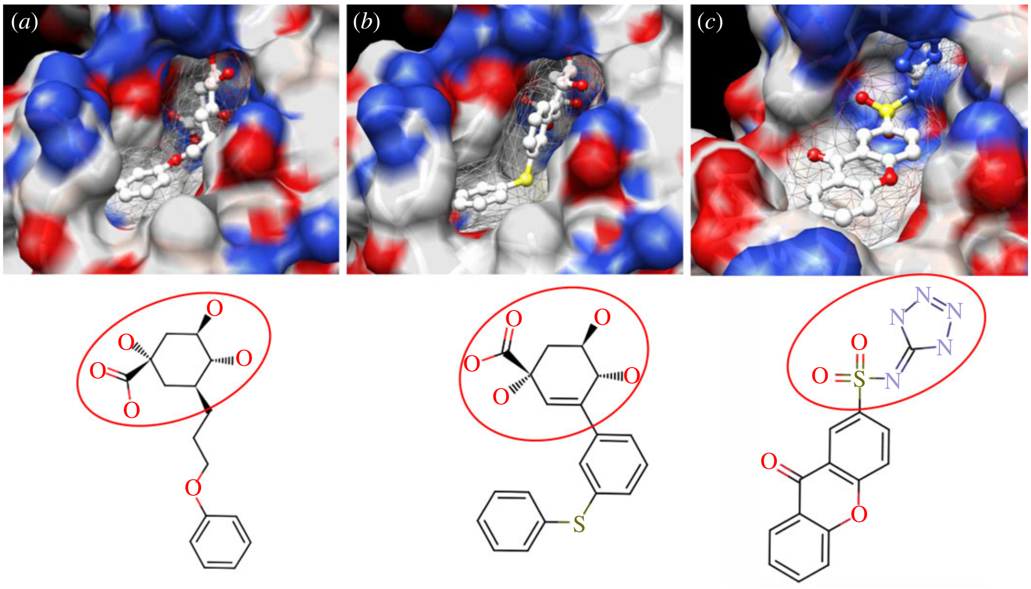


Figure 1. Visualization of the three co-crystallized ligands used as templates for the shape similarity screen ((*a*) CA2 complexed with *S. coelicolor* DHQase; (*b*) RP4 complexed with *S. coelicolor* DHQase; (*c*) GAJ complexed with *H. pylori* DHQase). The van der Waals surface of each bound molecule is represented as a grid to show the high degree of shape complementarity between the ligands and their receptors. The core scaffold, defined as that closest to the catalytic residues, is circled. CA2 and RP4 are derivatives of the transition state structure (core scaffold 2,3-anhydroquinic acid which is also the crystallographic ligand FA1), whereas the innovative structure of GAJ was identified with HTS [18].

*S. coelicolor*[1] DHQase ($K_\mathrm{i} = 230\ \mu\mathrm{M}$) but only residual activity against the *M. tuberculosis* enzyme (10% inhibition at 200 µM). The ChEMBL database (https://www.ebi.ac.uk/chembl/ last accessed on 31 January 2012), which has been estimated [9] to contain 90 per cent of the published medicinal chemistry structure–activity data, shows that practically all existing DHQase inhibitors are derivatives of the same core scaffold (2,3-anhydroquinic acid or anhydroquinate ring, the reaction intermediate), consistent with the successful use of rational drug design approaches and the typically low performance of HTS on antibacterial targets. Figure 1 shows the chemical structures of these active scaffolds as well as the high degree of shape complementarity between these molecules and their respective receptors.

Our search for new classes of DHQase inhibitors was carried out on a molecular database built from the ZINC resource [19]. With almost nine million commercially available molecules, its size is between 17 and 59 times higher than those previously used for large-scale HTS campaigns (from 150 000 to 530 000 compounds [3,18]) and, to the best of our knowledge, the largest that has ever been used in a successful prospective virtual screen. Such a wealth of chemical diversity is a key component of our screen, as a smaller database generated with the same procedure would have contained a lower number of innovative scaffolds. In order to compile a subset of molecules likely to fit the active site, we searched for molecules that are similarly shaped to known inhibitors using USR [20]. USR is an unusually rapid descriptor-based shape similarity technique [21], which is particularly suited for scaffold hopping and has already been successfully applied to the identification of brand new active scaffolds within very large molecular databases [22]. It is well known that using several molecules as search templates results in a broader exploration of different regions of chemical space and thus we ran USR using each of the DHQase ligands shown in figure 1 as templates (CA2 from PDB entry 2BT4, RP4 from 2CJF and GAJ from 2C4W). This process resulted in the identification of 4379 diverse molecules that are similar in shape to these inhibitors, and thus fit the DHQase active site, from the nine million molecules initially considered.

These similarly shaped molecules were thereafter inspected by an *in silico* model intended to remove those predicted to be toxic. Toxicity screens during the early stages of drug discovery should prioritize specificity (the proportion of correctly identified non-toxins) over sensitivity (the proportion of correctly identified toxins), in order to reduce the likelihood of erroneously discarding good hits at a stage when attrition costs are low [23]. Indeed, our aim is to remove obviously toxic molecules rather than performing the exhaustive toxicity analysis that would be required at later stages. Thus, our model removed compounds predicted to be both carcinogenic and mutagenic, which are recognized as serious forms of toxicity in the context of drug discovery [24]. By discarding compounds predicted to exhibit both types of toxicity, we sought to reduce the loss of non-toxins as a result of imperfect predictions of each individual type. This screen removed 471 molecules from the 4379 molecules analysed.

The resulting 3908 molecules were docked into a panel of DHQase protein structures. We selected five ligand-bound DHQase X-ray crystal structures to be used for docking: three structures of the *S. coelicolor* version of the enzyme (PDB codes 1GU1, 2BT4 and

[1] *Streptomyces coelicolor* is a close non-pathogenic relative of *M. tuberculosis*, and thus is often used as an *in vivo* model for the characterisation of compounds that show promise *in vitro*.

Table 1. Performance of each virtual screening protocol against *M. tuberculosis* DHQase. For each protocol, the number of tested compounds, confirmed inhibitors with $\mathrm{IC}_{50} \leq 250\ \mu\mathrm{M}$ (hit rate between brackets), confirmed inhibitors with $\mathrm{IC}_{50} \leq 500\ \mu\mathrm{M}$ (hit rate between brackets) and indication of potency distribution of these discovered inhibitors (L, lowest $\mathrm{IC}_{50}$; M, median $\mathrm{IC}_{50}$; H, highest $\mathrm{IC}_{50}$).

| virtual screening protocol | tested | $\mathrm{IC}_{50} \leq 250\ \mu\mathrm{M}$ | $\mathrm{IC}_{50} \leq 500\ \mu\mathrm{M}$ | (L, M, H) ($\mu$M) |
|---|---|---|---|---|
| 1. USR-3>GOLD::ChemScore>GOLD-3 | 71 | 20 (28.2%) | 40 (56.3%) | (48, 245, 475) |
| 2. USR-3>GOLD::ChemScore>RF-Score | 67 | 25 (37.3%) | 38 (56.7%) | (89, 223, 426) |
| 3. USR-3>GOLD::ChemScore>Top500-5 | 8 | 2 (25.0%) | 5 (62.5%) | (94, 268, 417) |
| 4. USR-RP4 | 5 | 0 (0%) | 5 (100%) | (320, 354, 409) |
| overall performance | 148 | 47 (31.8%) | 88 (59.5%) | (48, 243, 475) |

2CJF), one structure for *H. pylori* DHQase (2C4W) and the fifth structure for *M. tuberculosis* DHQase (1H0R). The purpose of considering more than one X-ray structure is twofold. First, because each structure represents at least a slightly different conformation of the protein in the crystal, we will be addressing the flexibility of the protein to some extent. Second, because the structures refer to three closely related versions of the same enzyme coming from three different organisms (two pathogenic bacteria and a commonly used non-pathogenic model organism), these data will enable the search for molecules potentially able to kill all three organisms. 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). Pose generation quality was investigated by re-docking the three largest co-crystallized ligands (CA2 from 2BT4, RP4 from 2CJF and GAJ from 2C4W) back to their respective receptors, with root-mean-square deviation (r.m.s.d.) values between co-crystallized and re-docked poses being 0.59 (2CJF), 0.79 (2BT4) and 2.28 (2C4W). The r.m.s.d. over 2 Å indicates that there may be significant errors in pose generation, and thus a more exhaustive search should lead to improved performance. In addition to ChemScore [25] values arising from the pose generation process, each of these five sets of docking poses were re-scored with GoldScore [26] and ASP [27]. Protocol 1 is a consensus scoring [28] strategy that considered the three sets of docking poses containing the three largest co-crystallized ligands (CA2 from PDB code 2BT4, RP4 from 2CJF and GAJ from 2C4W), and each set was sorted with the average rank of the pose according to ChemScore, GoldScore and ASP (this consensus score will be henceforth referred to as GOLD-3). High ranking poses by three different scoring functions represent by construction a more reliable prediction than any of the constituent scoring functions alone. In practice, consensus scoring has been generally found to improve virtual screening performance dramatically with respect to the individual scoring functions [28,29].

The RF-Score [30] is a member of a new class of scoring functions that use non-parametric machine learning to build predictive models of binding affinity in an entirely data-driven manner. RF-Score has been rigorously shown [30,31] to perform better than 16 standard scoring functions in ranking protein–ligand complexes according to predicted binding affinity. Protocol 2 used RF-Score alone to re-score and rank the same three sets of docking poses as in protocol 1. The reason for restricting our study to three of the five docking sets was that we wanted to determine the rank of each co-crystallized ligand according to both protocols, and considered that the ligand in 1GU1 and 1H0R (FA1) was too small to have a competitive potency and thus rank high against the predominantly larger docked molecules. Protocol 1 ranked ligands 2nd (GAJ), 5th (RP4) and 19th (CA2) with respect to the 3908 molecules docked against its corresponding co-crystallized protein (2C4W, 2CJF and 2BT4, respectively), whereas protocol 2 ranked these ligands much lower in each list: 2679th (GAJ), 2865th (RP4) and 3878th (GAJ). Interestingly, the binding affinity prediction of these inhibitors by RF-Score is particularly accurate for a scoring function: $\mathrm{p}K_{\mathrm{i}}^{\text{RF-Score}}(\text{2BT4}) = 6.30$ (+1.82 with respect to measured $K_{\mathrm{i}} = 33\ \mu\mathrm{M}$ [32]), $\mathrm{p}K_{\mathrm{i}}^{\text{RF-Score}}(\text{2CJF}) = 7.35$ (+0.20 with respect to measured $K_{\mathrm{i}} = 70$ nM [33]) and $\mathrm{p}K_{\mathrm{i}}^{\text{RF-Score}}(\text{2C4W}) = 6.28$ (+1.58 with respect to measured $K_{\mathrm{i}} = 20\ \mu\mathrm{M}$ [18])[2]. In §3, these seemingly conflicting retrospective results will be discussed in the light of the prospective performance of both protocols. Protocol 3 identified all molecules that ranked in the top 500 against all five targets to encourage the discovery of broad-spectrum inhibitors. Lastly, protocol 4 simply consisted of searching for the most similar molecules to the RP4 ligand using USR, so as to investigate the advantages of additionally exploiting protein structure as in protocols 1–3.

The next step was to purchase the compounds highlighted by these protocols to test them *in vitro* against *M. tuberculosis* DHQase and *S. coelicolor* DHQase. With the modest budget assigned to this proof of concept (£5000), we could purchase 148 compounds (full details on this process can be found in electronic supplementary material, Materials and methods). Tables 1 and 2 show the performance of each virtual screening protocol against each target (all $\mathrm{IC}_{50}$ measurements are included in the electronic supplementary material). Protocol 2 performed better than protocol 1 and much better than protocol 3, both in terms of hit rate at $\mathrm{IC}_{50} \leq 250\ \mu\mathrm{M}$ and median $\mathrm{IC}_{50}$. As protocol 3 was the only strategy exploiting the *M. tuberculosis* DHQase structure while using the same consensus scoring as protocol 1, the difference in performance suggests that focusing on the compounds at the top of the ranked list is more important

[2] Note that the same validation is not possible with any of the three GOLD scores, as these provide dimensionless units.

Table 2. Performance of each virtual screening protocol against *S. coelicolor* DHQase. For each protocol, the number of tested compounds, confirmed inhibitors with $\mathrm{IC}_{50} \leq 250\ \mu\mathrm{M}$ (hit rate between brackets), confirmed inhibitors with $\mathrm{IC}_{50} \leq 500\ \mu\mathrm{M}$ (hit rate between brackets) and indication of potency distribution of these discovered inhibitors (L: lowest $\mathrm{IC}_{50}$, M: median $\mathrm{IC}_{50}$, H: highest $\mathrm{IC}_{50}$).

| virtual screening protocol | tested | $\mathrm{IC}_{50} \leq 250\ \mu\mathrm{M}$ | $\mathrm{IC}_{50} \leq 500\ \mu\mathrm{M}$ | (L, M, H) (μM) |
|---|---|---|---|---|
| 1. USR-3>GOLD::ChemScore>GOLD-3 | 71 | 24 (33.8%) | 38 (53.5%) | (57, 208, 490) |
| 2. USR-3>GOLD::ChemScore>RF-Score | 67 | 32 (47.8%) | 43 (64.2%) | (8, 203, 478) |
| 3. USR-3>GOLD::ChemScore>Top 500-5 | 8 | 2 (25.0%) | 6 (75.0%) | (177, 321, 496) |
| 4. USR-RP4 | 5 | 0 (0%) | 4 (80.0%) | (295, 322, 401) |
| overall performance | 148 | 58 (39.2%) | 91 (61.5%) | (8, 215, 496) |

than using their exact crystal structure. Lastly, while protocol 4 did not identify any inhibitor with $\mathrm{IC}_{50} \leq 250\ \mu\mathrm{M}$, it obtained the best performance in terms of hit rate at a higher cut-off ($\mathrm{IC}_{50} \leq 500\ \mu\mathrm{M}$).

In order to assess the overall hit rate and potency of these inhibitors, we calculated their inhibition constant ($K_\mathrm{i}$) as explained in electronic supplementary material, Materials and methods (see table 3 for a summary). This is necessary to have a more accurate comparison with hit rates in the literature and with the potency of previously known inhibitors. Our protocols identified 89 inhibitors for *M. tuberculosis* DHQase and 91 inhibitors for *S. coelicolor* DHQase with $K_\mathrm{i} \leq 250\ \mu\mathrm{M}$ (a total of 100 new inhibitors with activity against at least one of the targets). Among the rest of the tested compounds, many showed a small percentage of inhibition at a higher concentration. Overall, hit rates are unusually high, with the confirmed hit rate at the low micromolar range for *S. coelicolor* DHQase (27.0% molecules with $K_\mathrm{i} \leq 100\ \mu\mathrm{M}$; a total of 40 inhibitors) being noticeably higher than that for *M. tuberculosis* DHQase (23.6%; 35 inhibitors). The trend is still observed when considering a less restrictive activity cut-off ($K_\mathrm{i} \leq 250\ \mu\mathrm{M}$) and in terms of median potency (114 μM for *M. tuberculosis* versus 108 μM for *S. coelicolor*). This difference might be due to the fact that our protocols are primarily exploiting crystal structures for *S. coelicolor* DHQase. Figure 2 shows examples of these new inhibitors, which are characteristic of the high chemical diversity of the new core scaffolds. Table 4 shows the shape and chemical structure of these five new inhibitors in comparison with the search template used (RP4).

As discussed in §1, the single most important requirement for new inhibitors is having different core scaffolds from those previously known. For our case study, we retrieved all inhibitors with measured $K_\mathrm{i}$ from the ChEMBL database (ChEMBL target IDs 18038 and 20064 for *S. coelicolor* and *M. tuberculosis* DHQase, respectively) and observed that all but two of these ChEMBL inhibitors have FA1 (a transition state analogue shown in figure 1) as the core scaffold. In order to investigate how different our new inhibitors are compared with those previously known, we cluster all these molecules in terms of chemical structure similarity (figures 3 and 4). The clustering dendrogram evidences the striking diversity of new inhibitors. In figure 3 (*M. tuberculosis*), the first cluster at the top groups all the ChEMBL inhibitors that have the same core scaffold. By using the clustering as a preliminary classification, we manually inspected the clusters and found 48 new core scaffolds (all inhibitors sharing the same core scaffold constitute a chemical series by definition). Representatives from these chemical series are pictured on the left of the figure. The results for figure 4 (*S. coelicolor*) are very similar except that one of the ChEMBL inhibitors has a very different chemical structure (chembl438436, which is actually GAJ; the HTS hit in figure 1) and thus appears separated from the FA1 derivatives that are clustered at the top. Five additional chemical series were found for this target, which adds to a total of 53 new molecular scaffolds contained in these 100 DHQase inhibitors.

Table 3. Overall virtual screening performance in terms of calculated $K_\mathrm{i}$ (148 compounds tested against each version of the enzyme). For each target, confirmed inhibitors with $K_\mathrm{i} \leq 100\ \mu\mathrm{M}$ (hit rate between brackets), confirmed inhibitors with $K_\mathrm{i} \leq 250\ \mu\mathrm{M}$ (hit rate between brackets) and the $K_\mathrm{i}$ of the three most potent inhibitors found ($\mathrm{L}^1$–$\mathrm{L}^3$).

| overall performance | $K_\mathrm{i} \leq 100\ \mu\mathrm{M}$ | $K_\mathrm{i} \leq 250\ \mu\mathrm{M}$ | ($\mathrm{L}^1$, $\mathrm{L}^2$, $\mathrm{L}^3$) (μM) |
|---|---|---|---|
| against Mtb DHQase | 35 (23.6%) | 89 (60.1%) | (23, 24, 40) |
| against Scl DHQase | 40 (27.0%) | 91 (61.5%) | (4, 21, 29) |

## 3. DISCUSSION AND CONCLUSIONS

Four virtual screening protocols have been used in this prospective study. This approach was intended to investigate the relative performance of ligand-based and structure-based techniques on the same target and screening database (these parallel prospective applications constitute the most rigorous type of virtual screening method validation, but have not been carried out until now, as discussed in a recent comprehensive survey [34]). Protocols 1 and 2 differ only in that each uses a different re-scoring function to rank docking poses according to predicted strength of binding (GOLD-3 and RF-Score, respectively). Unlike protocol 2, protocol 1 was able to rank the three crystal poses very highly. On the basis of these retrospective results, one might conclude that GOLD-3 is a much more appropriate re-scoring procedure for this target. However, prospectively, RF-Score led to the identification

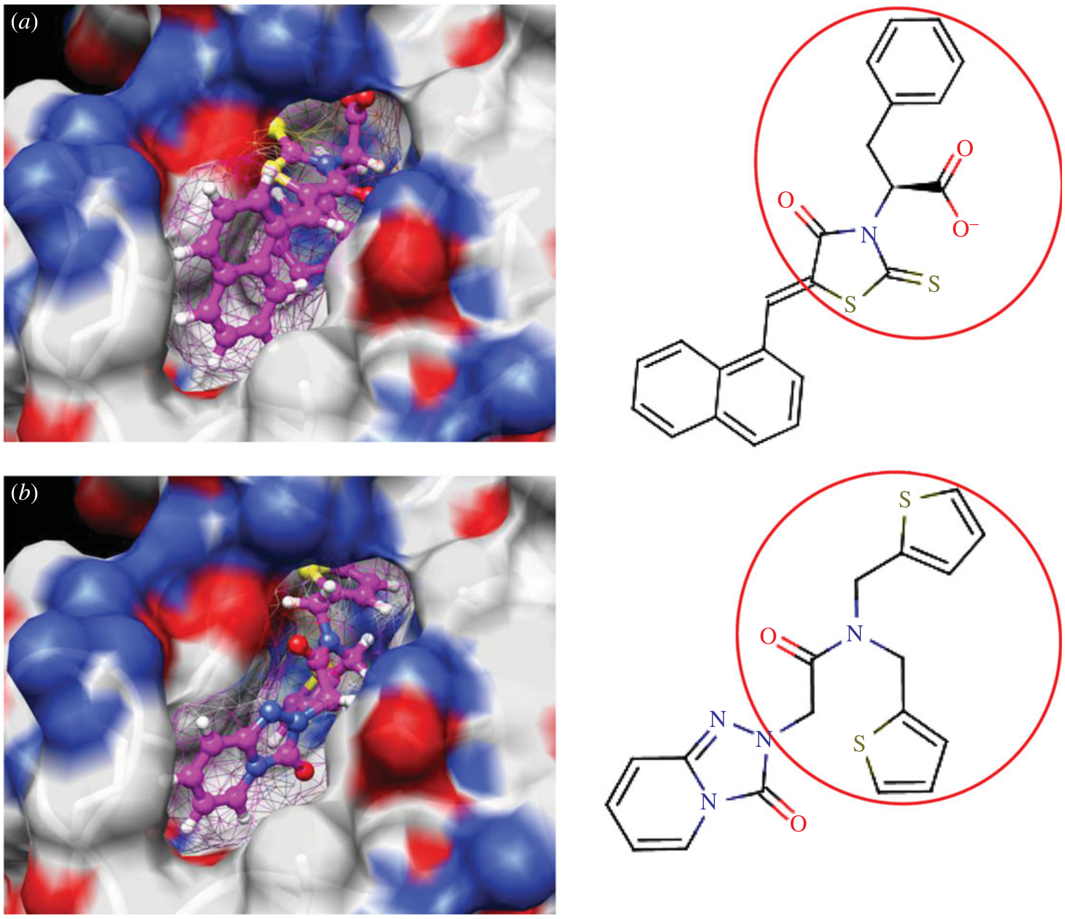


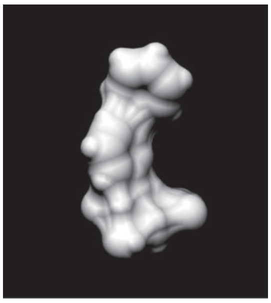
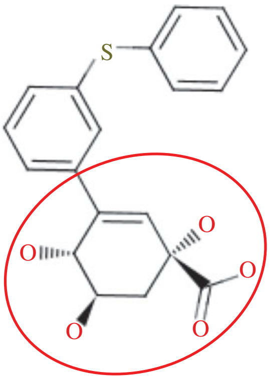
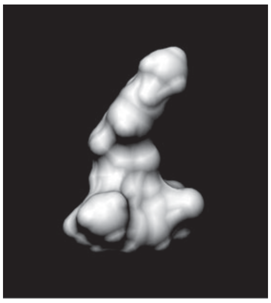
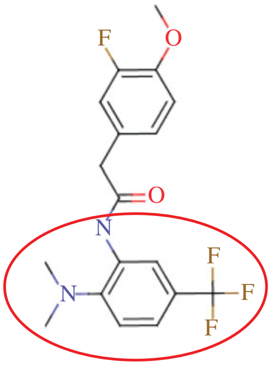
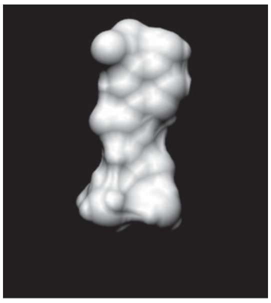
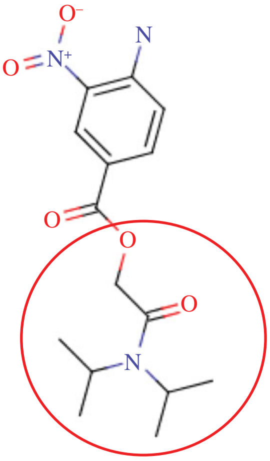
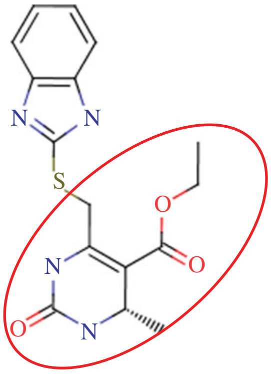
Figure 2. Docking pose and chemical structure of two new inhibitors of *S. coelicolor* DHQase identified with our virtual screening protocols ((*a*) ZINC00978022 with $K_\mathrm{i} = 4\ \mu\mathrm{M}$; (*b*) ZINC24469052 with $K_\mathrm{i} = 21\ \mu\mathrm{M}$). The new core scaffolds, circled in the structure, are strikingly different from those previously known (figure 1).

of an outstanding proportion of new inhibitors, and performed even better than GOLD-3 both in terms of hit rate and median potency. Interestingly, RF-Score was ranking different, previously unknown inhibitors higher than the known inhibitors, with no overlap with the inhibitors identified by GOLD-3. This is a reminder of an important limitation of retrospective studies: many of the highly ranked assumed inactives used as decoys could actually be active (especially if their potency is at the level typical of hit identification, not that of optimized leads or even drugs, and decoys are chosen to have physico-chemical properties similar to those of known actives).

Of course, not all the performance can be attributed to the re-scoring functions, as the initial shape similarity screen was used to enrich the few thousand docked molecules with actives. For instance, GOLD-3 ranks the low nanomolar RP4 ligand fifth of the 3908 similarly shaped molecules. However, a simple arithmetic rule of three tells us that this molecule would be expected to be ranked 11 239th if GOLD::ChemScore > GOLD-3 had been applied directly to the entire database of 8 784 580 molecules. As pointed out by Schneider [35], the particularly successful application of USR alone to prospective virtual screening in the past [22] underscores the utility of coarse-grained models for first-pass compound selection. The present application constitutes another example. In order to investigate what exactly was gained by exploiting structural information via docking, we look at the results of protocol 4 in table 4 (tested hits from the USR-RP4 screen). Compared with protocols 1 and 2, which perform docking on the subset preselected by USR (tables 1 and 2), protocol 4 (USR alone) obtained a significantly higher hit rate at the cost of providing a much higher median $K_\mathrm{i}$ (lower potency). These results show that USR leads to a smaller proportion of false positives, which suggests that shape similarity alone is more reliable than pose generation and scoring in docking, at least for this target. By contrast, the addition of estimated binding affinity makes the true positives from docking more potent than those from USR alone. Motivated by these results and just like in lead optimization [36], we propose hit identification to be regarded as a multi-objective problem where screening outcomes (potency, hit rate, chemical diversity, speed and cost of operation) are optimized by searching for the best protocol (methods, templates, structures, database composition and size). This will help us to establish which virtual screening techniques are more appropriate for the requirements of a particular hit identification project.

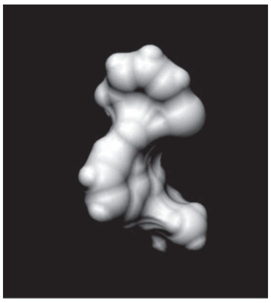
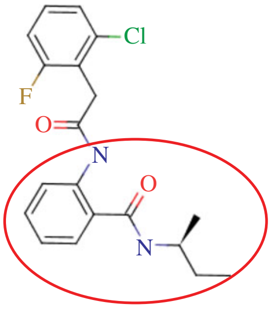
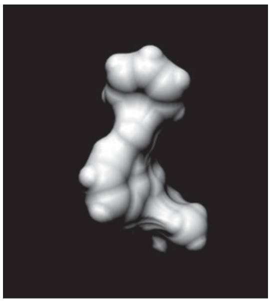
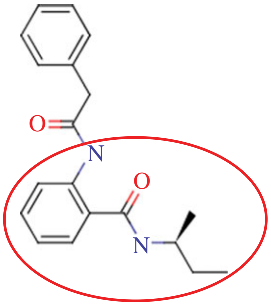
Beyond methodological findings, the most important contribution of this study is the outstanding diversity in new molecular scaffolds found (a total of 53 new chemical series at a nominal cost of £5000). This is particularly valuable, taking into account that, with

Table 4. Results for the USR screen of the nine-million single-conformer database using the co-crystallized pose of the highly optimized RP4 as the template, which was the highest ranked molecule as expected. We selected five top-ranking compounds on the basis of their prompt availability from the supplier and tested them *in vitro* against both *M. tuberculosis* DHQase and *S. coelicolor* DHQase. All these molecules showed mid-micromolar activity against both targets, but most importantly new core scaffolds to be used as starting points for optimization were identified. Each core scaffold is circled as predicted by the corresponding docking pose against *S. coelicolor* DHQase (2CJF).

| ZINC ID | rank | USR score | shape | two-dimensional structure | $K_i$ (Mtb; μM) | $K_i$ (Sc; μM) |
|---|---|---|---|---|---|---|
| RP4 | first | 1.000 |  |  | 0.38 | 0.07 |
| ZINC18086350 | fifth | 0.9404 |  |  | 166 | 45% at 224 |
| ZINC05657290 | 13th | 0.9356 |  |  | 192 | 201 |
| ZINC06024496 | 15th | 0.9341 | |  | 159 | 174 |

(*Continued.*)

Table 4. (*Continued.*)

| ZINC ID | rank | USR score | shape | two-dimensional structure | $K_i$ (Mtb; μM) | $K_i$ (Sc; μM) |
|---|---|---|---|---|---|---|
| ZINC15608616 | 24th | 0.9319 |  |  | 150 | 149 |
| ZINC10025084 | 26th | 0.9315 |  |  | 167 | 148 |

the exception of GAJ and ChEMBL340769 (*S. coelicolor*), all previously known inhibitors are derivatives of the transition state. Overall, we have found 100 new DHQase inhibitors comprising a total of 180 bioactivity endpoints against both targets. These results contrast with those from an HTS of 150 000 compounds against the *H. pylori* version of the enzyme, which reported a single confirmed inhibitor with $K_i = 20$ μM [18]. In calculating hit rates, we have adopted the same minimum definition of a hit as in industrial HTS campaigns for antibacterial hit identification[3]. As noticed in a comprehensive survey of prospective virtual screening applications [34], low- and mid-micromolar binders against novel targets are commonly reported in leading journals if there is limited ligand information available. This is because after discovery of a new scaffold, its derivatives can be tested to improve potency for that target. Although this task can be challenging, large potency improvements have been achieved in antibacterial targets such as *M. tuberculosis* adenosine 5′-phosphosulfate reductase [37] (10-fold) and *S. coelicolor* DHQase [33] (3000-fold). The latter figure suggest that in principle, a DHQase hit with $K_i = 100$ μM (the upper threshold of our definition for low micromolar) could potentially reach 30 nM once optimized.

Several factors have contributed to this level of performance. First, the use of ligand-based shape similarity to enrich the docking library with likely binders has been shown to be instrumental. Three USR searches, each using a different instance of shape complementarity as template, account to some extent for partial shape complementarity between ligand and its receptor as well as shape variation in the binding site due to induced fit effects. Second, while the notion that a certain degree of shape complementarity is necessary for binding has been recognized and implemented in drug design tools for decades, it is only now that the efficiency, effectiveness and widespread availability of such tools are making a large impact in hit identification [22,38–40]. Third, the exploration of such a large molecular database, effortlessly enabled by the use of USR, means that we can quickly search a much larger region of chemical space than previously possible. Fourth, the application of RF-Score, the first scoring function based on non-parametric machine learning, has been shown to result in more potent inhibitors than shape similarity alone, while maintaining an excellent hit rate. It is also encouraging that this initial version of RF-Score is already very competitive not only with established scoring functions, but with the consensus formed by all of them. These results are consistent with the observed superior performance in estimating binding affinity of diverse protein–ligand complexes using RF-Score [30,31] and more recently by other machine-learning-based scoring functions [41–43].

In practice, virtual screening methodologies are limited by the quality and availability of relevant experimental data (e.g. known active molecules and X-ray structures) as well as the degree to which their underlying assumptions align with the properties of the screened molecules and target protein (e.g. the scoring function performs well for that target). In the context of our study, there are a number of modifications that are likely to result in improved virtual screening protocols. First, research into more suitable intermolecular interaction descriptors and the exploitation of higher volumes of high-quality structural and interaction data should result in improved RF-Score predictability, as

[3] Hit means a chemically tractable, low micromolar inhibitor of the target and, where appropriate, at least 10-fold selectivity against the mammalian version of the target (none in the case of DHQase) [3].

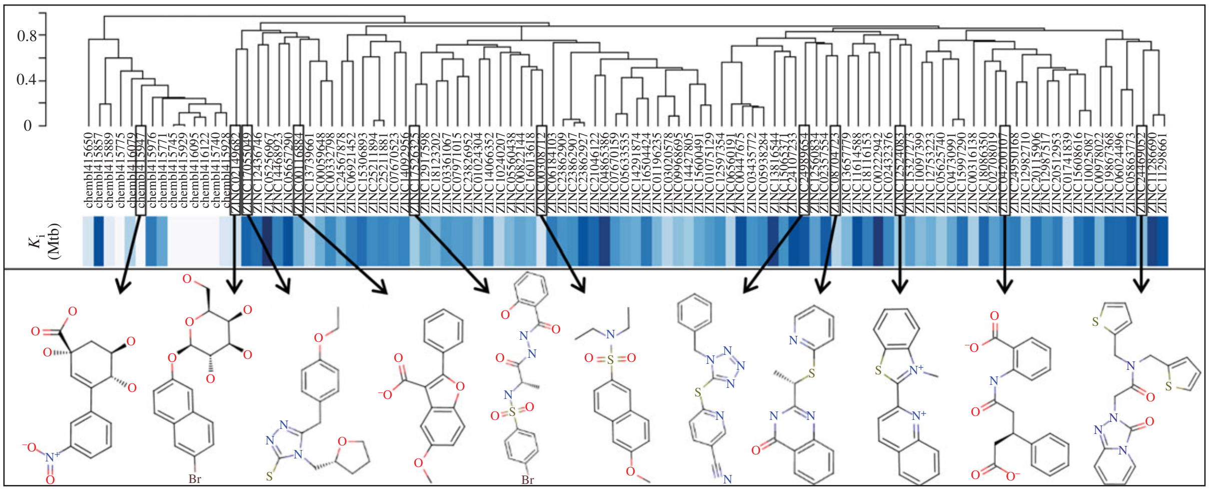


Figure 3. Diversity of new inhibitors of *M. tuberculosis* DHQase. The hierarchical clustering plot at the top includes all the new inhibitors for this version of the target arising from our study (89 molecules identified by their ZINC codes) as well as all the previously known confirmed inhibitors for this target in the ChEMBL database (14 molecules identified by their ChEMBL codes). The dendrogram shows the results of the clustering of these molecules in terms of their chemical structure similarity. The relative $K_i$ of these molecules is shown as different shades of blue (the darker the shade, the higher the $K_i$). $K_i$ values for the 14 ChEMBL inhibitors range from 54 nM to 200 μM (these are mostly optimized hits unlike our new structurally diverse inhibitors, which come directly from virtual screening). Below the $K_i$ band, representatives from the 48 new core scaffolds are shown. These are strikingly diverse compared to the previously known core scaffold (first cluster on the left).

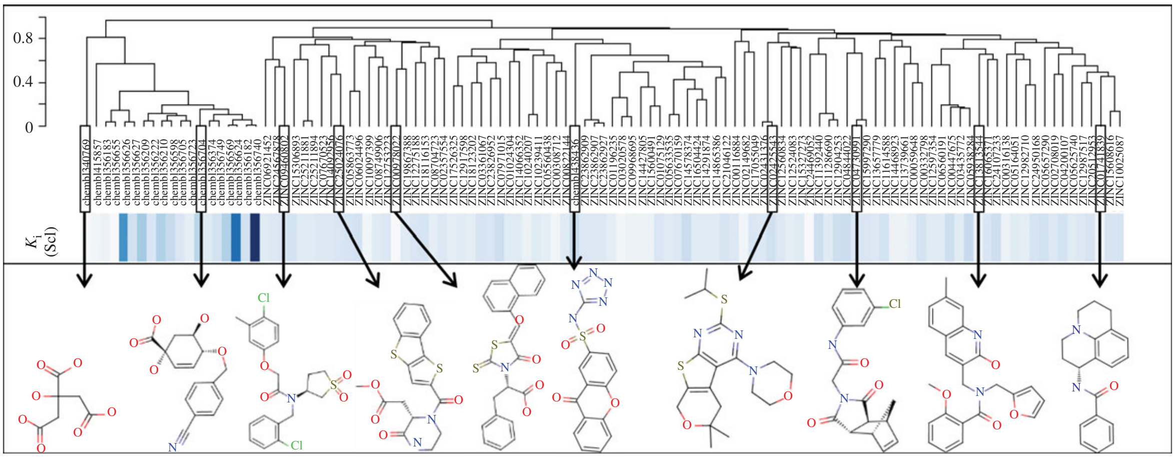


Figure 4. Diversity of new inhibitors of *S. coelicolor* DHQase. The hierarchical clustering plot at the top includes all the new inhibitors for this version of the target arising from our study (91 molecules identified by their ZINC codes) as well as all the previously known confirmed inhibitors for this target in the ChEMBL database (20 molecules identified by their ChEMBL codes). The dendrogram shows the results of the clustering of these molecules in terms of their chemical structure similarity. The relative $K_i$ of these molecules is shown as different shades of blue (the darker the shade, the higher the $K_i$). $K_i$ values for the 20 ChEMBL inhibitors range from 7.3 to 2000 μM (these are mostly optimized hits unlike our new structurally diverse inhibitors that come directly from virtual screening). Five additional chemical series were found for *S. coelicolor* DHQase, which adds to a total of 53 new molecular scaffolds contained in these 100 DHQase inhibitors. These are strikingly diverse compared to most of the previously known core scaffolds (first cluster on the left).

discussed in Ballester & Mitchell [30]. Also, because RF-Score identified different inhibitors than those from GOLD scoring functions, a consensus score based on all four scoring functions is likely to provide a better ability to identify binders, or at the very least a different set of new inhibitors. Furthermore, shape similarity and docking exclusively exploiting *M. tuberculosis* DHQase structures and inhibitors, instead of our mixed use of data for *S. coelicolor* and *H. pylori* DHQase, should allow better recognition of inhibitors of this version of the enzyme. On the other hand, Fischbach & Walsh [5] have argued that given current uncertainties as to how antibiotics get into bacterial cells and the structural diversity of antibacterial targets (and thus the molecules that bind them), libraries developed for other therapeutic areas may be just as likely to harbour hits as compound libraries developed for antibacterial screening. Consequently, we did not focus on the library design issue, although quantitative frameworks to design optimal HTS screening compound collections could be also beneficial for virtual screening performance in the future [44]. Most importantly, the extreme efficiency of USR permits one to exploit the chemical diversity contained in many billions of molecules in a time- and resource-efficient manner. Therefore, the use of much larger databases of molecules, each represented by a comprehensive ensemble of energy-accessible conformations, should lead to an even higher proportion of diverse inhibitors. Incidentally, we did not generate multiple conformers per molecule in our study, partly because the flexibility of the putative ligand in the binding site was going to be sampled during the pose generation stage. However, applying USR on a multi-conformer database instead would have led to the retrieval of more similarly shaped molecules that otherwise would have been missed.[4] The application of such improvements is expected to result in additional DHQase inhibitors.

Overall, our results are directly relevant for those medicinal chemists who are interested in studying the selectivity of these new inhibitors with a view to using them as *in vivo* chemical probes [45]. These compounds could also be of interest as starting points towards generating lead compounds for rational drug design. These new inhibitors can now be evaluated for whole-cell antibacterial activity, and also undesirable properties such as hitting unwanted enzyme targets, to assess whether these are likely to be useful for developing leads. The hit-to-lead process is likely to require testing chemical modifications around each core scaffold, as none of the antibacterial hits found in intensive HTS campaigns were directly a lead [3]. However, both the number, and especially the variety of scaffolds of the hits maximizes the chances of at least some of the resulting series ultimately becoming a lead. We are disclosing the structures of all these new classes of DHQase inhibitors in the electronic supplementary material of this article to make such follow-up studies possible.

Finally, the reported methodology and proposed improvements can be applied to any other antibacterial target with resolved protein structures and known binders. This will be particularly advantageous for those targets for which no, or only structurally similar, active molecules are known. Surprisingly, despite the need for fast and cost-effective tools for hit identification, virtual screening is currently underused [46] and the antibacterial area is not an exception [6]. Virtual screening is, we believe, an excellent strategy for generating hits, though these may require significant experimental work to develop into useful leads. We hope that our study will contribute to encouraging budget-constrained academic laboratories to collaborate with virtual screening experts to find novel active scaffolds for the many validated molecular targets already available. Beyond antibacterials, validated virtual screening tools are also promising to tackle emerging antimicrobial drug resistance in other infectious diseases such as those caused by viruses and multicellular parasites (e.g. drug-resistant mutant influenza viruses [47] or resistance to existing antimalarial drugs [48]). Looking more broadly, the integration of virtual screening and HTS, which is still underdeveloped in any therapeutic area [34], can be extremely productive. For a new target with known substrate and structural model, virtual screening can be used to quickly identify potential chemical probes to investigate the target's druggability and *in vivo* response to dose-dependent modulation (both are often necessary to secure the funding for a full-scale HTS). For validated targets, virtual screening can help to greatly reduce the costs, increase the hit rate and speed up the operation of HTS [49] by enriching the screening library with a large set of molecules likely to be active. We believe that this practise would be very valuable in these targets where HTS performance has been unsatisfactory, as the latter may simply be due to the selected screening library containing few inhibitors in the first place. Lastly, a minority of targets are not amenable to HTS [50] due to problems associated with generating sufficient purified protein or substrate, or because the assay cannot be miniaturized to a robust HTS format. Such cases provide an obvious niche for the exclusive application of prospective virtual screening.

P.J.B., M.M., J.B.O.M. and J.B. conceived the computational part of this work, and together with R.L.M.R. designed the *in silico* workflows. P.J.B. carried out the USR screens, re-scored docking poses with RF-Score, evaluated ranked lists for each virtual screening protocol, purchased top-ranked compounds for testing and analysed the diversity of the new inhibitors. M.M. carried out docking pose generation, re-scoring with GOLD and generated ranked lists for all ligands based on the three scores within GOLD. R.L.M.R. performed the *in silico* toxicity screen and promiscuity filters. N.I.H. tested the purchased compounds *in vitro* and analysed the results with C.A. P.J.B. wrote the manuscript using the information provided by M.M., R.L.M.R. and N.I.H. All authors discussed results and commented on the manuscript. We thank the CCDC for a licence of the GOLD software, Dr Adrian Lapthorn (University of Glasgow) for providing the pET-15a/aroQ (*S. coelicolor*) plasmid, Chemaxon for a licence for InstantJChem and MarvinView granted to P.J.B. and the developers of the ZINC database for making this resource freely available as well as Lhasa Ltd., and Ideaconsult Ltd., for licences and feedback on the use of their toxicology software. The following sources of funding are gratefully acknowledged: the Medical Research Council for a Methodology Research Fellowship (grant no. G0902106 awarded to P.J.B.), the Biotechnology and Biological Sciences Research Council (grant no. BB/G000247/1 awarded to J.B.O.M.), the Scottish Universities Life Sciences Alliance (J.B.O.M.), the European Commision for an EST PhD studentship under the Marie Curie Actions Programme (M.M.), the Bill and Melinda Gates Foundation (N.I.H.), the Engineering and Physical Sciences Research Council and Unilever PLC (R.L.M.R.) and the Royal Society for a University Research Fellowship (J.B.).

## REFERENCES

1 Koul, A., Arnoult, E., Lounis, N., Guillemont, J. & Andries, K. 2011 The challenge of new drug discovery for tuberculosis. *Nature* **469**, 483–490. (doi:10.1038/nature09657)

2 Balganesh, T. S., Alzari, P. M. & Cole, S. T. 2008 Rising standards for tuberculosis drug development. *Trends Pharmacol. Sci.* **29**, 576–581. (doi:10.1016/j.tips.2008.08.001)

[4] A conformer registered in a database may be different from the bioactive conformation and thus multiple conformers per molecule are usually beneficial.

3 Payne, D. J., Gwynn, M. N., Holmes, D. J. & Pompliano, D. L. 2007 Drugs for bad bugs: confronting the challenges of antibacterial discovery. *Nat. Rev. Drug Discov.* **6**, 29–40. (doi:10.1038/nrd2201)

4 Chan, P. F., Holmes, D. J. & Payne, D. J. 2004 Finding the gems using genomic discovery: antibacterial drug discovery strategies: the successes and the challenges. *Drug Discov. Today Ther. Strateg.* **1**, 519–527. (doi:10.1016/j.ddstr.2004.11.003)

5 Fischbach, M. A. & Walsh, C. T. 2009 Antibiotics for emerging pathogens. *Science* **325**, 1089–1093. (doi:10.1126/science.1176667)

6 Simmons, K. J., Chopra, I. & Fishwick, C. W. G. 2010 Structure-based discovery of antibacterial drugs. *Nat. Rev. Microbiol.* **8**, 501–510. (doi:10.1038/nrmicro2349)

7 Wright, G. D. 2000 Resisting resistance: new chemical strategies for battling superbugs. *Chem. Biol.* **7**, R127–R132. (doi:10.1016/S1074-5521(00)00126-5)

8 Ambur, O. H., Davidsen, T., Frye, S. A., Balasingham, S. V., Lagesen, K., Rognes, T. & Tønjum, T. 2009 Genome dynamics in major bacterial pathogens. *FEMS Microbiol. Rev.* **33**, 453–470. (doi:10.1111/j.1574-6976.2009.00173.x)

9 Hopkins, A. L., Bickerton, G. R., Carruthers, I. M., Boyer, S. K., Rubin, H. & Overington, J. 2011 Rapid analysis of pharmacology for infectious diseases. *Curr. Top. Med. Chem.* **11**, 1292–1300. (doi:10.2174/156802611795429130)

10 Keseru, G. M. & Makara, G. M. 2006 Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **11**, 741–748. (doi:10.1016/j.drudis.2006.06.016)

11 Zhang, J., Wu, X. & Sills, M. A. 2005 Probing the primary screening efficiency by multiple replicate testing: a quantitative analysis of hit confirmation and false screening results of a biochemical assay. *J. Biomol. Screen.* **10**, 695–704. (doi:10.1177/1087057105279149)

12 Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J. & Nadon, R. 2006 Statistical practice in high-throughput screening data analysis. *Nat. Biotechnol.* **24**, 167–175. (doi:10.1038/nbt1186)

13 Nwaka, S. & Hudson, A. 2006 Innovative lead discovery strategies for tropical diseases. *Nat. Rev. Drug Discov.* **5**, 941–955. (doi:10.1038/nrd2144)

14 Gourley, D. G., Shrive, A. K., Polikarpov, I., Krell, T., Coggins, J. R., Hawkins, A. R., Isaacs, N. W. & Sawyer, L. 1999 The two types of 3-dehydroquinase have distinct structures but catalyze the same overall reaction. *Nat. Struct. Biol.* **6**, 521–525. (doi:10.1038/9287)

15 Roszak, A. W., Robinson, D. A., Krell, T., Hunter, I. S., Fredrickson, M., Abell, C., Coggins, J. R. & Lapthorn, A. J. 2002 The structure and mechanism of the type II dehydroquinase from *Streptomyces coelicolor*. *Structure* **10**, 493–503. (doi:10.1016/S0969-2126(02)00747-5)

16 Blomberg, L. M., Mangold, M., Mitchell, J. B. O. & Blumberger, J. 2009 Theoretical study of the reaction mechanism of *Streptomyces coelicolor* type II dehydroquinase. *J. Chem. Theory Comput.* **5**, 1284–1294. (doi:10.1021/ct800480d)

17 Parish, T. & Stoker, N. G. 2002 The common aromatic amino acid biosynthesis pathway is essential in *Mycobacterium tuberculosis*. *Microbiology* **148**, 3069–3077.

18 Robinson, D. A., Stewart, K. A., Price, N. C., Chalk, P. A., Coggins, J. R. & Lapthorn, A. J. 2006 Crystal structures of *Helicobacter pylori* type II dehydroquinase inhibitor complexes: new directions for inhibitor design. *J. Med. Chem.* **49**, 1282–1290. (doi:10.1021/jm0505361)

19 Irwin, J. J. & Shoichet, B. K. 2005 ZINC: A free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model* **45**, 177–182. (doi:10.1021/ci049714+)

20 Ballester, P. J. & Richards, W. G. 2007 Ultrafast shape recognition for similarity search in molecular databases. *Proc. R. Soc A* **463**, 1307–1321. (doi:10.1098/rspa.2007.1823)

21 Ballester, P. J. 2011 Ultrafast shape recognition: method and applications. *Future Med. Chem.* **3**, 65–78. (doi:10.4155/fmc.10.280)

22 Ballester, P. J., Westwood, I., Laurieri, N., Sim, E. & Richards, W. G. 2010 Prospective virtual screening with ultrafast shape recognition: the identification of novel inhibitors of arylamine N-acetyltransferases. *J. R. Soc. Interface* **7**, 335–342. (doi:10.1098/rsif.2009.0170)

23 Simon-Hettich, B., Rothfuss, A. & Steger-Hartmann, T. 2006 Use of computer-assisted prediction of toxic effects of chemical substances. *Toxicology* **224**, 156–162. (doi:10.1016/j.tox.2006.04.032)

24 Johnson, D. E. & Rodgers, A. D. 2006 Computational toxicology: heading toward more relevance in drug discovery and development. *Curr. Opin. Drug Discov. Dev.* **9**, 29–37.

25 Eldridge, M. D., Murray, C. W., Auton, T. R., Paolini, G. V. & Mee, R. P. 1997 Empirical scoring functions. I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput. Aided Mol. Des.* **11**, 425–445. (doi:10.1023/A:1007996124545)

26 Jones, G., Willett, P. & Glen, R. C. 1995 Molecular recognition of receptor sites using a genetic algorithm with a description of desolvation. *J. Mol. Biol.* **245**, 43–53. (doi:10.1016/S0022-2836(95)80037-9)

27 Mooij, W. T.M. & Verdonk, M. L. 2005 General and targeted statistical potentials for protein–ligand interactions. *Proteins* **61**, 272–287. (doi:10.1002/prot.20588)

28 Charifson, P. S., Corkery, J. J., Murcko, M. A. & Walters, W. P. 1999 Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **42**, 5100–5109. (doi:10.1021/jm990352k)

29 Oda, A., Tsuchida, K., Takakura, T., Yamaotsu, N. & Hirono, S. 2006 Comparison of consensus scoring strategies for evaluating computational models of protein–ligand complexes. *J. Chem. Inf. Model* **46**, 380–391. (doi:10.1021/ci050283k)

30 Ballester, P. J. & Mitchell, J. B. O. 2010 A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking. *Bioinformatics* **26**, 1169–1175. (doi:10.1093/bioinformatics/btq112)

31 Ballester, P. J. & Mitchell, J. B. O. 2011 Comments on 'leave-cluster-out cross-validation is appropriate for scoring functions derived from diverse protein data sets': significance for the validation of scoring functions. *J. Chem. Inf. Model* **51**, 1739–1741. (doi:10.1021/ci200057e)

32 Toscano, M. D., Stewart, K. A., Coggins, J. R., Lapthorn, A. J. & Abell, C. 2005 Rational design of new bifunctional inhibitors of type II dehydroquinase. *Org. Biomol. Chem.* **3**, 3102–3104. (doi:10.1039/b507156a)

33 Payne, R. J., Riboldi-Tunnicliffe, A., Kerbarh, O., Abell, A. D., Lapthorn, A. J. & Abell, C. 2007 Design, synthesis, and structural studies on potent biaryl inhibitors of type II dehydroquinases. *ChemMedChem* **2**, 1010–1013. (doi:10.1002/cmdc.200700062)

34 Ripphausen, P., Nisius, B., Peltason, L. & Bajorath, J. 2010 Quo vadis, virtual screening? A comprehensive survey of prospective applications. *J. Med. Chem.* **53**, 8461–8467. (doi:10.1021/jm101020z)

35 Schneider, G. 2010 Virtual screening: an endless staircase? *Nat. Rev. Drug Discov.* **9**, 273–276. (doi:10.1038/nrd3139)

36 Nicolaou, C. A., Brown, N. & Pattichis, C. S. 2007 Molecular optimization using computational multi-objective methods. *Curr. Opin. Drug Discov. Dev.* **10**, 316–324.

37 Cosconati, S., Hong, J. A., Novellino, E., Carroll, K. S., Goodsell, D. S. & Olson, A. J. 2008 Structure-based virtual screening and biological evaluation of Mycobacterium tuberculosis adenosine 5′-phosphosulfate reductase inhibitors. *J. Med. Chem.* **51**, 6627–6630. (doi:10.1021/jm800571m)

38 Markt, P. *et al.* 2008 Discovery of novel PPAR ligands by a virtual screening approach based on pharmacophore modeling, 3D shape, and electrostatic similarity screening. *J. Med. Chem.* **51**, 6303–6317. (doi:10.1021/jm800128k)

39 Li, H. *et al.* 2009 Identification of novel falcipain-2 inhibitors as potential antimalarial agents through structure-based virtual screening. *J. Med. Chem.* **52**, 4936–4940. (doi:10.1021/jm801622x)

40 Naylor, E. *et al.* 2009 Identification of a chemical probe for NAADP by virtual screening. *Nat. Chem. Biol.* **5**, 220–226. (doi:10.1038/nchembio.150)

41 Das, S., Krein, M. P. & Breneman, C. M. 2010 Binding affinity prediction with property-encoded shape distribution signatures. *J. Chem. Inf. Model.* **50**, 298–308. (doi:10.1021/ci9004139)

42 Kinnings, S. L., Liu, N., Tonge, P. J., Jackson, R. M., Xie, L. & Bourne, P. E. 2011 A machine learning-based method to improve docking scoring functions and its application to drug repurposing. *J. Chem. Inf. Model.* **51**, 408–419. (doi:10.1021/ci100369f)

43 Li, L., Wang, B. & Meroueh, S. O. 2011 Support vector regression scoring of receptor-ligand complexes for rank-ordering and virtual screening of chemical libraries. *J. Chem. Inf. Model.* **51**, 2132–2138. (doi:10.1021/ci200078f)

44 Clemons, P. A., Wilson, J. A., Dančík, V., Muller, S., Carrinski, H. A., Wagner, B. K., Koehler, A. N. & Schreiber, S. L. 2011 Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proc. Natl Acad. Sci. USA* **108**, 6817–6822. (doi:10.1073/pnas.1015024108)

45 Garcia-Serna, R. & Mestres, J. 2011 Chemical probes for biological systems. *Drug Discov. Today* **16**, 99–106. (doi:10.1016/j.drudis.2010.11.004)

46 Vasudevan, S. R. & Churchill, G. C. 2009 Mining free compound databases to identify candidates selected by virtual screening. *Exp. Opin. Drug Discov.* **4**, 901–906. (doi:10.1517/17460440903190953)

47 Das, K., Aramini, J. M., Ma, L.-C., Krug, R. M. & Arnold, E. 2010 Structures of influenza A proteins and insights into antiviral drug targets. *Nat. Struct. Mol. Biol.* **17**, 530–538. (doi:10.1038/nsmb.1779)

48 Kongsaeree, P., Khongsuk, P., Leartsakulpanich, U., Chitnumsub, P., Tarnchompoo, B., Walkinshaw, M. D. & Yuthavong, Y. 2005 Crystal structure of dihydrofolate reductase from Plasmodium vivax: Pyrimethamine displacement linked with mutation-induced resistance. *Proc. Natl. Acad. Sci. USA* **102**, 13 046–13 051. (doi:10.1073/pnas.0501747102)

49 Peakman, T., Franks, S., White, C. & Beggs, M. 2003 Delivering the power of discovery in large pharmaceutical organizations. *Drug Discov. Today* **8**, 203–211. (doi:10.1016/S1359-6446(03)02612-6)

50 Drewry, D. H. & Macarron, R. 2010 Enhancements of screening collections to address areas of unmet medical need: an industry perspective. *Curr. Opin. Chem. Biol.* **14**, 289–298. (doi:10.1016/j.cbpa.2010.03.024)