

Identification of Novel Antimalarial Chemotypes via Chemoinformatic Compound Selection Methods for a High-Throughput Screening Program against the Novel Malarial Target, PfNDH2: Increasing Hit Rate via Virtual Screening Methods

Raman Sharma,[†] Alexandre S. Lawrenson,[†] Nicholas E. Fisher,[‡] Ashley J. Warman,[‡] Alison E. Shone,[‡] Alasdair Hill,[‡] Alison Mbekeani,[‡] Chandrakala Pidathala,[†] Richard K. Amewu,[†] Suet Leung,[†] Peter Gibbons,[†] David W. Hong,[†] Paul Stocks,[†] Gemma L. Nixon,[‡] James Chadwick,[†] Joanne Shearer,^{||} Ian Gowers,^{||} David Cronk,^{||} Serge P. Parel,[§] Paul M. O'Neill,[†] Stephen A. Ward,[‡] Giancarlo A. Biagini,[‡] and Neil G. Berry^{*†}

[†]Robert Robinson Laboratories, Department of Chemistry, University of Liverpool, Liverpool L69 7ZD, United Kingdom

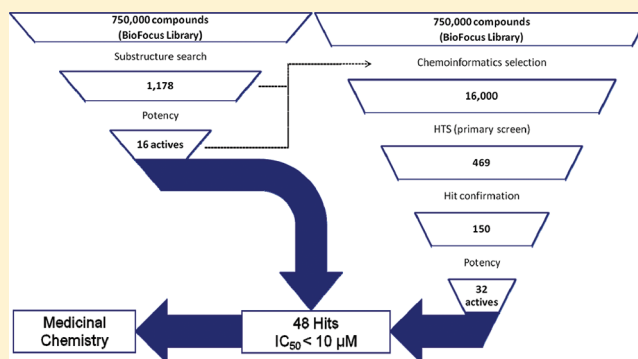
[‡]Liverpool School of Tropical Medicine, Pembroke Place, Liverpool, L3 5QA, United Kingdom

[§]BioFocus, Gewerbestrasse 16, CH-4123, Allschwil, Switzerland

^{||}BioFocus, Chesterford Research Park, CB10 1XL, United Kingdom

Supporting Information

ABSTRACT: Malaria is responsible for approximately 1 million deaths annually; thus, continued efforts to discover new antimalarials are required. A HTS screen was established to identify novel inhibitors of the parasite's mitochondrial enzyme NADH:quinone oxidoreductase (PfNDH2). On the basis of only one known inhibitor of this enzyme, the challenge was to discover novel inhibitors of PfNDH2 with diverse chemical scaffolds. To this end, using a range of ligand-based chemoinformatics methods, ~17000 compounds were selected from a commercial library of ~750000 compounds. Forty-eight compounds were identified with PfNDH2 enzyme inhibition IC₅₀ values ranging from 100 nM to 40 μM and also displayed exciting whole cell antimalarial activity. These novel inhibitors were identified through sampling 16% of the available chemical space, while only screening 2% of the library. This study confirms the added value of using multiple ligand-based chemoinformatic approaches and has successfully identified novel distinct chemotypes primed for development as new agents against malaria.



INTRODUCTION

Malaria is a life-threatening disease, which is responsible for roughly 1 million deaths each year.¹ Approximately 40%² of the world's population is exposed to the risk of malaria, particularly those in tropical and subtropical countries.³ Malaria also poses a huge economic burden in countries where the disease is endemic, cutting economic growth rates by as much as 1.3% in countries with high disease rates.^{1,4}

Previous successes in attempting to eradicate the disease were only relatively short-lived due to increasing resistance of the mosquito to insecticides⁵ and of the parasite to established drugs.⁶ In many parts of the world, the parasites have developed resistance to a number of drug classes.^{2,7} Emerging resistance is responsible for a recent increase in malaria mortality, particularly in countries that had previously eliminated its presence. The disease has worldwide implications due to the increase in air travel, with travelers from malaria-free areas of the world

especially vulnerable;¹ therefore, the development of new and more effective antimalarial chemotherapy has never been more important.

The *Plasmodium falciparum* parasite, which is the most deadly form of the malaria parasite,¹ has developed resistance to chloroquine in many parts of the world. There are strenuous and continued efforts to identify novel small molecules that either circumvent chloroquine resistance or act on alternative stages of the malaria parasite lifecycle.⁸ One target that has received attention is the mitochondrial respiratory chain of *P. falciparum*. Atovaquone (part of combination therapy Malarone) targets the cytochrome bc₁ complex (complex III) in the mitochondrial electron transport chain of *P. falciparum*.⁹ The electron transport chain is an attractive chemotherapeutic

Received: November 25, 2011

Published: March 1, 2012

target in that it differs from the human host in that it lacks a canonical protonmotive NADH:ubiquinone oxidoreductase (complex I); rather, it has a single subunit, nonprotonmotive NDH2.¹⁰ Using an *Escherichia coli* NADH dehydrogenase knockout strain (ANN0222, *ndh::tet nuoB::nptI-sacRB*), we have developed a heterologous expression system for PfNDH2 facilitating its physicochemical and enzymological characterization.^{10b} PfNDH2 is a metabolic choke point in the respiratory chain of the parasite's mitochondria and is the focus of the discovery program toward the development of novel therapy for uncomplicated malaria. We have previously described a miniaturized spectrophotometric assay for recombinant PfNDH2 (steady state NADH oxidation and ubiquinone reduction monitored at 340 and 283 nm, respectively) with robust assay performance measures.¹¹ This assay forms the basis of the high-throughput screen (HTS) sequential screening program.

The objective of this program was to identify novel chemotypes that act as selective inhibitors of PfNDH2. Upon commencement of the program, there was only one molecule that was known to exhibit PfNDH2 activity, 1-hydroxy-2-dodecyl-4-(1*H*)quinolone (HDQ) (Figure 1), which has an

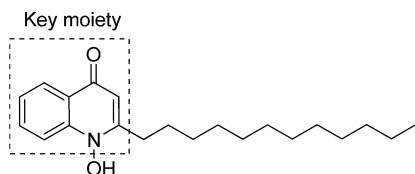


Figure 1. Structure of HDQ and the identity of the proposed key moiety in the structure.

IC₅₀ value of 70 nM.^{10b,11,12} Our discovery program used virtual screening to select a 16050 compound subset of the ~750000 compound BioFocus library for a high-throughput in vitro screening campaign. The authors describe here the use of chemoinformatics, virtual screening, and computational methods to identify the “best” subset of compounds from the BioFocus database and the screening of these compounds using the PfNDH2 HTS assay to identify novel chemical hit series, with diverse chemotypes, for subsequent medicinal chemistry development.

Virtual screening emerged in the 1990s as a way of predicting bioactive compounds using computational methods.¹³ There are many examples of successful use of such approaches in the literature for both hit finding and hit-to-lead optimization stages of the drug discovery process.¹⁴ One example of virtual screening successfully influencing the discovery of products now on the market is that of Aggrastat.¹⁵ The methods of virtual screening are usually defined as either structure-based or ligand-based. Structure-based approaches use knowledge of the 3D structure of the biological target, whereas ligand-based approaches rely on the knowledge of the structure of compounds exhibiting the desired activity.¹⁶ This current work adopts a ligand-based approach as there is no crystal structure of the PfNDH2 target and it displays poor homology with any published structure in the PDB.^{10b} Ligand-based virtual screening approaches rely on the complementary area of chemoinformatics, which has been defined as the “The application of informatics methods to solve chemical problems”.^{14c} There have been an increasing number of publications that have made use of chemoinformatics in recent years,¹⁷ partly driven by the

increasing pressures on the pharmaceutical industry to increase productivity, while decreasing costs. Using computational approaches to expedite the identification of candidate molecules that are predicted to possess the desired properties is an efficient approach to discovery. With the incorporation of computational filters for properties synonymous with unfavorable PK/PD profiles, the aim of chemoinformatics is to reduce compound attrition rates at all stages in the discovery process.

In this work, our definition of the “best” set of compounds is a multidimensional challenge. The first dimension of the metric to identify the “best” set of compounds is whether a compound is predicted by virtual screening and chemoinformatic methods to possess the desired activity against PfNDH2 in the HTS assay. The second dimension of this quality metric is that the compounds assayed possess desirable lead/druglike characteristics. Over the past few years, there have been many publications concerning the identification of molecular properties of compounds that are common among drugs and drug candidates.¹⁸ This has led to many rules of thumb to guide discovery of molecules that are likely to possess the appropriate absorption, distribution, metabolism, excretion, and toxicity (ADMET) characteristics and be amenable for further development along the drug discovery pipeline.^{19,14a,20} The third dimension of the metric for “best” set of compounds is that of optimal sampling of chemical space. In the project, we had the resources to assay approximately 17000 compounds in total; therefore, we had to explore chemical space in the most efficient way possible to yield the most information-rich set of results that would be the most informative for both the sequential screening stage and the final compound selection phase.^{14b} In summary, the “best” set of compounds are those that are predicted to be active against the target, possess desirable drug/lead like characteristics, and sample chemical space the most effectively.

In this paper, we described the use of chemoinformatics and virtual screening methods to select an optimal set of compounds for HTS screening leading to the identification of new chemotypes that display activity against the PfNDH2 malarial target. Overall, our approach consisted of four fundamental stages: (i) transfer and optimization of assay from in-house to BioFocus, (ii) substructure search of quinolin-4-one in BioFocus library followed by preliminary HTS screening, (iii) chemoinformatic compound selection from BioFocus library based on results of preliminary screening, and (iv) HTS screening of selected compounds followed by hit confirmation.

RESULTS AND DISCUSSION

Preliminary Screening. Initial substructure searching of the BioFocus library for compounds, which contained the “key moiety”, the core of HDQ (Figure 1), revealed 1175 compounds. These compounds underwent screening using the spectrophotometric assay previously developed and validated by us, which had subsequently been transferred to BioFocus.^{10b} Duplicate five point dose–response curves ranging from 20 μM to 78 nM revealed the presence of 54 new active compounds (IC₅₀ < 20 μM) that had confirmed purity by LCMS of >70%. The identity of these active compounds and HDQ along with the identity of the inactive compounds provided the basis of the virtual screening and compound selection.

Compound Selection Using Chemoinformatics. As the primary goal of this work was to identify novel chemotypes, we used chemoinformatic methods to achieve a “scaffold hop”, that is, identification of structurally diverse compounds that exhibit

the desired biological activity.²¹ We applied seven different virtual screening methods in parallel to identify compounds from the BioFocus library.

The first three virtual screening methods used molecular fingerprints. Molecular fingerprints have been successfully used in similarity searching many times. Indeed, 2D fingerprint searching has been shown to outperform 3D methods in some situations.²² Recently, it has been reported that fingerprints can be used successfully to achieve a “scaffold hop”.²³ The use of fingerprints is particularly attractive for large scale screening as the calculations are quick (as compared with other methods such as protein:ligand docking), and 3D structures of the molecules do not need to be generated or stored.

We applied a range of different fingerprinting methods to navigate chemical space in a variety of different ways. We employed MDL molecular access system (MACCS) keys, ECFP2, and FCFP2 to identify compounds similar to any of the 55 actives (54 from substructure screening plus HDQ) previously identified. The MDL keys are a set of predefined substructural fragments that have been used in similarity searching due their speed of calculation and comparison and broad experience in application.²⁴ The publicly available 166 bit key set has found a variety of uses in the drug discovery workflow.²⁵ Extended connectivity fingerprints (ECFPs) have been shown to have a number of strengths that make them useful for similarity searching. ECFPs are a fingerprint methodology explicitly designed to capture molecular features relevant to molecular activity. They can be quickly calculated, as they are not defined a priori (in contrast with MDL MACCS keys) and they can represent novel structural classes.²⁶ Functional class fingerprints (FCFPs) are a related fingerprint to ECFPs but instead of using a specific atom identifier for the initial atom in the algorithm to generate the fingerprint, FCFPs use a more abstract pharmacophoric set of initial atom identifiers based on properties such as hydrogen bond acceptor and donor, negatively and positively ionizable, aromatic, and halogen.²⁶

For all the similarity searches, the Tanimoto coefficient was employed. A study of a wide variety of similarity coefficients (22 in total) that may be employed in similarity searching supports the use of the popular Tanimoto coefficient as compared with the alternatives.²⁷ There is little literature precedent for the cutoff value that should be used for a given fingerprint and similarity coefficient to identify a compound displaying the desired activity. There is, however, a study that used the Daylight fingerprint methodology and Tanimoto coefficient and suggested that a coefficient of greater than or equal to 0.85 would result in a 30% chance of the compound exhibiting similar biological activity to the query molecule.²⁸ However, the fingerprints employed in the current study are different to the Daylight methodology in the aforementioned study so these analyses may not apply. Consequently, we have chosen similarity thresholds to seek to obtain a “balance” between the numbers of compounds retrieved for each fingerprint used.

Using a threshold of ≥ 0.8 for the MDL MACCS keys found 8784 compounds, whereas for ECFP2, a Tanimoto threshold of ≥ 0.6 gave 333 compounds, and FCFP2 with a Tanimoto coefficient ≥ 0.75 highlighted 435 compounds. Interestingly, these threshold values match up extremely well with a paper (published after this work was performed), which indicates that the probability assignment of a particular similarity measure of finding 50% of all possible actives using a given measure for

MACCS keys should use a value of 0.82, whereas 0.52 should be used for ECFP2 and 0.75 for FCFP2—values very similar to the ones we had selected.²⁹ Thus, the combination of these three fingerprint approaches would hopefully maximize the chances of identifying compounds that were sufficiently similar to one of the query compounds to display inhibition of PfNDH2.

The fourth chemoinformatic method that was used to identify compounds with the desired biological profile was turbo similarity searching.³⁰ Turbo similarity searching has been developed to increase the effectiveness of virtual screening when there is little information available. The approach uses the multiple databases searches using the nearest neighbors resulting from an initial similarity search.³¹ The results from these searches are then combined together using group fusion and the molecules ranked by the fusion score. This approach has been shown in simulated virtual screening to show impressive performance in retrieving active molecules. For our work, turbo similarity searching was performed using the 55 active compounds as initial query molecules and identifying the 50 nearest neighbors to each compound using ECFP4 fingerprints and the Tanimoto coefficient. These molecules were then used as the queries searching the database utilizing the ECFP4 fingerprints. The top 250 compounds from each nearest neighbor search were summed using the Tanimoto coefficient and identified 4891 unique compounds.

The fifth method employed was that of substructural searches. The concept for these searches was to identify bioisosteres of the key moiety headgroup of HDQ [4-(1H)quinolone] (Figure 1), thus employing the principle of isosterism, in that similar molecules usually possess similar properties.³² Thirty-eight scaffold isosteres of the 4-(1H)-quinolone headgroup (HDQ) were identified using a variety of 2D topological pharmacophores.³³ The pharmacophores used were based on the Cahart strategy where a feature refers to two chemical groups separated by a certain 2D path length. Atom-based, fragment-based, and pharmacophore-based binary descriptors were investigated using the Tanimoto coefficient. The atom-based approach is likely to retrieve very close analogues to the query compound, while fragment and pharmacophore searches were more likely to find more diverse analogues. Continuous descriptors were also employed to identify isosteres of the 4-(1H)quinolone headgroup using a modified Burden³³ number using the Euclidean distance. Both binary and continuous approaches were used to search a range of databases (e.g., ACL and NCI diversity database³³) using 4-(1H)quinolone as the query. Compounds identified by these methods were examined by eye by medicinal chemists before inclusion in the list of scaffold isosteres. The thirty-eight substructures were used to search the BioFocus database, and 137693 compounds were found.

This number of compounds was too large for this current work, and a simple selection procedure was employed. For those isosteres that retrieved less than 1000 compounds, all were kept, but for searches that highlighted more than 1000 compounds, a maximally diverse selection procedure was used to sample the compounds identified, while retaining the maximum coverage of chemical space possible. To achieve this, we used FCFP4 fingerprints together with diversity selection to achieve a maximum similarity selection. These procedures resulted in the identification of 5247 molecules.

The sixth method used was a naïve Bayesian classifier. Bayesian methods rely on the estimation of probability distributions of numerical representations of compounds

based on molecular properties of fingerprints.³⁴ Each descriptor value of a molecule is considered in turn, and the probability of activity is considered to be proportional to the ratio of actives to inactives that share that descriptor value. The overall probability of activity is the product of all of the probabilities. This naïve approach assumes statistical independence of descriptors; however, theoretical results suggest that large deviations from independence can be tolerated.³⁵ A Bayesian model was built using the 1175 molecules screened initially (55 “actives” the rest “inactive”) using a variety of physicochemical properties (AlogP, MW, number of hydrogen bond donors, number hydrogen bond acceptors, number of rotatable bonds, polar surface area) and fingerprints (ECFP2) as the molecular descriptors. The descriptors were binned, and low-information content bins were removed due to a poor normalized estimate. A leave-one-out cross-validation process was employed to build the model in which each sample was left out one at a time and a model built using the results of the samples and that model used to predict the left-out sample. The area under the receiver operator curve for the cross-validation set was 0.881. Thus, given an active molecule and an inactive molecule and one used the model to guess which one was the hit, one would be right 88% of the time. Using this split a contingency table revealed the number of true positives, true negatives, false positives, and false negatives. The figure of accuracy of 0.924 indicates that the model generated was very good. The performance of this model using the whole data set was assessed through examination of its enrichment performance. In the first 25% of the molecules tested, over 80% of the active molecules were retrieved. Application of the Bayesian model to the BioFocus collection identified 11702 compounds that were predicted to be active.

Principle component analysis (PCA) was used as the seventh method for compound selection. PCA is a simple, non-parametric method of extracting relevant information from complex data sets that are often confusing, clouded, or even redundant. PCA transforms a number of possibly interdependent or correlated variables into a smaller number of significant, independent, and uncorrelated orthogonal components.³⁶ A principal component model of the initial 55 active compounds was constructed using 20 physicochemical descriptors (see the Experimental Section details for a full list). The model constructed explained 88.5% of the overall variance of the active compounds in three principal components. The Euclidean distance was used to identify from the BioFocus library the 5000 closest compounds in the three-dimensional principle component space to any of the active molecules. Thus, we were identifying the nearest neighbors to the active compounds in PCA space.

The results of the seven virtual screening methods were combined and identified 34356 unique compounds that were predicted by one or more methods to be similar in some way to one or more of the initial set of 55 active compounds. This number of compounds was not able to be screened in the *in vitro* assay; hence, we undertook scoring and diversity sampling protocols to select the best subset of 16000 compounds that achieved the optimal coverage of chemical space while biasing the selection such that compounds that had been selected by more than one cheminformatics approach were more likely to be included in the final 16000 molecular set.

The concept that we employed was that of consensus scoring, which can lead to tremendous improvements in virtual screening through the improved quality of the results obtained.

Ligand-based virtual screening consensus scoring can show improved performance over a single scoring protocol³⁷ due to the fact that the mean of repeated samplings is closer to the true value than one single measurement. Also, different methods agree more on the ranking of actives than inactives, which arises from the fact that different ligand-based virtual screening protocols focus on different aspects of the ligand binding process and thus lead to different false positives. Also, it has been suggested that actives are clustered more tightly than inactives; thus, multiple samplings will recover more actives than inactives. A scoring function was applied to each compound accounting for how often any of the seven virtual screening methods had identified a compound and also took into account key druglike properties (*vide infra*).

The druglike properties that were calculated for each compound were solubility, octanol/water partition coefficient, and molecular weight. Scoring functions were applied for these molecular characteristics displayed in Table 1 and Figure 2.

Table 1. Scoring Functions Ranges

property	more desirable range	less desirable range
log <i>S</i>	>−5	<−6
log <i>P</i>	−1 < log <i>P</i> < 4	log <i>P</i> ≤ −2.5; log <i>P</i> ≥ 5.5
MW	<400	>600

For molecular weight, the lower bound was chosen due to the fact that during hit to lead development and onward through the preclinical discovery pipeline, there is, in general, an increase in molecular weight of a potential candidate. The second influence on our choice of these boundaries was the Lipinski guidelines for passive absorption of drugs.³⁸ Solubility is a key factor in any drug discovery program, and as such, compound predicted solubility was assessed.³⁹ These values were selected as previous work has suggested that for a compound showing high permeability and a potency of 0.1 mg/kg, the aqueous solubility needs to be 1 μg/mL to be completely absorbed.^{18a} For example, for a compound with a molecular weight of 400, 5 μg/mL corresponds to a log *S* value of −5.6. The octanol/water partition coefficient is one of the key molecular characteristics for any compound as it plays a key determinant in preclinical ADMET and the increasing body of evidence that suggests that molecules with optimal lipophilicity might have increased chances of success in development.^{20b} For example, it has been shown that the promiscuity of a given compound increases dramatically if log *P* is greater than 3,^{20a} and other work has suggested that compounds with a log *P* value of less than 4 (and molecular weight less than 400) have a greatly increased chance of success against a comprehensive set of ADMET tests.¹⁹ Taking these into account, a compound scoring function was derived as displayed in Figure 2 and Table 1. Thus, each compound was assigned a score according to its druglikeness considering its solubility, lipophilicity, and aqueous solubility.

Each compound was scored using the seven virtual screening methods described above using range-scaled scores. The results from the three fingerprint methods used the calculated Tanimoto coefficients unaltered. The compounds selected by the turbo similarity search were scored using the Tanimoto coefficient of the nearest neighbor identified in the turbo search. Molecules chosen by the bioisostere substructure search all scored 1. Molecules predicted to be active via the Bayesian model (Bayesian score cutoff >5) were scaled between 0 and 1.

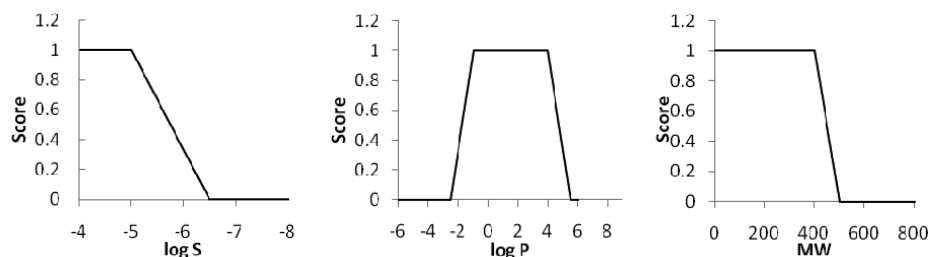


Figure 2. Scoring functions for calculated molecular properties.

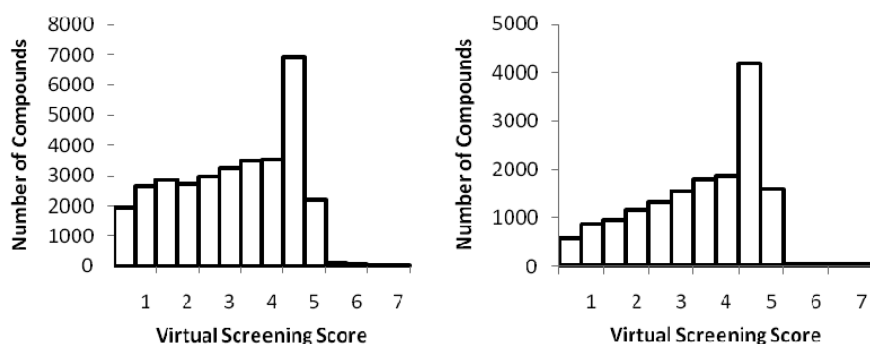


Figure 3. Comparison of the virtual screening scores: (left) 32727 compounds and (right) 16050 compounds.

The PCA distances of the 5000 compounds selected were scaled between 0.5 and 1 with the closest compound scoring 1 and most distant 0.5. These range-scaled virtual screening scores were combined with the physicochemical molecular properties to give a final “virtual screening score” as defined in eq 1. As we sought to favor selection of compounds that were predicted to be active according to our virtual screening models, the range-scaled scoring function was multiplied by four. The score for molecular weight was multiplied by two to enhance the likelihood of selecting smaller compounds, while the solubility and lipophilicity function were applied unchanged. Upon collation of all of the data, there were in total 34356 unique compounds selected, each of which had a virtual screening score calculated. The virtual screening score is our numerical assessment of how “good” a molecule is in terms of likelihood of possessing the desired profile in terms of activity and ADMET properties. The form of the virtual screening score assigned to each compound is as follows:

$$\begin{aligned} \text{virtual screening score} \\ = 4 \times \left(\frac{\text{sum of range scaled scores}}{7} \right) + f(\log S) \\ + f(\log P) + 2 \times f(\text{MW}) \end{aligned} \quad (1)$$

These 34356 compounds were then subjected to a series of filters with hard cut-offs to eliminate compounds that did not possess desirable druglike properties. The filters that were used may be described as “relaxed” Lipinski^{18a} and Veber^{18b} guidelines for oral bioavailability. Compounds were discarded if they failed more than one of the following properties: molecular weight > 600, Alog *P* > 6, number of hydrogen bond donors > 6, number of hydrogen bond acceptors > 11, number of rotatable bonds > 14, and polar surface area > 150 Å². The reason that we slightly relaxed the published guidelines was that the original publications were based on properties of the majority of compounds that passed certain criteria. Thus, by relaxing these guidelines slightly, we hope to maximize our chances of finding

novel chemotypes so we did not want to preclude molecules “too early”. When these cut-offs were applied, there were 32727 compounds that were taken forward.

To select 16000 compounds for HTS, a procedure was employed to choose the most diverse 16000 compound set from the 32727 compounds. This approach was chosen as we desired to identify new scaffolds active against PfNDH2; thus, selecting the most diverse subset was considered advantageous over selecting the top 16000 highest scoring compounds. To achieve the diversity selection, BCUT descriptors were used, as these have previously been used in diversity selection.⁴⁰ All but two compounds were successfully calculated; these two compounds contained a positively charged sulfur atom and a positively charged phosphorus, both types of atoms unlikely to produce a suitable druglike compound.

Histograms of the final virtual screening score were examined for the set of 32727 compounds and the 16000 compound set. It is noteworthy that upon compound selection the distribution of virtual screening scores altered to favor compounds with higher virtual screening scores (Figure 3). Thus, any additional biasing of high-scoring compounds was not considered necessary. Compounds in the lower scoring bins were included, as recent work has highlighted the possibility of achieving a successful scaffold hop may come from compounds that have low similarity scores as assessed using molecular fingerprints.³⁶ Thus, low-scoring compounds overall may well still be able to achieve a scaffold hop.

The effect on this compound selection upon the coverage of chemical space available in the BioFocus library was assessed. The diversity of compounds selected was examined using fingerprints with the Tanimoto coefficient to compare the coverage of the chemical space within the entire BioFocus library. The whole BioFocus library of ~750000 compounds displayed 37881 clusters (ranging from having 1 member to 2927 members), whereas the 16000 compounds selected gave 5913 clusters (having between 1 and 164 members). Thus, 15.6% of the overall diversity within the library was covered by 2.1% of the compounds.

Three potential false negatives were identified from the initial screen from our previous work. A similarity search using ECFP2 was performed against the BioFocus library, and the 50 most similar compounds were added to 16000 compounds to be screened. Thus, in total, there were 16050 compounds to be screened.

SCREENING

A total of 16050 compounds were screened against PfNDH2 over 2 days. Compounds were screened at a concentration of 20 μM generating one data point per compound. The mean and median Z' factors⁴¹ on day one were 0.77 and 0.76, respectively, and on day two were 0.81 and 0.81. A total of 395 compounds (2.5% of the total screened) showed >20% inhibition in the primary screen. To best resolve compounds with weak activity from the noise, a runwise multiplicative correction factor available within Genedata Screener software (Assay Analyzer module, Genedata Inc., Switzerland) was applied to the data.⁴² This correction factor is created by a sophisticated pattern detection algorithm that detects subtle recurring patterns within the data set, and when applied, the median of the distribution of activity is brought in line with zero % inhibition, thereby tightening the data and potentially rescuing false negatives. Using the corrected data, 333 compounds showed >20% inhibition. Of these 333 compounds, 24 had not been identified using noncorrected data. To maximize the chances of identifying weakly active compounds, any compounds that showed >20% inhibition using either the corrected or the noncorrected data were progressed to retesting. This resulted in the progression of 419 compounds (395 + 24) to retest analysis. A total of 469 compounds were tested in triplicate for the ability to inhibit the PfNDH2-catalyzed reaction. A total of 419 of these were identified from the primary screen, and a further 50 were included as structurally related to three potential false negatives in the 1175 screen. Prior to initiating this phase of the HTS, the pharmacology of HDQ was reconfirmed to ensure comparable performance between the primary and the hit confirmation screens. The mean and median Z' factors for each plate were 0.84 and 0.81.

The samples showed good correlation between the primary and the retest data. These compounds were progressed to potency analysis to ensure capture of all compounds that showed activity in either the primary or the retest stages. The hit calling criteria were set such that if any one of the three percentage inhibition measurements for a given compound was >30%, the compound was progressed for potency testing. This captured 108 compounds. In addition, compounds that showed >50% inhibition in the primary screen but whose maximum triplicate retest values was <30% inhibition were included, capturing a further 36 compounds. A further six compounds were selected that showed a maximum retest value of 25–30% and whose screening concentration was <20 μM . This hit-calling approach was designed to ensure that no compounds of interest or weak hits were missed, and 150 compounds met these criteria.

The IC_{50} was determined for these 150 compounds with each compound screened at 10 concentrations (with 3-fold dilution steps between points) as replicated duplicates; that is, each compound was replicated on a plate, and each plate was screened twice. As for previous phases of the HTS, the Z' factor and signal to background obtained were within the accepted range, and the correlation between the two runs was strong with minimal numbers of outliers. The potency screen

produced 32 compounds with an IC_{50} < 10 μM . The most potent compound showed an IC_{50} of 292 nM.

ANALYSIS

The chemoinformatics methods selected 16050 compounds from a library of over 750000 that were subjected to a screening cascade involving primary, retest, and potency screens. The screen resulted in the identification of 32 compounds with an IC_{50} < 10 μM . Interestingly, analysis of the chemoinformatics approaches that selected the hits identified revealed that only 2 of the 32 compounds were selected by more than one virtual screening approach (Table 2), justifying our use of a range of

Table 2. Hit Compounds Identified via Each Chemoinformatic Method

method	no. of hits
MACCS	1
FCFP2	2
ECFP2	0
bioisosteres	15
Bayesian	8
turbo	8
PCA	1

virtual screening approaches. This result indicates that different screening methods probe different areas of chemical similarity space. The two compounds that were identified are displayed in Figure 4 together with the methods that identified them.

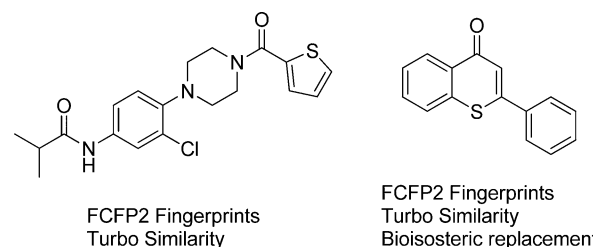


Figure 4. Two compounds identified by more than one chemoinformatic method.

To examine which selection method out of bioisosteres, Bayesian and turbo, gave the most molecular diversity, the diversity of the compounds as selected was assessed using ECFP4 fingerprints and the Tanimoto coefficient (Table 3).

Table 3. Diversity of the Compounds Selected by Three Chemoinformatic Methods

	minimum distance	maximum distance	average distance
bioisostere	0.747	0.950	0.876
Bayesian	0.2697	0.919	0.692
turbo	0.4627	0.934	0.848

As can be seen, the bioisosteric method gave the highest average diversity with turbo only slightly behind, with Bayesian compound selection displaying the least diversity.

In addition to these 32 compounds identified by the screening cascade, 16 compounds from the initial screening displayed IC_{50} < 10 μM against the PfNDH2 enzyme; thus, in total, 48 hit compounds have been identified. Inspection of the

Table 4. Example Chemotypes Discovered Together with their PfNDH2 and Pf(3D7) IC₅₀ Values and Ligand Efficiency (3D7)

Entry	Structure	IC ₅₀ PfNDH2 (nM)	IC ₅₀ 3D7 (nM)	Ligand Efficiency (3D7)
1		78	3800 ± 400	0.25
2		295	5000 ± 1500	0.14
3		512	5560 ± 123	0.17
4		1039	2080 ± 620	0.24
5		1357	2080	0.20

48 hits identified revealed several chemotypes indicating that using many complementary virtual screening methods is a favorable choice when embarking on a hit discovery campaign. Example compounds from some of the chemotypes are displayed in Table 4 together with the PfNDH2 and 3D7 antimalarial activities.

Our approach did provide hits that would not have been identified from simple 2D similarity approaches. To exemplify this, compounds 3, 4, and 5 in Table 4 would not have been identified if we had only relied on substructure searches based on bioisosteric replacement of the core of HDQ. Thus, we believe that our approach of using many cheminformatics methods is justified. This stance is further strengthened through a very recent retrospective study in which a multipronged virtual screening approach very similar to the one employed here was one of the best performing “consensus” approaches.⁴³

The activity range of the example five chemotype compounds against the enzyme is in the tens to hundreds nanomolar. Their whole cell 3D7 growth inhibition activity is in the micromolar range presumably due to the large number of membranes that the compound has to cross to reach its site of action in the mitochondria. Crucially, however, the ligand efficiency (pIC₅₀/number of non hydrogen atoms) of these classes of compounds compares favorably with current therapies (e.g., chloroquine has a value ~0.33 and atovaquone ~0.30), and our compounds are yet to undergo optimization.

After this work was performed, the results from large scale screening campaigns were released from GSK,⁴⁴ St. Jude's,⁴⁵ and Novartis.⁴⁶ Comparing our hits with those from those published, there are no exact duplicate compounds, and using FCFP4 fingerprints, there are no compounds with a Tanimoto value of >0.9. Thus, our suite of cheminformatics methods has

enabled the discovery of novel antimalarial chemotypes as compared with the mass screening campaigns, and the results display extremely promising levels of inhibition for both the target enzyme and the whole cell inhibition. As such, these molecules are very attractive scaffolds on which to base a medicinal chemistry lead development program. Consequently, these chemotypes are currently undergoing further medicinal chemistry investigation and optimization.^{47,48}

CONCLUSIONS

We have employed a wide range of ligand-based cheminformatics methods in the rational selection of 16050 compounds that were predicted to possess activity against PfNDH2 and possess favorable ADMET properties. These compounds were subjected to high-throughput screening triage against the *P. falciparum* NDH2 enzyme target. Hit confirmation and potency determination revealed 48 compounds with IC₅₀ values ranging from 100 nM to 10 μM. Analysis of these hits revealed several novel distinct chemotypes primed for development as new agents against malaria. To our knowledge, hitherto, HDQ and the phenothiazines are the only selective compounds known to inhibit type II NADH:ubiquinone oxidoreductase homologues from any organism. The hits discovered in this HTS therefore represent a significant advancement and should enable chemical biology research into these enzymes from a number of important organisms including *Plasmodium*, *Mycobacterium*, *Trypanosomes*, and yeast.

EXPERIMENTAL SECTION

NADH Oxidation Assay. All reagents were prepared in glassware pre-rinsed with deionized water followed by ethanol (EtOH) and finally deionized water unless otherwise stated. With the exception of assay buffer and 1 M KCN, all reagents were stored on ice. The assay buffer was prepared on the day of the assay and was comprised of 20 mM HEPES (free acid, Sigma H3375) in H₂O, pH 7.4. One molar KCN was prepared in a fume hood using buffer and 10 M HCl to adjust the pH to 7.5. Ten millimolar CoQ (Sigma C7956) was prepared by dissolving 10 mg in 4 mL of EtOH. This was then further diluted 10-fold in 25% DMSO to give 1 mM working solution. One millimolar HDQ was prepared fresh every 2–3 days by diluting a fresh weighing into methanol (MeOH); this was then further diluted 20-fold in buffer to obtain the 50 μM used in the assay (5 μM in 0.5% MeOH final assay concentration). A 0.1 M concentration of NADH (Sigma N8129) was prepared in buffer contained in a light protective eppendorf tube; the reagent was not used more than 3 h after preparation. A Perkin-Elmer EnVision with a 340/25 nm emission filter was used to measure absorbance. Assay plates were black sided and clear bottomed, and the final well volume was 100 μL. Because absorbance is dependent on the path length, the reaction was initiated with a minimal volume of CoQ (2 μL), and a 2% change in volume itself was deemed to have negligible effects on the absorbance and was consistent for all wells. Control wells that received no CoQ received 2 μL of diluent to maintain the path length and concentrations. The expected absorbance drop corresponding to 100% conversion of Coenzyme Q from the quinone to quinol form is 0.15 abs units.

Prior to the reaction, 11.36 mM KCN was prepared in assay buffer. For each 384-well plate to be screened, 83 μL of 0.1 M NADH was added to 37.6 mL of 11.36 mM KCN/buffer followed by 108 μL of 1:1 membrane (8 mg/mL). The glass container was then swirled to ensure thorough mixing of the assay mixture, and 88 μL was added to each well containing 10 μL of compound using a Matrix multichannel pipet. A “pre-read” 340 nm absorbance value was then obtained prior to the addition of 2 μL of 1 mM CoQ using a Velocity11 Bravo to initiate the reaction. The reaction was then mixed (20 μL × 3 cycles fast speed), and a “post-read” 340 nm absorbance value was obtained 1 min after reaction initiation. The final assay concentrations were

therefore 200 μM NADH, 10 mM KCN, 1 μg/well membrane, 20 μM CoQ, and 5 μM HDQ. The assay mixture was only stable for a few minutes (NADH oxidizes when in contact with the membrane) and was therefore prepared fresh for each plate where a 1 min reaction time was used or to support a batch of three plates where 3 min of reaction time was used. Because the reaction was not stopped, it was important to keep the timings very tightly controlled.

HDQ was used as the assay standard. A fresh weighing of lyophilized HDQ was dissolved in MeOH to a concentration of 1 mM. This stock was stored at –20 °C and used for a maximum of 3 days. Following data collection, the results were analyzed, and IC₅₀ was calculated in GraphPad Prism (GraphPad Software, Inc., United States) using nonlinear regression followed by a sigmoidal dose–response calculation.⁴⁹

Preliminary assaying of the 1175 set of compounds was performed in replicate five-point concentration curves with a 4-fold dilution step in 100% DMSO. The 1 μL stock concentration curve was diluted 25-fold by the addition of 24 μL of buffer to give a top concentration of 200 μM in 4% DMSO, that is, 20 μM final in the assay. Plates were formatted such that test compounds were present in columns 3–22; columns 1, 2, 23, and 24 contained 4% DMSO with the exception of wells M-P1 and A-D24, which contained 50 μM HDQ as a positive control for tracking purposes. Columns 1 (A–L) and 24 (E–P) contained the no-CoQ positive control used for data calculations; instead of 2 μL of 1 mM CoQ, these wells received 2 μL of the carrier (22.5% DMSO/10% EtOH). The final well contents for phase 2 screen were 20 μM top concentration of compound, 20 μM CoQ (or diluent for the 100% inhibition controls), 0.85% DMSO, and 0.2% EtOH.

Assaying of the 16000 compounds was performed at 20 μM for both the primary and the hit confirmation screens. Plates were formatted such that test compounds were present in columns 1–22, and columns 23 and 24 contained 4% DMSO with the exception of wells A-D24, which contained 50 μM HDQ as a positive control for tracking purposes. Column 24 (E-P) contained the positive control used for data calculations; these wells received 2 μL of the carrier (22.5% DMSO/10% EtOH) instead of 2 μL of 1 mM CoQ. The final well contents for phase 4 screen were approximately 20 μM compound, 20 μM CoQ (or diluent for the 100% inhibition controls), 0.85% DMSO, and 0.2% EtOH.

The potency determination on the compounds of interest was performed as 10-point concentration–response curves with a 3-fold serial dilution step. Plates were formatted such that test compounds were present in columns 3–22, and columns 1, 2, 23, and 24 contained 4% DMSO with the exception of wells M-P1 and A-D24, which contained 50 μM HDQ as a positive control for quality control purposes. Each plate also contained a 10-point concentration–response curve of HDQ which had a final assay top concentration of 5 μM in 0.5% MeOH; the curve was contained in rows O13–22 and P13–22. Columns 1 (A–L) and 24 (E–P) contained the positive control used for data calculations, and these wells received 2 μL of carrier (22.5% DMSO/10% EtOH) instead of 2 μL of 1 mM CoQ. The final well contents for potency determination were 40 μM top concentration of compound, 20 μM CoQ (or diluent for 100% inhibition controls), 1.25% DMSO, and 0.2% EtOH.

Data were recorded as absorbance units prior to initiation of the reaction and after reaction completion. The pre- and postreaction absorbance values were imported into Genedata Screener Assay Analyzer (Genedata Inc., Switzerland), and having defined the control populations, the % inhibition was calculated for each well. Control wells used in the calculations did not receive compound (but were diluent controlled), and % inhibition was calculated using wells that had not received CoQ as 100% inhibition. Five micromolar HDQ wells were monitored to ensure that a high % inhibition was being achieved. The option of correcting the data for plate effects using Assay Analyzer's runwise multiplicative correction factor (Genedata Inc.) was investigated, and this data set was taken into account for screen hit calling. The concentration–response data were processed through Genedata Screener as described, then calculated, and plotted using XLFit 4.2.1. (ID Business Solutions, Guildford, United Kingdom).

Curve fitting was carried out using a 4-point parameter logistic equation, model 204.⁵⁰

Cheminformatic Methods. All of the methods below were performed in parallel on the BioFocus compound collection.

Substructure Search. The BioFocus compound collection of ~750000 compounds was searched for compounds that contained the “key moiety” (Figure 1) of HDQ that we chose as 1*H*-quinolin-4-one using Pipeline Pilot.⁵¹

Molecular Fingerprints. Three types of molecular fingerprints were calculated using Pipeline Pilot.⁵¹ MDL MACCS keys,²⁵ FCFP2, and ECFP2.²⁶ The similarity between the 55 active compounds and the BioFocus library compounds was assessed using the Tanimoto coefficient.

Turbo Similarity. The turbo similarity algorithm³⁰ was implemented in Pipeline Pilot.⁵¹ Using the 55 active compounds, ECFP4²⁶ were used as descriptors, and for each search, 250 of the highest ranking compounds were kept using the BioFocus database. The 5000 top ranked compounds were selected.

PCA. A PCA model was built using 20 descriptors in Pipeline Pilot.⁵¹ The descriptors calculated were ALog P, Molecular_Weight, Num_Atoms, Num_Bonds, Num_Hydrogens, Num_PositiveAtoms, Num_NegativeAtoms, Num_RingBonds, Num_RotatableBonds, Num_AromaticBonds, Num_BridgeBonds, Num_Rings, Num_AromaticRings, Num_RingAssemblies, Num_Chains, Num_ChainAssemblies, Num_Fragments, Num_H_Acceptors, Num_H_Donors, and Molecular_Solubility. The 5000 closest compounds to any of the hits as assessed by Euclidean distance in PCA were selected from the BioFocus database using Pipeline Pilot.

Bayesian Modeling. Bayesian modeling was achieved using seven descriptors in Pipeline Pilot.⁵¹ The descriptors calculated were as follows: ALog P, Molecular_Weight, Num_H_Donors, Num_H_Acceptors, Num_RotatableBonds, ECFP_2, and Molecular_PolarSurfaceArea. The model was developed using leave-one-out cross-validation and applied to the BioFocus database with compounds with a score of 5 or above being selected.

Bioisosteres. Bioisosteres were identified using PowerMV³³ using a variety of descriptors, similarity measures, and databases: Atom pair/Tanimoto, atom pair (Cahart)/Tanimoto, fragment pair/Tanimoto, pharmacophore fingerprints/Tanimoto, and weighted Burden number/Euclidean methods were used against the ACL, Gene Logic, and NCI databases within PowerMV. The bioisosteres identified were used as substructure searches to identify compounds in the BioFocus database using Pipeline Pilot.⁵¹

Scoring. The scoring functions were implemented in Pipeline Pilot using the built in algorithms for molecular solubility³⁹ and log P.⁵²

Diversity Selection. BCUT descriptors⁴⁰ were calculated using the CDK descriptor calculator.⁵³ A total of 16000 diverse compounds were selected using a protocol in Pipeline Pilot.⁵¹

Diversity Assessment. ECFP4 fingerprints, using a similarity cut off of 0.7 with the Tanimoto coefficient, were used together with clustering to compare the coverage of the chemical space of the 16000 selected compounds within in the entire BioFocus library using protocols in Pipeline Pilot.⁵¹

Parasite Culture. *Plasmodium* blood stage cultures⁵⁴ and drug sensitivity⁵⁵ were determined by established methods. IC₅₀ values (50% inhibitory concentrations) were calculated by using the four-parameter logistic method (Grafit program; Erithacus Software, United Kingdom).

■ ASSOCIATED CONTENT

● Supporting Information

Proof of dose–response behavior (PfNDH2 and 3D7), description of controls in the HTS, ¹H NMR and MS data for the compounds in Table 4, and the ranks of the compounds selected. This material is available free of charge via the Internet at <http://pubs.acs.org>.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +44(0)151 794 3877. Fax: 44(0)151 794 3588. E-mail: ngberry@liv.ac.uk. URL: <http://www.liv.ac.uk/~ngberry>.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by grants from the Leverhulme Trust, the Wellcome Trust, National Institute of Health Research (NHIR, BRC Liverpool), and EPSRC.

■ ABBREVIATIONS USED

HTS, high throughput screen; Pf, *Plasmodium falciparum*; HDQ, 1-hydroxy-2-dodecyl-4-(1*H*)quinolone; ADMET, absorption distribution metabolism excretion toxicity; ECFP, extended connectivity fingerprints; FCFP, functional class fingerprints; MACCS, molecular access system; PCA, principal component analysis

■ REFERENCES

- (1) WHO. Malaria; *Fact Sheet Number 94*; WHO: Geneva, Switzerland, 2009.
- (2) Turschner, S.; Efferth, T. Drug Resistance in Plasmodium: Natural Products in the Fight Against Malaria. *Mini-Rev. Med. Chem.* **2009**, *9* (2), 206–214.
- (3) Snow, R. W.; Guerra, C. A.; Noor, A. M.; Myint, H. Y.; Hay, S. I. The global distribution of clinical episodes of *Plasmodium falciparum* malaria. *Nature* **2005**, *434* (7030), 214–217.
- (4) Gallup, J. L.; Sachs, J. D. The economic burden of malaria. *Am. J. Trop. Med. Hyg.* **2001**, *64* (1–2), 85–96.
- (5) Weill, M.; Lutfalla, G.; Mogensen, K.; Chandre, F.; Berthomieu, A.; Berticat, C.; Pasteur, N.; Philips, A.; Fort, P.; Raymond, M. Insecticide resistance in mosquito vectors. *Nature* **2003**, *423* (6936), 136–137.
- (6) White, N. J. Antimalarial drug resistance. *J. Clin. Invest.* **2004**, *113* (8), 1084–1092.
- (7) Fidock, D. A.; Eastman, R. T.; Ward, S. A.; Meshnick, S. R. Recent highlights in antimalarial drug resistance and chemotherapy research. *Trends Parasitol.* **2008**, *24* (12), 537–544.
- (8) Muregi, F. W.; Kirira, P. G.; Ishih, A. Novel rational drug design strategies with potential to revolutionize malaria chemotherapy. *Curr. Med. Chem.* **2011**, *18* (1), 113–143.
- (9) Barton, V.; Fisher, N.; Biagini, G. A.; Ward, S. A.; O'Neill, P. M. Inhibiting Plasmodium cytochrome bc1: A complex issue. *Curr. Opin. Chem. Biol.* **2010**, *14* (4), 440–446.
- (10) (a) Biagini, G. A.; Viriyavejakul, P.; O'Neill, P. M.; Bray, P. G.; Ward, S. A. Functional characterization and target validation of alternative complex I of *Plasmodium falciparum* mitochondria. *Antimicrob. Agents Chemother.* **2006**, *50* (5), 1841–1851. (b) Fisher, N.; Bray, P. G.; Ward, S. A.; Biagini, G. A. The malaria parasite type II NADH:quinone oxidoreductase: An alternative enzyme for an alternative lifestyle. *Trends Parasitol.* **2007**, *23* (7), 305–310.
- (11) Fisher, N.; Warman, A. J.; Ward, S. A.; Biagini, G. A. Type II NADH: Quinone Oxidoreductases of *Plasmodium falciparum* and *Mycobacterium tuberculosis* Kinetic and High-Throughput Assays. *Methods in Enzymology*; Allison, W. S., Scheffler, I. E., Eds.; Academic Press: New York, 2009; Vol. 456, Chapter 17, pp 303–320.
- (12) (a) Reil, E.; Hofle, G.; Draber, W.; Oettmeier, W. Quinolones and their n-oxides as inhibitors of mitochondrial complexes I and III. *Biochim. Biophys. Acta, Bioenerg.* **1997**, *1318* (1–2), 291–298. (b) BohneW.; Gross, U. Use of quinoline derivatives as anti-protozoal agent and in combination preparations. WO 2007131488 A2 20071122, 2007.
- (13) Gardiner, E. J.; Gillet, V. J.; Haranczyk, M.; Hert, J.; Holliday, J. D.; Malim, N.; Patel, Y.; Willett, P. Turbo similarity searching: Effect of fingerprint and dataset on virtual-screening performance. *Stat. Anal. Data Mining* **2009**, *2* (2), 103–114.
- (14) (a) Ripphausen, P.; Nisius, B.; Peltason, L.; Bajorath, J. Quo Vadis, Virtual Screening? A Comprehensive Survey of Prospective Applications. *J. Med. Chem.* **2010**, *53* (24), 8461–8467. (b) Koeppen,

- H. Virtual screening—What does it give us? *Curr. Opin. Drug Discovery Dev.* **2009**, *12* (3), 397–407. (c) Jorgensen, W. L. Efficient Drug Lead Discovery and Optimization. *Acc. Chem. Res.* **2009**, *42* (6), 724–733.
- (15) Clark, D. E. What has virtual screening ever done for drug discovery? *Expert Opin. Drug Discovery* **2008**, *3* (8), 841–851.
- (16) Seifert, M. H. J.; Wolf, K.; Vitt, D. Virtual high-throughput in silico screening. *Biosilico* **2003**, *1* (4), 143–149.
- (17) Muchmore, S. W.; Edmunds, J. J.; Stewart, K. D.; Hajduk, P. J. Cheminformatic Tools for Medicinal Chemists. *J. Med. Chem.* **2010**, *53* (13), 4830–4841.
- (18) (a) Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *J. Pharmacol. Toxicol. Methods* **2000**, *44* (1), 235–249. (b) Veber, D. F.; Johnson, S. R.; Cheng, H. Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular properties that influence the oral bioavailability of drug candidates. *J. Med. Chem.* **2002**, *45* (12), 2615–2623.
- (19) Gleeson, M. P. Generation of a set of simple, interpretable ADMET rules of thumb. *J. Med. Chem.* **2008**, *51* (4), 817–834.
- (20) (a) Hughes, J. D.; Blegg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physicochemical drug properties associated with in vivo toxicological outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18* (17), 4872–4875. (b) Waring, M. J. Lipophilicity in drug discovery. *Expert Opin. Drug Discovery* **2010**, *5* (3), 235–248.
- (21) Schneider, G.; Neidhart, W.; Giller, T.; Schmid, G. 'Scaffold-Hopping' by topological pharmacophore search: A contribution to virtual screening. *Angew. Chem. Int. Ed.* **1999**, *38* (19), 2894–2896.
- (22) Venkatraman, V.; Pérez-Nuño, V. I.; Mavridis, L.; Ritchie, D. W. Comprehensive comparison of ligand-based virtual screening tools against the DUD data set reveals limitations of current 3D methods. *J. Chem. Inf. Model.* **2010**, *50* (12), 2079–2093.
- (23) Gardiner, E. J.; Holliday, J. D.; O'Dowd, C.; Willett, P. Effectiveness of 2D fingerprints for scaffold hopping. *Future Med. Chem.* **2011**, *3* (4), 405–414.
- (24) Baringhaus, K. H.; Hessler, G. Fast similarity searching and screening hit analysis. *Drug Discovery Today: Technol.* **2004**, *1* (3), 197–202.
- (25) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci.* **2002**, *42* (6), 1273–1280.
- (26) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50* (5), 742–754.
- (27) Holliday, J. D.; Hu, C. Y.; Willett, P. Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2D fragment bit-strings. *Comb. Chem. High Throughput Screening* **2002**, *5* (2), 155–166.
- (28) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity? *J. Med. Chem.* **2002**, *45* (19), 4350–4358.
- (29) Muchmore, S. W.; Debe, D. A.; Metz, J. T.; Brown, S. P.; Martin, Y. C.; Hajduk, P. J. Application of belief theory to similarity data fusion for use in analog searching and lead hopping. *J. Chem. Inf. Model.* **2008**, *48* (5), 941–948.
- (30) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11* (23–24), 1046–1053.
- (31) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azaoui, K.; Jacoby, E.; Schuffenhauer, A. Enhancing the effectiveness of similarity-based virtual screening using nearest-neighbor information. *J. Med. Chem.* **2005**, *48* (22), 7049–7054.
- (32) Bender, A.; Glen, R. C. Molecular similarity: A key technique in molecular informatics. *Org. Biomol. Chem.* **2004**, *2* (22), 3204–3218.
- (33) Liu, K.; Feng, J.; Young, S. S. PowerMV: A software environment for molecular viewing, descriptor generation, data analysis and hit evaluation. *J. Chem. Inf. Model.* **2005**, *45* (2), 515–522.
- (34) Geppert, H.; Vogt, M.; Bajorath, J. Current trends in ligand-based virtual screening: molecular representations, data mining methods, new application areas, and performance evaluation. *J. Chem. Inf. Model.* **2010**, *50* (2), 205–216.
- (35) Melville, J. L.; Burke, E. K.; Hirst, J. D. Machine learning in virtual screening. *Comb. Chem. High Throughput Screening* **2009**, *12* (4), 332–343.
- (36) Vogt, M.; Stumpfe, D.; Geppert, H.; Bajorath, J. Scaffold Hopping Using Two-Dimensional Fingerprints: True Potential, Black Magic, or a Hopeless Endeavor? Guidelines for Virtual Screening. *J. Med. Chem.* **2010**, *53* (15), 5707–5715.
- (37) Baber, J. C.; Shirley, W. A.; Gao, Y.; Feher, M. The use of consensus scoring in ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (1), 277–288.
- (38) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **2000**, *46* (1–3), 3–26.
- (39) Tetko, I. V.; Tanchuk, V. Y.; Kasheva, T. N.; Villa, A. E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (6), 1488–1493.
- (40) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9–11*, 339–353.
- (41) Zhang, J. H.; Chung, T. D. Y.; Oldenburg, K. R. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J. Biomol. Screening* **1999**, *4* (2), 67–73.
- (42) Genedata, M., 4053 Basel, Switzerland.
- (43) Svensson, F.; Karlen, A.; Skold, C. Virtual screening data fusion using both structure- and ligand-based methods. *J. Chem. Inf. Model.* **2012**, *52* (1), 225–232.
- (44) Gamo, F.-J.; Sanz, L. M.; Vidal, J.; de Cozar, C.; Alvarez, E.; Lavandera, J.-L.; Vanderwall, D. E.; Green, D. V. S.; Kumar, V.; Hasan, S.; Brown, J. R.; Peishoff, C. E.; Cardon, L. R.; Garcia-Bustos, J. F. Thousands of chemical starting points for antimalarial lead identification. *Nature* **2010**, *465* (7296), 305–US6.
- (45) Guigemede, W. A.; Shelat, A. A.; Bouck, D.; Duffy, S.; Crowther, G. J.; Davis, P. H.; Smithson, D. C.; Connelly, M.; Clark, J.; Zhu, F.; Jimenez-Diaz, M. B.; Martinez, M. S.; Wilson, E. B.; Tripathi, A. K.; Gut, J.; Sharlow, E. R.; Bathurst, I.; El Mazouni, F.; Fowble, J. W.; Forquer, I.; McGinley, P. L.; Castro, S.; Angulo-Barturen, I.; Ferrer, S.; Rosenthal, P. J.; DeRisi, J. L.; Sullivan, D. J. Jr.; Lazo, J. S.; Roos, D. S.; Riscoe, M. K.; Phillips, M. A.; Rathod, P. K.; Van Voorhis, W. C.; Avery, V. M.; Guy, R. K. Chemical genetics of *Plasmodium falciparum*. *Nature* **2010**, *465* (7296), 311–315.
- (46) Meister, S.; Plouffe, D. M.; Kuhen, K. L.; Bonamy, G. M. C.; Wu, T.; Barnes, S. W.; Bopp, S. E.; Borboa, R.; Bright, A. T.; Che, J.; Cohen, S.; Dharia, N. V.; Gagaring, K.; Gettayacamin, M.; Gordon, P.; Groessl, T.; Kato, N.; Lee, M. C. S.; McNamara, C. W.; Fidock, D. A.; Nagle, A.; Nam, T.-g.; Richmond, W.; Roland, J.; Rottmann, M.; Zhou, B.; Froissard, P.; Glynne, R. J.; Mazier, D.; Sattabongkot, J.; Schultz, P. G.; Tuntland, T.; Walker, J. R.; Zhou, Y.; Chatterjee, A.; Diagona, T. T.; Winzeler, E. A. Imaging of Plasmodium Liver Stages to Drive Next-Generation Antimalarial Drug Discovery. *Science* **2011**, *334* (6061), 1372–1377.
- (47) Leung, S. C.; Gibbons, P.; Amewu, R.; Nixon, G. L.; Pidathala, C.; Hong, W. D.; Pacorel, B.; Berry, N. G.; Sharma, R.; Stocks, P. A.; Srivastava, A.; Shone, A. E.; Charoensutthivarakul, S.; Taylor, L.; Berger, O.; Mbekeani, A.; Hill, A.; Fisher, N. E.; Warman, A. J.; Biagini, G. A.; Ward, S. A.; O'Neill, P. M. Identification, design and biological evaluation of heterocyclic quinolones targeting plasmodium falciparum type II NADH:quinone oxidoreductase (PfNDH2). *J. Med. Chem.* **2012**, *55* (5), 1844–1857.
- (48) Pidathala, C.; Amewu, R.; Pacorel, B.; Nixon, G. L.; Gibbons, P.; Hong, W. D.; Leung, S. C.; Berry, N. G.; Sharma, R.; Stocks, P. A.; Srivastava, A.; Shone, A. E.; Charoensutthivarakul, S.; Taylor, L.; Berger, O.; Mbekeani, A.; Hill, A.; Fisher, N. E.; Warman, A. J.; Biagini, G. A.; Ward, S. A.; O'Neill, P. M. Identification, design and biological evaluation of bisaryl quinolones targeting plasmodium falciparum type II NADH:quinone oxidoreductase (PfNDH2). *J. Med. Chem.* **2012**, *55* (5), 1831–1843.
- (49) *GraphPad Software*, I., La Jolla, CA.

(50) ID Business Solutions Ltd., O. C., Surrey Research Park, Guildford, Surrey, GU2 7QB, United Kingdom.

(51) <http://accelrys.com/products/pipeline-pilot/>.

(52) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. Prediction of hydrophobic (lipophilic) properties of small organic molecules using fragmental methods: An analysis of ALOGP and CLOGP methods. *J. Phys. Chem. A* **1998**, *102* (21), 3762–3772.

(53) Guha, R. (accessed 27/07/2011).

(54) Trager, W.; Jensen, J. B. Human malaria parasites in continuous culture. *Science* **1976**, *193* (4254), 673–675.

(55) Smilkstein, M.; Sriwilaijaroen, N.; Kelly, J. X.; Wilairat, P.; Riscoe, M. Simple and inexpensive fluorescence-based technique for high-throughput antimalarial drug screening. *Antimicrob. Agents Chemother.* **2004**, *48* (5), 1803–1806.