A Novel *In Silico* Approach to Drug Discovery via Computational Intelligence

David Hecht*,† and Gary B. Fogel‡

Southwestern College, 900 Otay Lakes Road, Chula Vista, California 91910, and Natural Selection, Inc.,
9330 Scranton Road, Suite 150, San Diego, California 92121

Received February 23, 2009

A computational intelligence drug discovery platform is introduced as an innovative technology designed to accelerate high-throughput drug screening for generalized protein-targeted drug discovery. This technology results in collections of novel small molecule compounds that bind to protein targets as well as details on predicted binding modes and molecular interactions. The approach was tested on dihydrofolate reductase (DHFR) for novel antimalarial drug discovery; however, the methods developed can be applied broadly in early stage drug discovery and development. For this purpose, an initial fragment library was defined, and an automated fragment assembly algorithm was generated. These were combined with a computational intelligence screening tool for prescreening of compounds relative to DHFR inhibition. The entire method was assayed relative to spaces of known DHFR inhibitors and with chemical feasibility in mind, leading to experimental validation in future studies.

INTRODUCTION

Perhaps the greatest source of inefficiency in traditional drug discovery and development arises from the high percentage of evaluated compounds that have a low probability of ultimate success. To address the high attrition rate in lead optimization and early preclinical development processes, "focused" or "enriched" compound libraries are routinely generated in a virtual environment.[1,2] The selected compounds are then synthesized/purchased and experimentally screened. Each compound in the focused library is selected after considering structural and physical properties that will increase its probability of having activity for the specific target as well as its ultimate survivability through preclinical development. Quantitative structure−activity relationship (QSAR) model(s) and/or docking experiments are most often used for selection of these focused compound libraries.[3]

De novo ligand design methods are computational methods for designing molecules that complement a receptor or binding site structurally and/or energetically.[4] Successful *de novo* design results in proposed ligand structures that have high binding affinity for their desired binding sites. *De novo* design has been most successful where biological and experimental knowledge of the ligands and substrates exists. Two major approaches have been applied to the development of *de novo* ligand design software: molecular fragment approaches[5] and sequential growth approaches.[6]

The molecular fragment approaches dock molecular fragments to determine various energetically favorable positions on the active (binding) site that are then "linked" together. The first step is to identify key locations in the binding pocket and then bind small seeds or fragments to these locations. Once the seed fragments/functional groups are positioned, the next step is to link them together with scaffolds. Previous attempts in this area of research include the following: CONCERTS,[7] LUDI,[8,9] CAVEAT,[10] and NEWLEAD,[11] DLD,[12] BUILDER,[13] and SKELGEN.[14,15] In all of these approaches, optimization steps must occur where bonds can be broken and formed and scoring must take place. Monte Carlo or evolutionary algorithms are often used for optimization (e.g., Pro-Ligand,[16] ADAPT,[17] and Leapfrog[18]).

For sequential growth approach, molecules are "grown" into an active site starting from a seed (e.g., a small molecule or fragment) bound to the active site. The ligand grows atom by atom or fragment-by-fragment to complement the active site geometrically as well as energetically (e.g., electrostatics, van der Waals, "hydrophobicity").[19−21]

The "growth" approach also requires the docking or binding of a seed molecule or fragment to the active site. This is essentially the same process as the first step in the molecular fragment approach described above. Once positioned, the ligand is built via the sequential addition of atoms or fragments. Some of the earliest examples of this approach include the following: Legend[19] and Genstar[22] which grow ligands one atom at a time, SmoG,[23,24] GrowMol[21] (which use single atoms as well as functional groups), GroupBuild,[25] SPROUT[26] (which use fragments to grow molecules), and GROW[27] which builds peptides one amino acid at a time. At each additional step in the growth process, there is a selection process with scoring used to accept or reject the modifications.

Current *de novo* ligand design methodologies suffer from one or more of three major deficiencies that have severely limited their use and acceptance in drug discovery programs. The first and most important deficiency is the fact that a large number of the generated structures are synthetically unfeasible. This is particularly evident for the fragment methods where in many cases it is chemically infeasible to bridge the bound functional groups and join all fragments in their most favorable locations.

* Corresponding author phone: (619)421-6700; e-mail: dhecht@swccd.edu.
† Southwestern College.
‡ Natural Selection, Inc.

10.1021/ci9000647 CCC: $40.75
Published on Web 04/06/2009

The second major deficiency arises from the commonly observed differences in experimental and calculated binding affinities. *De novo* design methods utilize a scoring function to evaluate each step of the process. This scoring function is really a calculated measure of receptor−ligand binding affinity. Unfortunately, available scoring functions are limited in their abilities to accurately model/predict experimentally determined binding affinities and activities.[28,29] The third deficiency arises from the large combinatorial problem of quickly and efficiently searching diversity space for good solutions (e.g., structures with reasonable binding affinity).[25]

As a result of these limitations of *de novo* design, companies have turned to docking programs to screen small molecule, commercially available libraries.[1−3] A major drawback of these approaches is the limited chemical diversity available for *in silico* screening. Many relevant regions of structure space are simply not available in these screening libraries. *De novo* design has an advantage in its ability to efficiently cover a larger portion of structure space for a particular binding site.[21]

"Fragment" compound screening libraries are often used to address the concern that many of the "validated" hits arising from traditional HTS of commercially available compound libraries do not advance far beyond one or two rounds of lead optimization.[30] Primary causes for this lack of success include issues concerning "the Rule of 3", aqueous and plasma solubilities, and stabilities as well as the ability to permeate cell and intestinal membranes.[30,31] Although the fragments will each have lower affinities (e.g., high to low μM K_i or K_d values) for the target protein than the larger compounds found in the commercially available screening libraries, they will often have better physicochemical properties (e.g., "the Rule of 3"). This approach takes advantage of the exponential increase in potency that is often found when low affinity fragments are joined together through a linking region or one or more scaffolds or structural templates. While this experimental screening approach captures many of the advantages of *de novo* design, synthetic accessibility remains a major concern.

To address the issue of synthetic accessibility, combinatorial docking as well as *de novo* design methods have been created which filter for synthetic accessibility.[32−34] PRO_SELECT is an example which uses commercially available combinatorial library scaffolds and components.[35] A template is first placed in the active site with multiple attachment points. Functional group substituents are selected from databases for each attachment site based using the PRO_LIGAND *de novo* design package mentioned previously. A great deal of effort is spent on optimizing the positions and conformations of the modified templates and substituents. Additional filters are used in the selection process that include the following: molecular weight, calculated log P, number of atoms, and number of rotational bonds. The end product of PRO_SELECT is a ranked, small focused combinatorial library that is synthetically accessible for experimental evaluation.

In order to address the issue of quickly and efficiently searching diversity space for "good" solutions, *de novo* design approaches often integrate tools and techniques borrowed from the field of computational intelligence. Computational intelligence is a broad field that focuses on the development of machine learning approaches for the automatic selection of features and the optimization of models and includes tools and techniques such as artificial neural networks (ANNs), fuzzy logic, and evolutionary computation.[36] Some recent examples of applications to *de novo* design include the use of evolutionary algorithms to design peptidic[37−39] and nonpeptidic ligands for protein[16−18,40−44] as well as RNA[45] targets.

In this paper we present an approach to use computational intelligence for accelerated high-throughput drug screening and generalized protein-targeted drug discovery. This technology integrates various tools and techniques including evolutionary algorithms, evolved artificial neural nets,[46−48] and docking software as well as quantitative structure activity/property relationship (QSAR/QSPR) modeling. This approach not only results in collections of novel, "druglike," synthetically accessible, small molecule compounds predicted to bind to protein targets but also provides details on predicted binding modes and molecular interactions.

The approach was tested on dihydrofolate reductase (DHFR) for novel antimalarial drug discovery; however, the methods developed can be applied broadly in early stage drug discovery and development. For this purpose, an initial fragment library was defined, and an automated fragment assembly algorithm was generated. These were combined with a computational intelligence screening tool for prescreening of compounds relative to DHFR inhibition. The entire method was assayed relative to spaces of known DHFR inhibitors and with chemical feasibility in mind, leading to NMR validation in future studies.

The overall discovery pipeline envisioned for antimalarial drug discovery (Figure 1) begins with a small molecule fragment library derived from published DHFR inhibitors as well as commercially available compound libraries. These candidates serve as templates with various functional groups representing the small molecule space that is to be explored. The choice of candidates is made with respect to protecting groups, patentability, and clear and straightforward synthetic assembly pathways.

Candidate fragments are further filtered with respect to cost, the "Rule of 3",[31] predicted aqueous and plasma solubilities and stabilities,[49] and calculated partition coefficients as well as predicted cell membrane and intestinal permeabilities,[50−52] synthetic accessibility for fragment assembly, and structural diversity. The fragments and scaffolds are first assembled and then prefiltered using evolved neural networks based on predicted physicochemical properties and descriptors. Surviving compounds are screened for their binding affinity using *in silico* docking. The top scoring compounds are used to generate new compounds from the fragment library. After several rounds of optimization, the top scoring compounds can be synthesized and/or purchased from a vendor and tested in functional yeast assays. These experimental results are used to update the evolved neural nets through retraining. After several rounds of optimization through this discovery cycle, top scoring compounds will be selected for binding assays and NMR studies of binding modes to *E. coli* DHFR-TS.

This process of small molecule variation (fragment assembly) and selection (docking and/or assays) is repeated until putative lead compounds with sufficient binding affinities to both wild type and quadruple mutant forms of DHFR are discovered (see below and Figure 2). Penalties

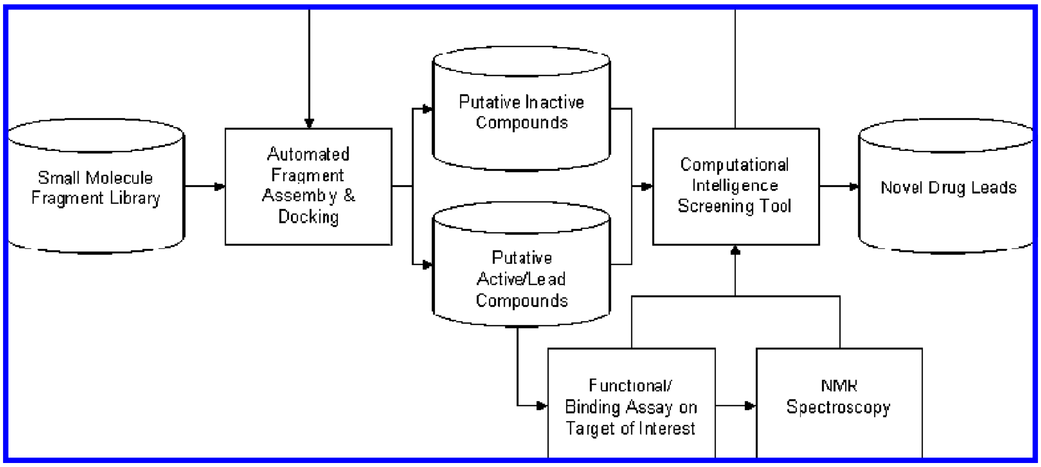

Small Molecule Fragment Library
Automated Fragment Assemby & Docking
Putative Inactive Compounds
Putative Active/Lead Compounds
Computational Intelligence Screening Tool
Novel Drug Leads
Functional/ Binding Assay on Target of Interest
NMR Spectroscopy

Figure 1. Overall pipeline for drug discovery.

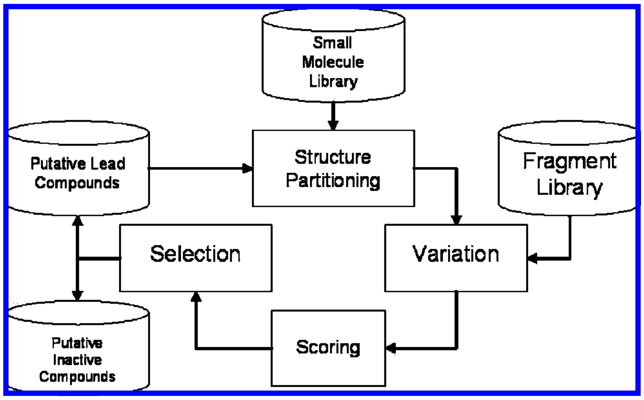


Figure 2. Schematic of automated fragment assembly using evolutionary computation.

for discovering previously known or patented compounds (and subfragments) can be applied so that discovery across the small molecule space is strictly novel and/or commercially viable. Additional penalties can also be applied to remove compounds with unfavorable predicted pharmacokinetic/dynamic/toxicological properties. Throughout this process, all screened molecules (either putative inactives or lead compounds) and their binding affinities are used as training data for a computational intelligence prescreening tool that can help identify active and inactive compounds prior to screening, saving valuable time and money.

As a suitable test of the software components of this system, compounds for the fragment library were generated from known DHFR inhibitors combined with an assortment of building blocks and templates. Each fragment in the library was scored using 1) *in silico* docking experiments and 2) artificial neural networks that combined small molecule fragment features such as experimental data (e.g., K_i, IC_{50}, K_d) with *in silico* docking experiments to produce a measure of DHFR inhibition. A series of computational experiments was conducted starting with a fragment library from known inhibitors in order to determine if the approach could "rediscover" known and/or novel DHFR inhibitors. In future studies we plan to use the resulting focused libraries from this study as a starting place for discovery of novel antimalarial compounds.

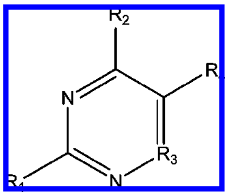


Figure 3. A typical combinatorial template (based on known (*Pf*) DHFR-TS inhibitors) with position dependent R-groups.[54−56]

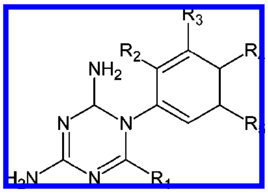


Figure 4. Template 1 with substituent (R-group) positions.[54−56]

MATERIALS AND METHODS

Scaffold and R-Group Selection and Initial Library Design. Three initial fragment libraries were generated by exploring compound libraries from multiple sources in the literature based on the core template common to the (*Pf*) DHFR-TS inhibitors: WR99210, pyrimethamine, and cycloguanil[53,54] (Figure 3). These libraries were decomposed *in silico* into scaffolds and R-group fragments using MOE (Chemical Computing Group, Inc., www.chemcomp.com). The first scaffold and associated R-groups, Template 1, derived from 56 published structures of pyrimethamine and its analogues[54−56] (Figure 4 and Table 1). The second, Template 2, derived from 34 published structures of cycloguanil[54−56] (Figure 5 and Table 2). The third, Template 3, derived from published pyrimidine compounds[57,58] (Figure 6 and Table 3). In the design of this third library, R-groups fragments from Templates 1 and 2 were added in to enhance diversity and the size of the structure space to be sampled.

The space of possible structures from the first two designed libraries was small, on the order of 10^2 to 10^3. The third library was much larger, on the order of 10^{12} possible compounds. An artificial neural network (described in more detail below) was applied as a filter to direct the evolutionary search by prescreening compounds for their activity. Neural network filters included predicted aqueous and plasma solubilities and stabilities,[49] calculated partition coefficients

Table 1. Position-Dependent R-Groups as Smiles Strings for Template 1[54−56]

R1	R2	R3	R4	R_5
C	Cl	Br	Br	Br
C(C)(C)C	Br	Cl	C	Cl
C(CCC)CC		OC	C(C)(C)C	
c1ccccc1		OCO	c1ccccc1	
CC			Cl	
O(C(=O)CCC)C			Clc1cc(ccc1Cl)CO	
O(C)c1ccc(cc1)CCC			OC	
O(CCC)COC			OCO	
O(CCCOCC)c1ccccc1				
O(COCCC)c1ccccc1				
OCc1ccc(cc1)CCC				
OCc1ccccc1				
OCCC				

and permeabilities,[50−52] synthetic accessibility for fragment assembly, and diversity and docking scores from *in silico* docking experiments. Software and tools from MOE, Sybyl (Tripos, Inc., www.tripos.com), Qikprop (Schrodinger, Inc., www.schrodinger.com), GOLD (Cambridge Crystallography Data Centre, www.ccdc.cam.ac.uk), and Molegro Virtual Docker (Molegro ApS, www.molegro.com) were used for this purpose.

From these lists of templates and R-groups, sets of smiles strings for each generation were constructed. The smile strings were then converted into chemical structures and exported into a *.sdf file using Chemfinder for Microsoft Excel (Cambridgesoft, Inc., www.cambridgesoft.com) as well as Accord for Excel (Accelrys, Inc., www.accelrys.com). In MOE, the "wash" function was used on the 2D structures to standardize bond lengths and angles. 3D structures were then generated and minimized using the MMFF94x potential.[59] This potential has been parametrized for gas phase small organic molecules in medicinal chemistry. The resulting minimized structures were then exported as a *.sdf file for docking in GOLD.

Docking Protocols. The X-ray crystal structures of both *Pf* WT DHFR-TS (1J3I.pdb)[53] and quadruple mutant DHFR-TS (1J3K.pdb)[53] were obtained from the literature. Both of these structures contain the third-generation *Pf*-DHFR inhibitor WR99210 bound to the active site in the presence of NADPH and are assumed to be representative of bound conformations *in vivo*.

We used Deepview (Swiss-PDBviewer) to remove all waters as well as other nonprotein, nonligand, or noncofactor molecules which did not participate in ligand binding and were located on the protein surface far from the active site. In order to simplify docking calculations, the DHFR crystal structures were truncated to a 10 Å radius from each atom in the bound WR99210 inhibitor. The inhibitor WR99210 was then removed, the NADPH cofactor was retained, and the *.pdb file was imported into MOE. The "Wash" function in MOE was used to refine the structure including addition of explicit hydrogen atoms. Lastly, the protein structures were exported as a *.pdb file for input into GOLD.

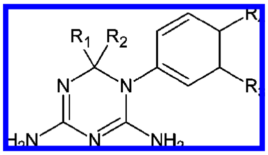


Figure 5. Template 2 with substituent (R-group) positions.[57,58]

Docking experiments were performed using the default GOLD fitness function (VDW = 4.0, H-bonding = 2.5) and default GOLD evolutionary parameters: population size = 100; selection pressure = 1.1; # operations = 100,000; # islands = 5; niche size = 2; migration = 10; mutation = 95; crossover = 95.[31] The carbonyl oxygen on Leu 164 (for 1J3K.pdb) and Ile 164 (for WT 1J3I.pdb) were selected as the binding site centers for all calculations. Ten docking runs were performed per structure unless 3 of the 10 poses were within 1.5 Å rmsd of each other. All poses were output into a single *.sdf file. A single desktop computer operating with WinXP Media edition (AMD 3800+, 64 bit processor (~2.65 GHz) and 1GB RAM) processed 25 compounds in ~3 h. The output from GOLD was imported into Molegro Virtual Docker for scoring. We previously evaluated several different scoring functions and selected the Protein−Ligand Interaction score from Molegro.[60,61]

Pose Validation and Evaluation. In order to verify that poses resulting from *in silico* docking represent correctly bound conformations, each pose was inspected visually and compared to the experimentally determined binding modes and conformations of WR99210. Key protein−ligand contacts and interactions were examined. These included hydrogen bonding interactions with Asp 54, Ile 14, and potentially Leu or Ile 164 as well as potential interactions with Ile 112 and Pro 113.[60,61]

Evolved Fragment Assembly for Directed Search of Novel Leads. Automated fragment assembly proceeded as follows (Figure 2). An initial generation, Generation 1, of 50 smile strings was generated from random coupling of the template and R-groups. Each ligand in the population was scored, as described above, via docking and was ranked by protein−ligand interaction score. The top 10 small molecules were used as "parents" for the generation of 40 new "offspring" small molecules (4 offspring generated at random from each parent solution) so that the overall population size remained at 50 solutions. This process was repeated iteratively for a total of 10 generations of compound evolution.

Artificial Neural Network Development. For the neural network models, 251 MOE descriptors and 36 Qikprop descriptors were generated from the *.sdf files containing the gas phase minimized structures of each generation. These descriptors were used to generate neural network models of the docking that had taken place for the evolution of each template. For the evolution of neural networks using Template 1 data, 450 structures scored through generations 1 through 9 were used for development. Any compounds that were predicted as "inactive" during Molegro docking were given a score of zero.

RESULTS AND DISCUSSION

As the goal of these initial studies was to validate that the approach could recover synthetically accessible active compounds from a decomposed fragment library of known DHFR inhibitors, the decomposition followed the synthetic strategy and mirrored the original SAR analysis of the published literature.[53−58] To ensure chemical feasibility, the selection of templates and R-groups was limited to the synthesis pathways for a compound vendor.

Template 1 Evolution. The top scoring compounds following 10 generations of *in silico* evolution are presented

Table 2. Position-Dependent R-Groups as Smiles Strings for Template 2[54−56]

R1	R2	R3	R4
C	C	C	C
NOTHING	C(C)C	Cl	Cl
	c2cc(Oc3cc(C(F)(F)(F))ccc3)ccc2	Br	Br
	c2cc(Oc3cc(Cl)cc(Cl)c3)ccc2	F	F
	c2cc(Oc3ccc(Cl)cc3)ccc2	NOTHING	NOTHING
	c2cc(Oc3ccccc3)ccc2		
	c2cc(Occ3ccccc3)ccc2		
	c2(OcccOc3c(Cl)cc(Cl)c(Cl)c3)ccc2(OcccOc3c(Cl)cc(Cl)c(Cl)c3)ccc2		
	c2ccc(Oc3ccccc3)cc2		
	c2ccccc2		
	CCC		
	CCCCCC		
	CCCCCCC		
	Oc2ccc(CCC)cc2		
	NOTHING		

in Table 4. Also presented are the three most similar compounds from the original set of 56 pyrimethamine derivatives (Generation 0). All three of these literature-derived structures are potent inhibitors to the DHFR quadruple mutant. The first round of evolution resulted in small molecule structures that would be expected to be potent relative to pyrimethamine.

An analysis was performed on the compounds resulting from 10 generations to evaluate the diversity of structures generated during the evolution. Tanimoto coefficients for all 556 compounds (including duplicates from the top 10 parents in each generation) were generated relative to pyrimethamine. The mean and standard deviation of the Tanimoto coefficient for each generation was then calculated to give a measure of the distance of the evolved compounds relative to the starting library. Table 5 illustrates that after an initial drop, the mean Tanimoto did not change greatly, but the standard deviation of the Tanimoto coefficient decreased, indicating the compounds were converging on similar structures. This convergence was more pronounced when examining the standard deviations of the Tanimoto coefficients for the top ten compounds in each generation (Table 5).

This preliminary analysis of the evolved fragment assembly software and docking method demonstrated a proof-of-concept of the approach. For this particular template, we searched only a small space of possible R-group substituents and recovered a final population after only 10 generations of evolution that fell within a structural class that is similar to known DHFR inhibitors. In fact, some of the top ten scoring compounds after 9 generations of evolution were discovered as early as generations 2, 3, and 5, indicating the speed at which useful solutions can be identified with this approach. This result demonstrated that despite the variability in the docking as a surrogate for experimental data, the process was able to utilize docking as a fitness metric to arrive at useful solutions.

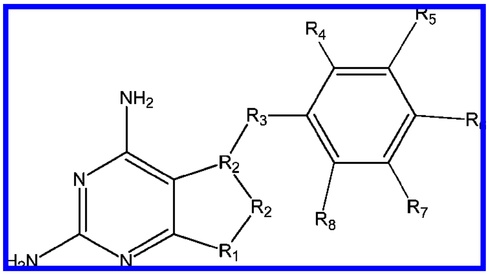


Figure 6. Template 3 with substituent (R-group) positions.

Template 2 Evolution. The top scoring compounds following 10 generations of *in silico* evolution consisted of two structural subclasses (as well as the most similar compounds from the original set of 34 cycloguanil derivatives - Generation 0) are presented in Tables 6 and 7.

The first subclass (Table 6) presents 6 compounds from generation 10 that were analogues of the known potent compound c313. Compound # 2_3_9 (Template #2, generation #3, compound #9) is a recreation of c313, demonstrating that within just 3 generations of evolutionary optimization, this compound had been discovered. During the evolution, we observed redundant structures (in early generations) and introduced a filter during Smiles generation to ensure unique structures throughout the population so that diversity could be maintained throughout the experiment. Table 7 consists of the second subclass of 4 compounds from generation 10 along with their structural analogues from the literature. Note that all are unique with the exception of compound 2_6_13 which is a recreation of c143.

As with the previous experimentation on the pyrimethamine derivatives of Template 1, this round resulted in two subclasses of compound structures expected to be potent. A diversity analysis was performed on the compounds from the 10 generations of evolution for Template 2 to evaluate the diversity of structures generated during the evolution (Table 8). Cycloguanil was used to generate Tanimoto coefficients for all compounds evaluated (including duplicates from the top 10 parents in each generation). As was the case for Template 1, following an initial drop, the mean and standard deviation of the Tanimoto index did not change greatly indicating the compounds had converging on a similar structure class. Again, this convergence was more pronounced when examining the standard deviations of the Tanimoto coefficients for the top ten compounds in each generation.

Computational Screening Tool: Evolved Neural Networks. Using evolution as a search mechanism across small molecule space and by using fragment libraries that are reduced to the space of possible organic synthesis approach serves as an efficient mechanism for searching vast numbers of small molecule assemblies for their potential as antimalarial candidate compounds. However, during the variation process, small molecules were generated with random

Table 3. Position-Dependent R-Groups as Smiles Strings for Template 3[57,58]

R1	R2	R3	R4,R5,R7,R8	R6
C	C=C	C	Br	Br
C(C(C)C)	CC	C[C]	C	C
C(C)		CC	C(C)C	C(C)C
C(c5cc(Oc6(C(F)(F)(F))ccc6)ccc5)(C(F)(F)(F))ccc6)ccc5)		CCC	C(O)=O	C(NC(CCC(O)=O)C(NC([H])C(O)=O)=O)=O
C(c5cc(Oc6 cc(Cl)cc(Cl)c6)ccc5)		CCO	c4cc(Oc5 cm^3(C(F)(F)(F))ccc5)ccc4	C(NC(CCC(O)=O)C(NC(CC4=CC=CC=C4)C(O)=O)=O)=O
C(c5cc(Oc6ccc(Cl)cc6)ccc5)		CCS	c4cc(Oc5 cm^3(Cl)cc(Cl)c5)ccc4	C(NC(CCC(O)=O)C(NC4=CC=C(C#N)C=C4)=O)=O
C(c5cc(Oc6ccccc6)ccc5)		CN	c4 cm^3(Oc5ccc(Cl)cc5)ccc4	C(NC(CCC(O)=O)C(NC4=CC=C(C(O)=O)C=C4)=O)=O
C(c5cc(Occ6ccccc6)ccc5)		CN[C]	c4 cm^3(Oc5ccccc5)ccc4	C(NC(CCC(O)=O)C(O)=O)=O
C(c5cc(OcccOc6c(Cl)cc(Cl)c(Cl)c6)ccc5)		CO	c4cc(Occ5ccccc5)ccc4	C(O)=O
C(c5ccc(Oc6ccccc6)cc5)		CS	c4cc(OcccOc5c(Cl)cc(Cl)c(Cl)c5)ccc4	c4cc(Oc5cc(C(F)(F)(F))ccc5)ccc4
C(c5ccccc5)		C(O)=O	c4ccc(Oc5ccccc5)cc4	c4cc(Oc5cc(Cl)cc(Cl)c5)ccc4
C(CC(C)C)		COC	c4ccccc4	c4cc(Oc5ccc(Cl)cc5)ccc4
C(CCC)		CC(O)=O	C4=CC=CC=C4	c4cc(Oc5ccccc5)ccc4
C(CCCC(=O)OC)		CC	c4c(Cl)c(Cl)ccc4CO	c4cc(Occ5ccccc5)ccc4
C(CCCc5ccc(cc5)OC)		CCCC	CC(C)C	c4cc(OcccOc5c(Cl)cc(Cl)c(Cl)c5)ccc4
C(CCCCCC)			CCC	c4ccc(Oc5ccccc5)cc4
C(CCCCCCC)			CCCCCC	c4ccccc4
C(CCCOCOC)			CCCCCCC	C4=CC=CC=C4
C(CCCOCOc5ccccc5)			Cl	c4c(Cl)c(Cl)ccc4CO
C(CCOCCCOc5ccccc5)			F	CC(C)C
C(COCCCOc5ccccc5)			NOTHING	CCC
C(Oc5ccc(CCC)cc5)			OC	CCCCCC
C(OCc5ccc(cc5)CCC)			Oc2ccc(CCC)cc2	CCCCCCC
C(OCc5ccccc5)			OC4=CC=CC=C4	Cl
N			OCC	F
N(C(C)C)			OCO	NOTHING
N(C)			C(O)=O	OC
N(c5cc(Oc6cc(C(F)(F)(F))ccc6)ccc5)				Oc2ccc(CCC)cc2
N(c5cc(Oc6cc(Cl)cc(Cl)c6)ccc5)				OC4=CC=CC=C4
N(c5cc(Oc6ccc(Cl)cc6)ccc5)				OCC
N(c5cc(Oc6ccccc6)ccc5)				OCO
N(c5cc(Occ6ccccc6)ccc5)				
N(c5cc(OcccOc6c(Cl)cc(Cl)c(Cl)c6)ccc5)				
N(c5ccc(Oc6ccccc6)cc5)				
N(c5ccccc5)				
N(CC(C)C)				
N(CCC)				
N(CCCC(=O)OC)				
N(CCCc5ccc(cc5)OC)				
N(CCCCCC)				
N(CCCCCCC)				
N(CCCOCOC)				
N(CCCOCOC)				
N(CCCOCOc5ccccc5)				
N(CCOCCCOc5ccccc5)				
N(COCCCOc5ccccc5)				
N(Oc5ccc(CCC)cc5)				
N(OCc5ccc(cc5)CCC)				
N(OCc5ccccc5)				
O				
S				

Table 4. Top 5 Compounds Resulting from Ten Generations of *in Silico* Evolution (Right Column; Generation 10) Compared to Similar Structures from the Literature[54−56] (Left Column; Generation 0)

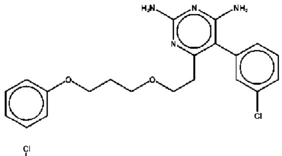
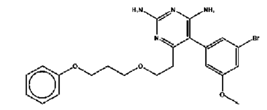
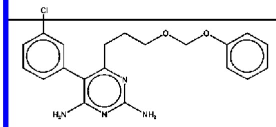
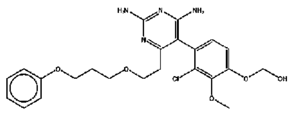
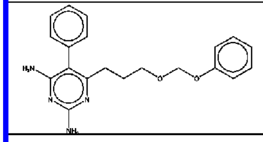
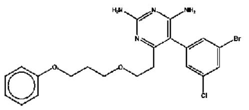
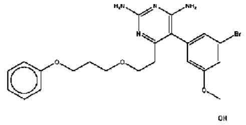
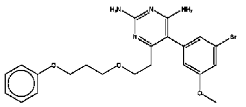
STRUCTURE	ID	pK_i (nM)	Fitness Score	STRUCTURE	ID	Fitness Score
	p40	8.77	-132.479		10_15	-147.552
	p43	8.44	-122.149		10_17	-143.32
	p47	7.85	-119.048		10_20	-142.754
					10_13	-141.508
					10_24	-140.215

Table 5. (a) Mean and Standard Deviation of Tanimoto Coefficients of Each Generation of *in Silico* Evolution vs Pyrimethamine and (b) Mean and Standard Deviation of Tanimoto Coefficients of Top 10 Scoring Compounds of Each Generation vs Pyrimethamine

(a)				(b)			
gen	mean	stdev	*n*	gen	mean	stdev	*n*
0	86.59	6.75	56	0	N/A	N/A	10
1	77.78	6.12	50	1	75.80	5.77	10
2	76.34	6.61	50	2	72.80	6.36	10
3	75.76	7.34	50	3	75.00	5.16	10
4	74.82	7.05	50	4	74.10	7.45	10
5	75.82	7.68	50	5	76.50	3.03	10
6	77.20	5.02	50	6	76.10	3.63	10
7	76.90	5.17	50	7	75.10	3.35	10
8	74.86	5.45	50	8	76.30	4.00	10
9	76.10	4.42	50	9	75.70	2.36	10
10	76.30	4.40	50	10	77.00	2.40	10

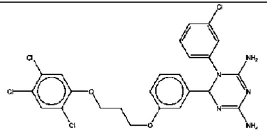
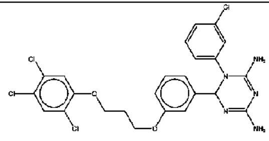
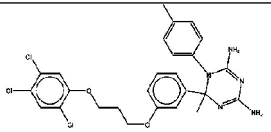
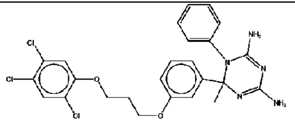
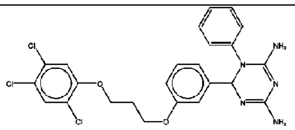
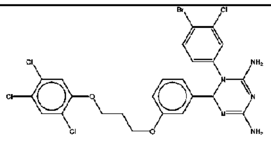
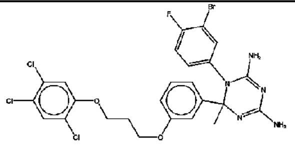
variation upon useful ligands discovered in previous iterations. While this process will work, it makes no use of the stored history of which fragments worked best with what scaffolds to generate potentially active ligands. A model of this process can be used to bias future fragment choices in the direction that worked well in previous rounds of the evolutionary process for the target at hand, further increasing the rate of discovery for novel, synthesizeable, patentable, active compounds. For this purpose we have used artificial neural networks. The overall design method for evolved neural network optimization is shown in Figure 7.

Descriptor Selection. As a first test of the evolved neural networks, we focused on the use of the pyrimethamine analogues identified from the literature.[54−56] As discussed previously, 251 MOE descriptors and 36 Qikprop descriptors were then generated from the *.sdf file containing the gas phase minimized structures. Linear regression was then conducted for each descriptor relative to the experimental pK_i. 74 descriptors had $R^2 > 0.200$ and were considered reasonable for use in developing a best linear model.

Stepwise regression was then used over all 74 descriptors relative to experimental pK_i with the best model having an R^2 of 0.728 and consisting of four descriptors: VDistMa, SMR_VSA5, pmiZ, and ASA (Figure 8). VDistMa is a parameter relative to distance and adjacency of heavy atoms. SMR_VSA5 is the subdivided surface areas based on molar reflectivity and is a common useful parameter for bioavailability. pmiZ describes the principle moment of inertia in the Z direction. ASA is the solvent accessible surface area.

Linear regressions such as this make use of all of the available data to generate a model of best fit. In that regard, the model is not "predictive" in that all of the data are used to generate the model, rather than test the model's accuracy in predicting samples held out of the model. To test the generalizability of the multiple linear regression, leave-one-out cross-validation was performed on this linear model for 1000, 5000, and 10000 generations of evolutionary optimization using all four features as input to a neural network, with no hidden nodes, and one output node (a linearized neural network). The best result from 5000 generations of evolutionary optimization of the weights associated with this model provide an R^2 of 0.614, which is quite lower than the 0.728 observed in the multiple linear regression, but gives some more reasonable measure of worth on held out samples.

Table 6. Results of Template 2 Evolution[a]

Structure	Gen0	pK_i (nM)	Fitness Score	Structure	Gen10	Fitness Score
	c313	8.3467	-119.462		2_3_9	-111.253
					2_4_7	-126.148
					2_9_1	-117.172
					2_9_15	-113.502
					2_8_8	-111.963
					2_10_10	-102.788

[a] Six of the top scoring compounds from generation 10 compared to the single structural analogue found in generation 0.[54−56] Note that 2_3_9 is a duplicate of the original c313.

For the purpose of evolving artificial neural networks, mean squared error (MSE) over the experimentally verified training exemplars was used as a measure of predictive accuracy where MSE was defined as

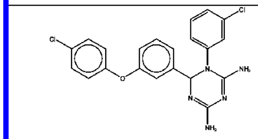
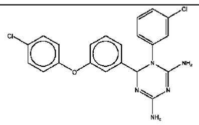
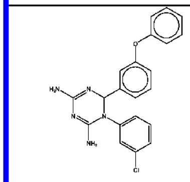
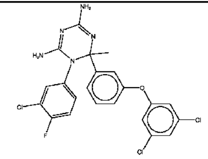
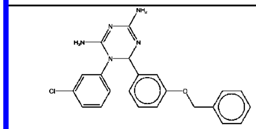
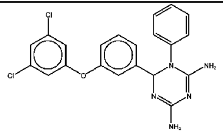
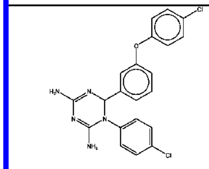
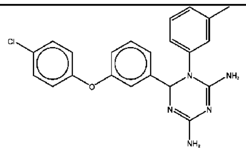
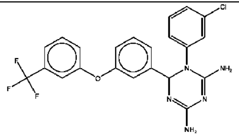
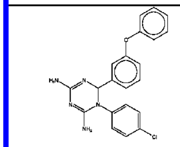
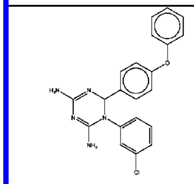
$$MSE = \frac{1}{N}\sum_{k=1}^{N}(P_k - O_k)^2$$

where P was the predicted activity, O was the observed activity, and N was the number of patterns in the training set. Convergence plots of the learning over the training examples provided a means to determine the most appropriate number of generations of evolution without overtraining for maximum performance on the held-out examples.

Leave-one-out cross-validation was used with a population of 50 parents and 50 offspring for generations [50, 5000] with initial sigma 0.1, initial weights 0.0, tournament selection with 4 opponents, all four features as input, with 2 hidden and 1 output nodes, and with normalization of input features to the range [0.1, 0.9]. These parameter choices were chosen in light of previous research using these approaches for QSAR problems. The best model discovered after 300 generations of evolution was slightly improved relative to the best leave-one-out cross-validation of the linear model. Additional investigations using different parameters such as 4 inputs, a range of 2−5 hidden nodes, and 1 output node have failed to identify a superior nonlinear model than the model presented in Table 9.

Template 1 - Pyrimethamine QSAR via Evolved Neural Networks. For the evolution of neural networks using Template 1 data, 450 structures scored through generations 1 through 9 were used for development. Stepwise regression was used to determine a best multiple linear regression for the Template 1 data. However, this new linear regression made use of 8 terms, rather than 4 terms. Using this simple

Table 7. Results of Template 2 Evolution[a]

Structure	Gen0	pK_i (nM)	Fitness Score	Structure	Gen10	Fitness Score
	c143	8.09691	-116.632		2_6_13	-110.885
	c110	8.236572	-117.617		2_10_13	-100.425
	c185	8.113509	-111.261		2_8_36	-119.192
	c188	6.917215	-109.408		2_8_37	-113.027
	c299	8.318759	-108.657			
	c133	6.621602	-101.524			
	c138	8.886057	-97.2156			

[a] Four of the top scoring compounds from generation 10, compared to similar structures from the literature (gen 0).[54−56] Note that compound 2_6_13 is a duplicate of the original c143.

linear model, it would be possible to place a threshold at approximately −90 and separate all predictive inactive compounds from those that were predicted to have high activity; however, many compounds with activity >−90 would, in this case, be considered false negatives for high throughput screening ($R^2 = 0.551$).

Given that the multiple linear regression represents a fitting of all of the data rather than a predictive model applied to a portion of the data that was held out from model development, we evaluated the worth of this linear model using neural networks wherein the 8 inputs were connected directly with the output node. This linear representation was optimized using evolutionary computation to adjust only the weights relative to training examples with performance evaluated on a held-out testing sample. Leave-one-out cross-validation was applied over all compounds, and the model's predictive performance was assayed relative to the held-out samples. The best neural network was generated after 5000 generations of optimization and had a correlation of 0.496 on the testing samples, which is lower than the 0.551 demonstrated when fitting all of the available data and more reasonable in terms of an estimate of predictive performance.

Using these same 8 parameters as input, neural networks including a hidden layer of 4 nodes were optimized using

Table 8. (a) Mean and Standard Deviation of Tanimoto Coefficients of Each Generation vs Cycloguanil and (b) Mean and Standard Deviation (σ) of Tanimoto Coefficients of Top 10 Scoring Compounds of Each Generation vs Cycloguanil

(a)				(b)			
gen	mean	σ	*n*	gen	mean	σ	*n*
0	84.79	8.82	34	0	77.9	7.28	10
1	75.46	5.82	50	1	73.3	2.63	10
2	73.18	3.98	50	2	71.9	2.85	10
3	72.86	5.00	50	3	70.00	0.00	10
4	75.9	7.73	50	4	71.70	5.85	10
5	75.64	7.96	50	5	70.00	1.94	10
6	74.9	7.49	50	6	70.2	2.39	10
7	74.64	7.20	50	7	71.5	2.46	10
8	76.16	7.36	50	8	71.7	3.13	10
9	74.92	7.11	50	9	71.8	3.22	10
10	76.18	7.68	50	10	72.1	3.25	10

Table 9. Nonlinear Neural Networks Using 4 MOE-Derived Descriptors for the Prediction of Experimental pK_i Using Leave-One-out Cross-Validation over Known DHFR Inhibitors Using 4 Inputs, 2 Hidden Nodes, and 1 Output Node

G	leave-one-out cross-validation R^2	leave-one-out cross-validation adj. R^2
50	0.169	0.140
100	0.314	0.291
150	0.496	0.479
200	0.447	0.427
250	0.453	0.435
300	0.513	0.496
350	**0.638**	**0.625**
400	0.540	0.524
450	0.504	0.487
500	0.543	0.527
750	0.613	0.600
1000	0.268	0.243
1250	0.493	0.476
1500	0.507	0.490
1750	0.621	0.608
2000	0.564	0.549
2500	0.634	0.621
3000	0.619	0.606
4000	0.533	0.517
5000	0.499	0.481

evolutionary computation. For this purpose, various generations of evolution were tested over the range [500, 5000]. The best neural network resulting from this process was identified after 2500 generations of evolution (Figure 9). Figure 9a shows the minimization of MSE (y-axis) over time in terms of the number of generations of evolution (x-axis) as reflected in the held-out testing data (78 samples) as measured every 50 generations. Rapid convergence to a useful model occurred within the first 1000 generations, followed by a slow increase in MSE (decreased performance) thereafter. The mean performance shown in Figure 9a is a mean of 30 trials with random starting weight assignments

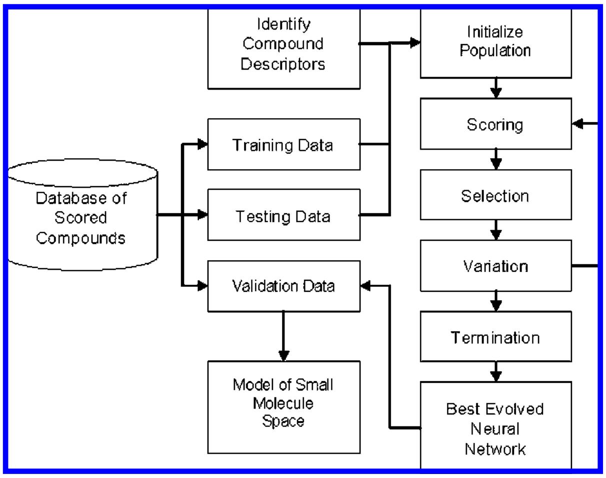

Identify Compound Descriptors
Initialize Population
Scoring
Training Data
Selection
Database of Scored Compounds
Testing Data
Variation
Validation Data
Termination
Model of Small Molecule Space
Best Evolved Neural Network

Figure 7. Evolved neural networks for compound screening. A database of compounds scored via automated fragment assembly is divided into training, testing, and validation data sets. Descriptors for compounds are identified from the literature and available computational resources. A population of neural network models is initialized and iterated evolution (scoring, selection, and variation) is used to generate superior neural network models. After termination of the optimization process, the best evolved neural network model is assayed on the held-out validation data and if determined to be suitable, considered as a QSAR model for the space regarding the target of interest.

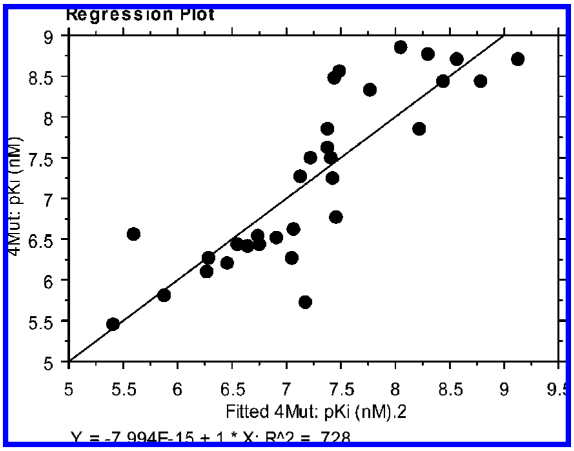


Figure 8. Best multiple linear regression.

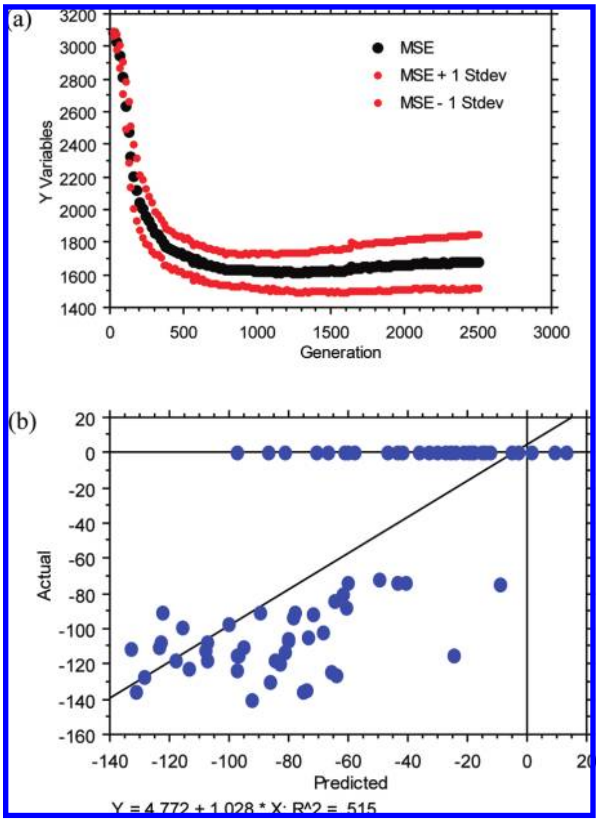


Figure 9. Performance of the best-evolved neural network using the 8 parameters of the linear model: (a) convergence of the evolutionary optimization and (b) regression for the best evolved neural network.

Table 10. Confusion Matrix for the Training Data Used To Develop the Best Neural Network for Template 1

	predicted actives	predicted inactives
actual actives (<−60 protein−ligand score)	85/96 = 89%	11/96 = 11%
actual inactives (>−60 protein−ligand score)	12/56 = 21%	44/56 = 79%

Table 11. Confusion Matrix for the Testing Data Resulting from the Best Neural Network for Template 1

	predicted actives	predicted inactives
actual actives (<−60 protein−ligand score)	37/43 = 86%	6/43 = 14%
actual inactives (>−60 protein- ligand score)	7/35 = 20%	28/35 = 80%

on all connections in the neural network. The performance of the best neural network of these 30 trials is shown in Figure 9b and was an improvement relative to the direct input-output connection neural network.

For the 152 samples used in training of these neural networks, the confusion matrix that resulted is shown in Table 10. The confusion matrix for the performance on the 78 samples of held-out testing data for this same best network is shown in Table 11. A threshold of −60 on the *x*-axis was chosen as a useful separator for active (those <−60) vs inactive (those >−60) compounds.

Using 4 hidden nodes and 1 output node (representing the prediction of protein−ligand interaction) the neural nets were evolved based on minimization of MSE over simulated evolution expressed as a mean and standard deviation over 30 trials. Separation of active and inactive compounds using this neural network shows improvement not only in terms of R^2 correlation but also separation of these two classes.

Tables 10 and 11 indicate the robustness of this approach. Not only the correlation is improved relative to the best multiple linear regression but also the separation of actives from inactives can be clear when using a threshold of 0.6 on the neural network output. In this condition, 86% of true actives are called correctly, and 80% of true inactives are called correctly. Thus, we had arrived at a neural network that could be used for the prescreening of compounds by modeling our previous experimentation with Template 1.

Template 2 - Cycloguanil QSAR via Evolved Neural Networks. This same process was repeated for the experiments that had been conducted with Template 2. A best 11-term multiple linear regression showed very little hope of separating actives from inactives in the data, despite a reasonable correlation (R^2 = 0.503). A best direct input-output connection neural network was evolved after 2500 generations using these same 11 input features and leave-one-out cross-validation. The correlation dropped substantially to R^2 = 0.385. 333 samples were divided randomly into 66% training samples and 33% testing samples resulting in clear separation of the data and better correlation (R^2 = 0.537). To achieve this best result, a variety of experiments were conducted with different generations of evolution [500, 5000] and neural network/evolutionary parameters.

Tables 12 and 13 indicate the performance of this best neural network for Template 2 on training, testing, and the combination of training and testing respectively. Note that the performance using a 0.6 threshold (same threshold as with the best neural network for Template 1) was similar to Template 1 on the testing data, although the number of actual actives predicted correctly was slightly lower for Template 2 (76% for Template 2 vs 86% for Template 1). This could simply be due to a difference in the parameters that were used during model development, or the nature of the compounds/ligand binding differences between Templates 1 and 2, or some combination thereof. Nonetheless, both the resulting best neural networks for Templates 1 and 2 were considered reasonable for testing in light of task 5, in terms of high throughput processing of compounds for truly novel discovery.

Table 12. Confusion Matrix for the Training Data Used To Develop the Best Neural Network for Template 2

	predicted actives	predicted inactives
actual actives (<−60 protein−ligand score)	108/137 = 79%	29/137 = 21%
actual inactives (>−60 protein- ligand score)	11/85 = 13%	74/85 = 87%

Table 13. Confusion Matrix for the Testing Data Used To Evaluate the Best Neural Network for Template 2

	predicted actives	predicted inactives
actual actives (<−60 protein−ligand score)	53/70 = 76%	17/70 = 24%
actual inactives (>−60 protein- ligand score)	8/41 = 20%	33/41 = 80%

Novel DHFR Inhibitor Discovery. The overall discovery process is indicated in Figure 10. Starting with 500 randomly generated Smiles strings using the template and R-groups identified previously, MOE was used to generate a set of descriptors suitable for use with the best evolved neural networks from for both Template 1 and Template 2. The predictions made by these neural networks were normalized to the lowest scoring compound for each net, and the normalized scores are summed, so as not to give any additional bias to one net or the other simply based on differences in predicted range. The summed normalized scores were ranked, and the top 50 compounds with the lowest normalized scores were selected for docking. These compounds were docked via GOLD and scored with Molegro and ranked by their predicted ligand binding affinity score.

The top 10 compounds with lowest predicted binding affinity were selected as parents, and 49 offspring were generated at random from each of these parent solutions generating a total of 490 offspring. The variation consisted of returning back to the possible lists of R-group substituents and selecting at random a position to vary and for that position an R-group change. In the case that the offspring was identical to a parent solution or identical to other offspring solutions, additional changes were made at random to ensure that redundant compounds were avoided. This process was iterated, and the Tanimoto index and predicted ligand binding affinity monitored to determine the rate of convergence on useful solutions. In parallel with this automated method for high-throughput screening, we pro-

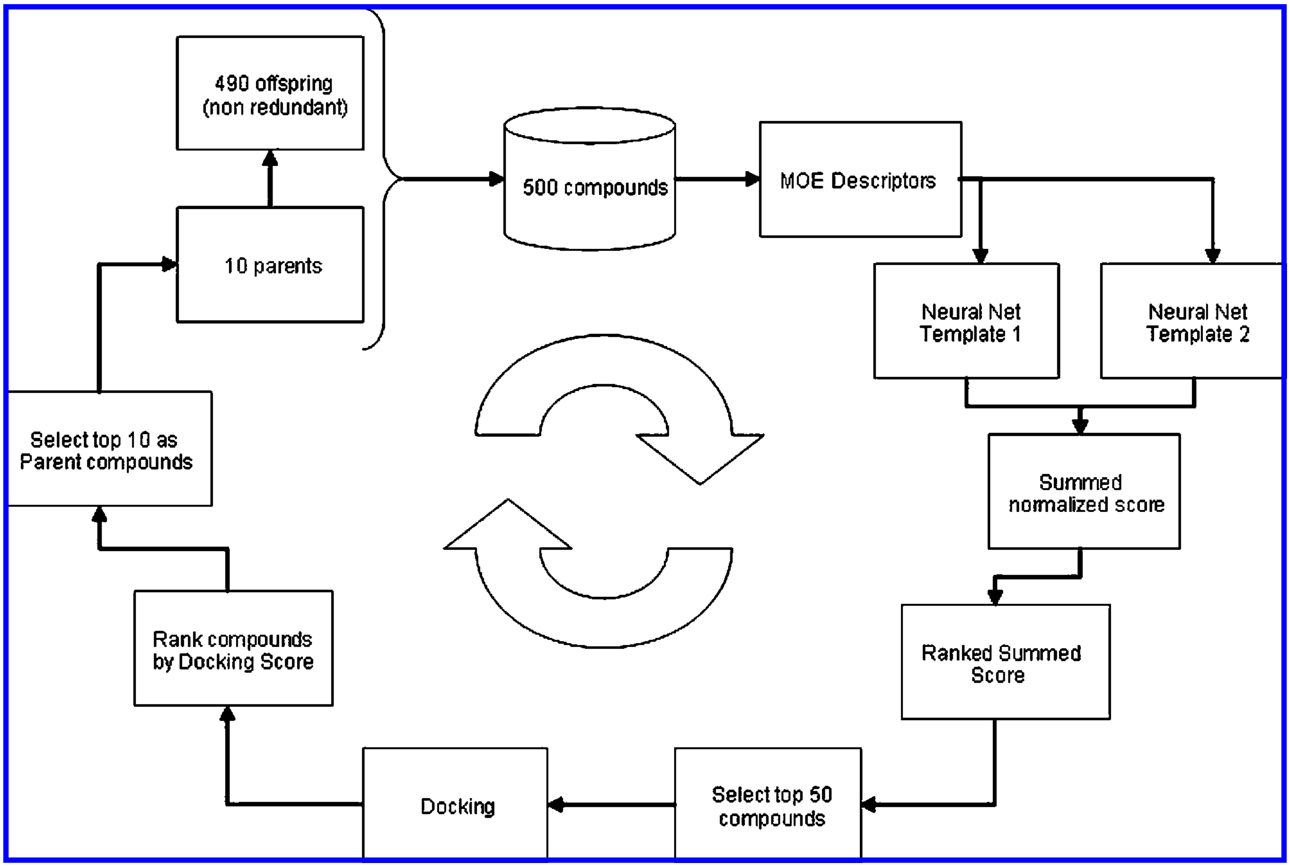


Figure 10. Strategy for evolutionary optimization of novel DHFR inhibitors incorporating the evolved neural nets for increased efficiency and exploration.

Table 14. Number of Compounds in Each Generation Scoring in the Top 50 (Out of 500) Total for the Two Evolved Neural Networks

generation	number of top scoring compounds in pyrimethamine neural net	number of top scoring compounds in cycloguanil neural net	number of top scoring compounds in both neural nets
0	0	19	0
1	25	20	5
2	17	22	3
3	13	29	4
4	19	27	6
5	26	34	13

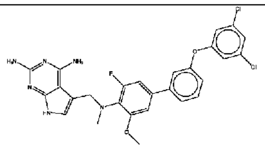
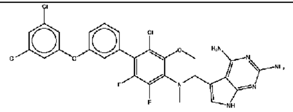
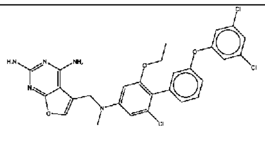
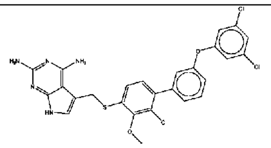
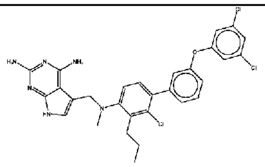
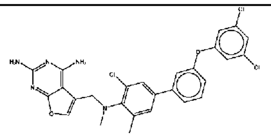
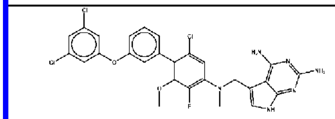
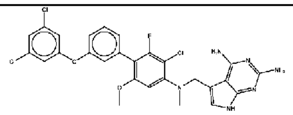
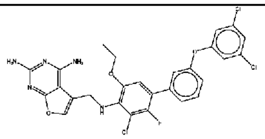
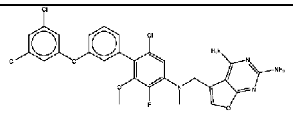
cessed 500 compounds without neural network reduction to better understand the performance improvement (in terms of both times savings and final solution quality) offered by this novel prescreening approach.

Template 3 Evolution. A third structural template consisting of a pyrimidine core was designed (Figure 6 and Table 3). This template was selected for multiple reasons. This template contained the largest number of published activity data vs quadruple mutant *Pf*DHFR.[57,58] Activities ranged from pIC_{50} (in nM) values of 4 to 7 (mean $\text{pIC}_{50} = 5.7$). Compared with the published activities of pyrimethamine (Template 1: mean $\text{p}K_i = 7.2$ nM) and cycloguanil (Template 2: mean $\text{p}K_i = 7.2$ nM) analogues these compounds were as a class significantly less potent (only 1 above 7 pIC_{50}) presenting an excellent opportunity to test the ability of the methodology to evolve more potent compounds. Lastly, selection of Template 3 provided an opportunity to evaluate the robustness, scalability, and search efficiency of the methodology.

The R-groups from Templates 1 and 2 were combined with those taken from pyrimidine compounds to be used for compound evolution. The resulting number of possible compounds in the Template 3 space was 25,206,246,630. From these, an initial set of 500 Smiles strings (termed "Pre-Generation 0") was generated at random (a common strategy for starting such evolutionary searches).

Unfortunately, the vast majority of the 500 generated structures were synthetically inaccessible, extremely large and bulky, and were, by visual inspection, obviously "non-druglike." To ensure synthetic accessibility, and to start the search from a more "druglike" region of the vast chemical space, the 500 compounds for generation 0 were regenerated. 42 compounds from the literature were chosen as parents, and >10 offspring were generated at random from each of these parent solutions generating a total of 500 offspring. For this proof of concept study, this procedure ensured sufficient sampling of structure space and introduced sufficient diversity to allow for evolution to sets of more potent compounds. In future applications we plan to generate structures using templates and position dependent R-groups defined by the capability of a vendor to actually synthesize

Table 15. Results of Template 3 Evolution[a]

Generation #5 Structures	ID	Fitness Score	Generation #5 Structures	ID	Fitness Score
	3_4_212	-188.124		3_5_193	-175.517
	3_5_409	-185.879		3_5_261	-174.331
	3_4_240	-183.2		3_5_330	-173.863
	3_5_203	-177.835		3_5_245	-173.631
	3_5_432	-176.962		3_5_352	-173.126

[a] 10 of the top scoring compounds are from generation 5.

Table 16. (a) Mean and Standard Deviation of Tanimoto Coefficients of Each Generation vs Pyrimethamine and (b) Mean and Standard Deviation (σ) of Tanimoto Coefficients of Top 10 Scoring Compounds Each Generation vs Pyrimethamine

(a)				(b)			
gen	mean	σ	n	gen	mean	σ	n
0	78.52	3.54	50	0	79.00	3.71	10
1	79.02	3.37	50	1	80.90	3.41	10
2	80.00	3.91	50	2	83.60	2.07	10
3	83.00	3.04	50	3	83.30	1.89	10
4	83.82	2.04	50	4	84.10	2.02	10
5	84.04	1.97	50	5	83.60	2.07	10

Table 17. (a) Mean and Standard Deviation of Tanimoto Coefficients of Each Generation vs Cycloguanil and (b) Mean and Standard Deviation (σ) of Tanimoto Coefficients of Top 10 Scoring Compounds of Each Generation vs Cycloguanil

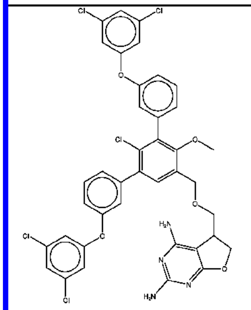
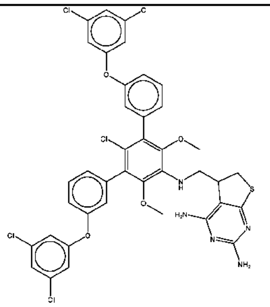
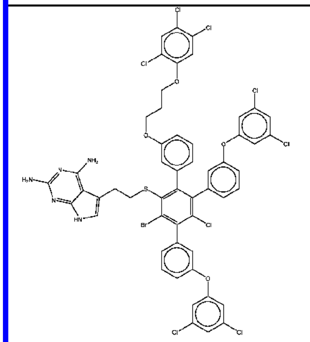
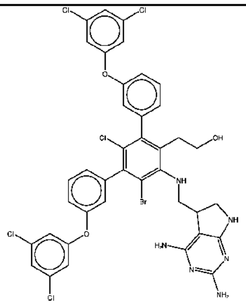
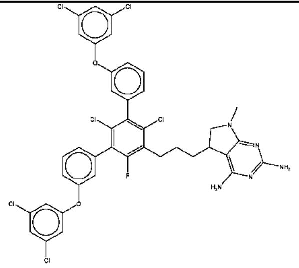
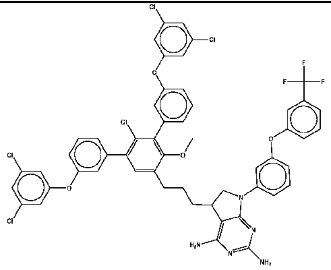
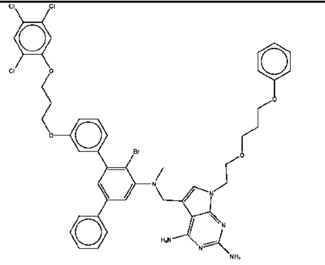
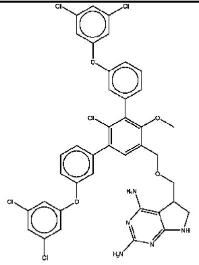
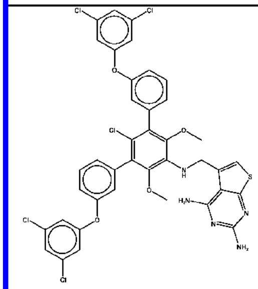
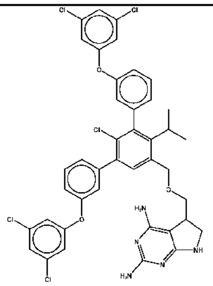
(a)				(b)			
gen	mean	σ	n	gen	mean	σ	n
0	67.84	5.55	34	0	66.50	6.98	10
1	65.90	5.05	50	1	69.00	4.59	10
2	68.42	4.33	50	2	72.80	2.39	10
3	71.56	3.48	50	3	73.60	2.27	10
4	72.44	3.00	50	4	72.70	2.45	10
5	72.72	2.47	50	5	73.10	2.56	10

them. Additionally, filters will be applied to remove "non-druglike" structures.

MOE descriptors for these 500 compounds were then generated and fed into the best evolved neural network models from Templates 1 and 2. The top 50 compounds with lowest normalized scores were selected for docking. This represented generation 0 for the strategy described above. Following this cycle, five generations of evolutionary discovery were completed.

As the number of evolutionary generations increased, a higher representation of top scorers from each neural net model resulted. In addition, the number of compounds scoring in the top 50 for both neural net models simultaneously increased from 0 in generation 0 to 13 in generation 5 (Table 14). This result is very important in that it suggests that multiple neural net models can be used to assist with the simulated evolution of compounds and help to focus the search for desired solutions in a very large space of possibilities. In this study, the two neural net models were based on predicted activity versus the quadruple mutant *Pf*DHFR-TS. This method can easily be expanded to include neural net models for predicted activity versus different mutant strains in order to evolve potential potent pan mutant

Table 18. Results of the Parallel Template 3 Evolution with No Preselection with the Neural Net Models[a]

Generation #5 Structures	ID	Fitness Score	Generation #5 Structures	ID	Fitness Score
	3_5_14	-231.12		3_4_13	-214.728
	3_5_20	-225.116		3_5_25	-211.333
	3_5_23	-224.755		3_4_27	-209.582
	3_5_2	-218.912		3_4_36	-209.285
	3_5_38	-217.608		3_5_15	-208.958

[a] 10 of the top scoring compounds from generation 5.

Table 19. Fitness Scores and Qikprop Number of Stars (# Stars) for the Top Ten Scoring Compounds from Generation 5 With and Without Neural Net Prescreening[a]

generation 5 w/neural net prescreening			generation 5 w/out neural net prescreening		
ID	Fitness Score	#Stars	ID	Fitness Score	#Stars
3_4_212	−188.12	4	3_5_14	−231.12	10
3_5_409	−185.88	5	3_5_20	−225.12	15
3_4_240	−183.20	6	3_5_23	−224.76	11
3_5_203	−177.84	5	3_5_2	−218.91	14
3_5_432	−176.96	4	3_5_38	−217.61	11
3_5_193	−175.52	5	3_4_13	−214.73	11
3_5_261	−174.33	6	3_5_25	−211.33	13
3_5_330	−173.86	5	3_4_27	−209.58	12
3_5_245	−173.63	6	3_4_36	−209.29	11
3_5_352	−173.13	4	3_5_15	−208.96	11
average	**−178.25**	**5**	**average**	**−217.14**	**11.9**

[a] The # Stars represents the number of property or descriptor values (calculated in Qikprop) that fall outside the 95% range of similar values for known drugs. A large number of stars suggests that a molecule is less druglike than molecules with few stars.

Table 20. Fitness Scores and Qikprop # Stars for the Top Ten Scoring Compounds from Template 1 and 2 Generation 10[a]

Template 1 Generation 10			Template 2 Generation 10		
ID	Fitness Score	#Stars	ID	Fitness Score	#Stars
10_15	−147.55	1	2_4_7	−126.15	4
10_17	−143.32	0	2_8_36	−119.19	1
10_20	−142.75	1	2_9_1	−117.17	3
10_13	−141.51	0	2_9_15	−113.50	2
10_24	−140.22	1	2_8_37	−113.03	1
9_25	−138.24	0	2_8_8	−111.96	5
8_9	−137.44	1	2_3_9	−111.25	4
10_4	−137.25	0	2_6_13	−110.89	1
10_6	−133.16	0	2_10_10	−102.79	4
10_15	−147.55	1	2_10_13	−100.43	3
average	**−139.45**	**0.45**	**average**	**−112.64**	**2.8**

[a] The #Stars represents the number of property or descriptor values (calculated in Qikprop) that fall outside the 95% range of similar values for known drugs. A large number of stars indicates that a molecule is less druglike than molecules with few stars.

DHFR inhibitors. The results of Template 3 evolution are presented in Table 15.

A diversity analysis was performed on the compounds from the 5 generations of evolution for Template 3 to evaluate the diversity of structures generated during the evolution. As a suitable representative compound in the Template 3 structure class was not readily available, both pyrimethamine and cycloguanil were used to generate Tanimoto coefficients for all 250 compounds evaluated (excluding duplicates from the top 10 parents in each generation). The mean and standard deviation for each generation was then calculated to provide a measure of how far the compounds evolved from the starting library. Tables 16 and 17 illustrate the increase in mean Tanimoto index (indicating increasing similarity to the reference compound) and decrease of the Tanimoto index standard deviation (indicating the compounds were converging on a similar structure class). Again this convergence was more pronounced when examining the standard deviations of the Tanimoto coefficients for the top ten compounds in each generation.

In order to evaluate the efficiency and effectiveness of the neural nets, five additional generations of evolution were performed in parallel. These were performed without using the Template 1 and 2 neural networks for compound preselection as a basis for true comparison. 50 compounds were generated from the 10 top-scoring parents of the previous generation (as described for Templates 1 and 2) (Table 18).

Most of the compounds resulting from this parallel evolution appeared by visual inspection to be large and bulky and to fall in “nondruglike” chemical space (Table 18). Qikprop was used to evaluate this more quantitatively. Qikprop is a software package widely used to evaluate compounds for their “druglikeness.” 34 Qikprop descriptors were calculated for the top ten scoring compounds from the Generation 5 of the two parallel evolutions. In general, 5 “Stars” are considered to be the upper-limit of “druglikeness” (fewer Qikprop stars are better).

Comparing generation 5 compounds from the nonfiltered evolution to those generation 5 compounds from the filtered evolution, the average fitness for the nonprefiltered compounds was −217.14 versus −178.25, while the number of Stars was 11.9 versus 5 respectively (Table 19). These results are consistent with the larger size and complexity of the nonfiltered compounds which will have increased van der Waals and hydrophobic contributions to the binding affinity at the price of additional loss of druglikeness. A similar analysis was performed on Templates 1 and 2 (see Table 20). As expected, these compounds were simpler in structure than those of Template 3 and had lower fitness scores as well as significantly lower number of Stars. In future studies we will include this druglikeness parameter as part of the fitness evaluation for small molecule discovery to ensure druglikeness of evolved small molecules.

CONCLUSIONS

In this paper we have presented several successful accomplishments for a novel high-throughput small molecule screening technology. These include 1) the ability to use nonlinear modeling approaches to outperform standard linear regression modeling of QSAR in light of published data, 2) automated feature reduction simultaneous to model optimization, 3) methods of evolved fragment assembly in light of a chemically synthesizable space, and 4) the use of best-evolved neural networks models as filters for directed *in silico* screening of compounds using evolved fragment assembly. Lists of putative, novel, DHFR inhibitors were generated from a decomposed fragment library of known DHFR inhibitors. To ensure synthetic accessibility in this proof of concept study, the decomposition followed the synthetic strategy and mirrored the original SAR analysis of the published literature.[53−58]

While *in vivo* experiments have yet to be performed to verify binding affinities, these compounds serve as potentially interesting compounds for future studies that will focus strongly on functional assays and structural confirmation via NMR so that the predicted activity and binding poses for these compounds can be confirmed. These experimental data can be fed back into the computational engine for further ligand evolution. For future studies we plan to select additional scaffolds and R-groups that will result in novel, synthetically accessible structures. Although this approach was tested on dihydrofolate reductase (DHFR) for novel antimalarial drug discovery, the methods presented in this paper can readily be applied to other targets of interest.

ACKNOWLEDGMENT

This work is supported by the U.S. Army Medical Research and Material Command under Contract No. W81XWH-06-C-0399. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision unless so designated by other documentation. The authors would also like to thank Mars Cheung for his technical assistance.

REFERENCES AND NOTES

(1) Schnecke, V.; Boström, J. Computational chemistry-driven decision making in lead generation. *Drug Discovery Today* **2006**, *11*, 43–50.

(2) Good, A. C.; Krystek, S. R.; Mason, J. S. High-throughput and virtual screening: core lead discovery technologies move towards integration. *Drug Discovery Today* **2000**, *5*, S61–S69.

(3) Anderson, A. C.; Wright, D. L. The design and docking of virtual compound libraries to structures of drug targets. *Curr. Comput.-Aided Drug Des.* **2005**, *1*, 103–127.

(4) Mauser, H.; Guba, W. Recent developments in de novo design and scaffold hopping. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 365–374.

(5) Jhoti, H. Fragment-based drug discovery using rational design. *Ernst Schering Found. Symp. Proc.* **2007**, *3*, 169–85.

(6) Honma, T. Recent advances in de novo design strategy for practical lead identification. *Med. Res. Rev.* **2003**, *23*, 606–632.

(7) Pearlman, D. A.; Murcko, M. A. CONCERTS: dynamic connection of fragments as an approach to de novo ligand design. *J. Med. Chem.* **1996**, *39*, 1651–1663.

(8) Böhm, H. J. The computer program LUDI: A new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 61–78.

(9) Böhm, H. J. On the use of Ludi to search the Fine Chemical Directory for ligands of proteins of known 3-dimensional structure. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 623–632.

(10) Lauri, G.; Bartlett, P. A. CAVEAT: a program to facilitate the design of organic molecules. *J. Comput.-Aided Mol.Des.* **1994**, *1*, 51–66.

(11) Tschinke, V.; Cohen, N. C. The NEWLEAD program: a new method for the design of candidate structures from pharmacophoric hypotheses. *J. Med. Chem.* **1993**, *36*, 3863–70.

(12) Miranker, A.; Karplus, M. An automated method for dynamic ligand design. *Proteins* **1995**, *23*, 472–490.

(13) Roe, D. C.; Kuntz, I. D. BUILDER v.2: Improving the chemistry of a de novo design strategy. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 269–282.

(14) Stahl, M.; Todorov, N. P.; James, T.; Mauser, H.; Boehm, H. J.; Dean, P. M. A validation study on the practical use of automated de novo design. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 459–78.

(15) Firth-Clark, S.; Kirton, S. B.; Willems, H. M.; Williams, A. De novo ligand design to partially flexible active sites: application of the ReFlex algorithm to carboxypeptidase A, acetylcholinesterase, and the estrogen receptor. *J. Chem. Inf. Model.* **2008**, *48*, 296–305.

(16) Westhead, D. R.; Clark, D. E.; Frenkel, D.; Li, J.; Murray, C. W.; Robson, B.; Waszkowycz, B. PRO_LIGAND: An approach to de novo molecular design. 3. A genetic algorithm for structure refinement. *J. Comput.-Aided Mol. Des.* **1995**, *9*, 139–148.

(17) Pegg, S. C.; Haresco, J. J.; Kuntz, I. D. A genetic algorithm for structure-based de novo design. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 911–933.

(18) *Leapfrog, 6.8 ed.*; Tripos, Inc.: St. Louis, MO.

(19) Nishibata, Y.; Itai, A. Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* **1991**, *47*, 8985–8990.

(20) Caflisch, A; Miranker, A; Karplus, M. Multiple copy simultaneous search and construction of ligands in binding sites: application to inhibitors of HIV-1 aspartic proteinase. *J. Med. Chem.* **1993**, *36*, 2142–2167.

(21) Bohacek, R. S.; McMartin, C. Multiple highly diverse structures complementary to enzyme binding sites:Results of extensive application of de novo design method incorporating combinatorial growth. *J. Am. Chem. Soc.* **1994**, *116*, 5560–5571.

(22) Rotstein, S. H.; Murcko, M. A. Genstar-a method for de novo drug design. *J. Comput.-Aided Mol. Des.* **1993**, *7*, 23–43.

(23) DeWitt, R.; Shaknovich, E. SmoG: de novo design method based on simple, fast, and accurate free energy estimates. 1. Methodology and supporting evidence. *J. Am. Chem. Soc.* **1996**, *118*, 11733–11744.

(24) DeWitt, R.; Shaknovich, E. SmoG: de novo design method based on simple, fast, and accurate free energy estimates. 2. Case studies on molecular design. *J. Am. Chem. Soc.* **1996**, *119*, 4608–4617.

(25) Rotstein, S. H.; Murcko, M. A. Groupbuild-a fragment-based method for de novo drug design. *J. Med. Chem.* **1993**, *36*, 1700–1710.

(26) Gillet, V. J.; Newell, W; Mata, P.; Myatt, G.; Sike, S.; Zsoldos, Z.; Johnson, A. P. SPROUT: recent developments in the de novo design of molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 207–17.

(27) Moon, J.; Howe, W. Computer design of bioactive molecules: a method for receptor-based de novo ligand design. *Proteins* **1991**, *11*, 314–328.

(28) Kitchen, D. B.; Decornez, H.; Furr, J. R.; Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discovery* **2004**, *3*, 935–949.

(29) Leach, A. R.; Shoichet, B. K.; Peishoff, C. E. Prediction of protein ligand interactions. Docking and scoring: successes and gaps. *J. Med. Chem.* **2006**, *49*, 5851–5855.

(30) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-Based Lead Discovery. *Nat. Rev. Drug Discovery* **2005**, *3*, 660–672.

(31) Congreve, M.; Carr, R.; Murray, C.; Jhoti, H. A ‘rule of three’ for fragment-based lead discovery. *Drug Discovery Today* **2003**, *8*, 876–877.

(32) Böhm, H. J.; Stahl, M. Combinatorial docking and combinatorial chemistry: design of potent non-peptide thrombin inhibitors. *J. Comput.-Aided Mol. Des.* **1999**, *13*, 51.

(33) Kick, E. K.; Roe, D. C.; Skillman, A. G.; Liu, G.; Ewing, T. J.; Sun, Y.; Kuntz, I. D.; Ellman, J. A. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chem. Biol.* **1997**, *4*, 297.

(34) Todorov, N. P.; Dean, P. M. Evaluation of a method for controlling molecular scaffold diversity in de novo ligand design. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 175–192.

(35) Murray, C. W.; Clark, D. E.; Auton, T. R.; Firth, M. A.; Li, J.; Sykes, R. A.; Waszkowycz, B.; Westhead, D. R.; Young, S. C. PRO_SELECT: Combining structure-based drug design and combinatorial chemistry for rapid lead discovery. 1. Technology. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 193–207.

(36) Fogel, G. B. Computational intelligence approaches for pattern discovery in biological systems. *Briefings Bioinf.* **2008**, *9*, 307–316.

(37) Budin, N.; Ahmed, S.; Majeux, N.; Caflisch, A. An Evolutionary Approach for Structure-based Design of Natural and Non-natural Peptidic Ligands. *Comb. Chem. High Throughput Screening* **2001**, *4*, 661–673.

(38) Belda, I.; Madurga, S.; Llorà, X.; Martinell, M.; Tarragó, T.; Piqueras, M. G.; Nicolás, E.; Giralt, E. ENDPA: an evolutionary structure-based *de novo* peptide design algorithm. *J. Comput.-Aided Mol. Des.* **2005**, *19*, 585–601.

(39) Belda, I.; Madurga, S.; Tarragó, T.; Llorà, X.; Giralt, E. Evolutionary computation and multimodal search: A good combination to tackle molecular diversity in the field of peptide design. *Mol. Diversity* **2007**, *11*, 7–21.

(40) Hou, T.; Xu, X. A new molecular simulation software package - Peking University Drug Design System (PKUDDS) for structure-based drug design. *J. Mol. Graphics Modell.* **2001**, *19*, 455–465.

(41) Douguet, D.; Munier-Lehmann, H.; Labesse, G.; Pochet, S. LEA3D: A Computer-Aided Ligand Design for Structure-Based Drug Design. *J. Med. Chem.* **2005**, *48*, 2457–2468.

(42) Bandyopadhyay, S.; Bagchi, A.; Maulik, U. Active Site Driven Ligand Design: An evolutionary Approach. *J. Bioinf. Comput. Biol.* **2005**, *3*, 1053–1070.

(43) Liu, Q.; Masek, B.; Smith, K.; Smith, J. Tagged Fragment Method for Evolutionary Structure-Based De Novo Lead Generation and Optimiziation. *J. Med. Chem.* **2007**, *50*, 5392–5402.

(44) Dey, F.; Calfisch, A. Fragment-Based de Novo Ligand Design by Multiobjective Evolutionary Optimization. *J. Chem. Inf. Model.* **2008**, *48*, 679–690.

(45) Schüller, A.; Suhartono, M.; Fechner, U.; Tanrikulu, Y.; Breitung, S.; Scheffer, U.; Göbel, M. W.; Schneider, G. The concept of template-based de novo design from drug-derived molecular fragments and its application to TAR RNA. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 59–68.

(46) Hecht, D.; Fogel, G. B. High-throughput ligand screening *via* preclustering and rvolved neural networks. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2007**, *4*, 476.

(47) Hecht, D.; Cheung, M.; Fogel, G. B. QSAR using evolved neural networks for the inhibition of mutant PfDHFR by pyrimethamine derivatives. *Biosystems* **2008**, *92*, 10–15.

(48) Cheung, M.; Johnson, S.; Hecht, D.; Fogel, G. B. Quantitative structure-property relationships for drug solubility prediction using evolved neural networks. *IEEE Congr. Evol. Comput., Hong Kong* **2008**, xx.

(49) Duffy, E. M.; Jorgensen, W. L. Prediction of Properties from Simulations: Free Energies of Solvation in Hexadecane, Octanol, and Water. *J. Am. Chem. Soc.* **2000**, *122*, 2878–2888.

(50) Yazdanian, M.; Glynn, S. L.; Wright, J. L.; Hawi, A. Correlating partitioning and Caco-2 cell permeability of structurally diverse small molecular weight compounds. *Pharm. Res.* **1998**, *15*, 1490–1494.

(51) Irvine, J. D.; Takahashi, L.; Lockhart, K.; Cheong, J.; Tolan, J. W.; Selick, H. E.; Grove, J. R. MDCK (Madin-Darby canine kidney) cells: a tool for membrane permeability screening. *J. Pharm. Sci.* **1999**, *88*, 28–33.

(52) Stenberg, P.; Norinder, U.; Luthman, K.; Artursson, P. Experimental and computational screening models for the prediction of intestinal drug absorption. *J. Med. Chem.* **2001**, *44*, 1927–1937.

(53) Yuvaniyama, J.; Chitnumsub, P.; Kamchonwongpaison, S.; Vanichtanankul, J.; Sirawaraporn, W.; Taylor, P.; Walkinshaw, M.; Yuthavong, Y. Insights into antifolate resistance from malarial DHFR-TS structures. *Nat. Struct. Biol.* **2003**, *10*, 357–365.

(54) Parenti, M. D.; Pacchioni, S.; Ferrari, A. M.; Rastelli, G. Three-Dimensional Quantitative Structure-Activity Relationship Analysis of a set of *Plasmodium falciparum* Dihydrofolate Reductase Inhibitors Using a Pharmacophore Generation Approach. *J. Med. Chem.* **2004**, *47*, 4258–4267.

(55) Kamchonwongpaison, S.; Quarrell, R.; Charoensetakul, N.; Ponsinet, R.; Vilaivan, T.; Vanichtanankul, J.; Tarnchampoo, W.; Sirawaraporn, W.; Lowe, G.; Yuthavong, Y. Inhibitors of multiple mutants of Plasmodium falciparum dihydrofolate reductase and their antimalarial activities. *J. Med. Chem.* **2004**, *47*, 673–680.

(56) Kamchonwongpaison, S.; Vanichtanankul, J.; Tarnchampoo, B.; Yuvaniyama, J.; Taweechai, S.; Yuthavong, Y. Stoichiometric selection of tightbinding inhibitors by wild-type and mutant forms of malarial (Plasmodium falciparum) dihydrofolate reductase. *Anal. Chem.* **2005**, *77*, 1222–1227.

(57) Santos-Filho, O. A.; Hopfinger, A. J. A Search for Sources of Drug Resistance by the 4D-QSAR Analysis of a Set of Antimalarial Dihydrofolate Reductase Inhibitors. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 1–12.

(58) Sutherland, J. J.; Weaver, D. F. Three-Dimensional Quantitative Structure-Activity and Structure-Selectivity Relationships of Dihydrofolate Reductase Inhibitors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 309–331.

(59) Halgren, T. A. The merck force field. *J. Comput. Chem.* **1996**, *17*, 490–512.

(60) Fogel, G. B.; Cheung, M.; Pittman, E.; Hecht, D. In Silico Screening Against Wild-Type and Mutant Plasmodium falciparum Dihydrofolate Reductase. *J. Mol. Graphics Modell.* **2008**, *26*, 1145–1152.

(61) Fogel, G. B.; Cheung, M.; Pittman, E.; Hecht, D. Modeling the Inhibition of Quadruple Mutant Plasmodium falciparum Dihydrofolate Reductase by Pyrimethamine Derivatives. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 29–38.

CI9000647