Summer 8-18-2010

Rational Drug Design for Neglected Diseases: Implementation of Computational Methods to Construct Predictive Devices and Examine Mechanisms

Catharine Jane Collar
Georgia State University

RATIONAL DRUG DESIGN FOR NEGLECTED DISEASES: IMPLEMENTATION OF COMPUTATIONAL METHODS TO CONSTRUCT PREDICTIVE DEVICES AND EXAMINE MECHANISMS

by

CATHARINE JANE COLLAR

Under the Direction of Dr. W. David Wilson

ABSTRACT

Over a billion individuals worldwide suffer from neglected diseases. This equates to approximately one-sixth of the human population. These infections are often endemic in remote tropical regions of impoverished populations where vectors can flourish and infected individuals cannot be effectively treated due to a lack of hospitals, medical equipment, drugs, and trained personnel. The few drugs that have been approved for the treatments of such illnesses are not widely used because they are riddled with inadequate implications of cost, safety, drug availability, administration, and resistance. Hence, there exists an eminent need for the design and development of improved new therapeutics. Influential world-renowned scientists in the Consortium for Parasitic Drug Development (CPDD) have preformed extensive biological testing for compounds active against parasites that cause neglected diseases. These data were acquired through several collaborations and found applicable to computational studies that examine quantitative structure-activity relationships through the development of predictive

models and explore structural relationships through docking. Both of these *in silico* tools can contribute to an understanding of compound structural importance for specific targets. The compilation of manuscripts presented in this dissertation focus on three neglected diseases: trypanosomiasis, Chagas disease, and leishmaniasis. These diseases are caused by kinetoplastid parasites *Trypanosoma brucei*, *Trypanosoma cruzi*, and *Leishmania spp.*, respectively. Statistically significant predictive devices were developed for the inhibition of the: (1) *T. brucei* P2 nucleoside transporter, (2) *T. cruzi* parasite at two temperatures, and (3) two species of *Leishmania*. From these studies compound structural importance was assessed for the targeting of each parasitic system. Since these three parasites are all from the Order Kinetoplastida and the kinetoplast DNA has been determined a viable target, compound interactions with DNA were explored to gain insight into binding modes of known and novel compounds.

INDEX WORDS: Trypanosomiasis, Leishmaniasis, Chagas Disease, Nucleoside transporter, DNA, Molecular Modeling, 3D-QSAR, CoMFA, CoMSIA, Docking

RATIONAL DRUG DESIGN FOR NEGLECTED DISEASES: IMPLEMENTATION OF COMPUTATIONAL METHODS TO CONSTRUCT PREDICTIVE DEVICES AND EXAMINE MECHANISMS

by

CATHARINE JANE COLLAR

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

in the College of Arts and Sciences

Georgia State University

2010

RATIONAL DRUG DESIGN FOR NEGLECTED DISEASES: IMPLEMENTATION OF COMPUTATIONAL METHODS TO CONSTRUCT PREDICTIVE DEVICES AND EXAMINE MECHANISMS

by

CATHARINE JANE COLLAR

Committee Chair: W. David Wilson

Committee: David W. Boykin
Donald Hamelberg

Electronic Version Approval:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

August 2010

DEDICATION

This dissertation is dedicated to individuals suffering from parasitic neglected diseases, as well as the researchers and doctors engaged in related drug discovery and treatment efforts.

ACKNOWLEDGEMENTS

Although a multitude of people have supported me throughout my doctoral research, I could not have accomplished the work presented in this dissertation without the encouragement and exceptional guidance provided by my advisor and mentor Dr. W. David Wilson for which I am eternally grateful. Dr. Wilson is an outstanding professor and a personable professional from whom I have learned much about academia and research. Invaluable conversations with Dr. Wilson led to extensive and creative computational implementation to improve predictive and mechanistic modeling efforts.

My educational experience was enhanced by interdisciplinary collaborations formed through the Consortium for Parasitic Drug Development (CPDD) and funded by the Bill and Melinda Gates foundation. Compounds were synthesized in the exceptional organic and medicinal chemistry laboratories of Dr. David W. Boykin at Georgia State University and Dr. Richard R. Tidwell at the University of North Carolina at Chapel Hill, while extensive biological assays were performed by researchers in the laboratories of Dr. Harry de Koning and Dr. Michael P. Barrett at the University of Glasgow, Dr. Karl Werbovetz at the Ohio State University, Dr. Maria de Nazare Correia Soeiro at the Oswaldo Cruz Institute, and Dr. Reto Brun at the Swiss Tropical Institute. The wealth of compounds and biological data gave way to in-depth quantitative structure-activity relationship (QSAR) studies. My CPDD QSAR research efforts would not have been possible or successful without the data and priceless explanations of the biological assays.

Further mechanistic computational studies were made possible through collaborations with Dr. Moses N. F. Lee at Hope College and Dr. Donald Hamelberg at Georgia State University; research efforts were funded by the National Science Foundation and National Institutes of Health. These studies produced structural data and insight into compound binding interactions in relation to experimentally obtained binding constants. Binding constant data were gathered by Dr. Binh Nguyen, Dr. Manoj Munde, and Dr. Yang Liu of the W. David Wilson laboratory at Georgia State University. Members of the W. David Wilson laboratory, including Carol Wilson, provided perspectives and editing that led to high quality presentations and manuscripts.

Last but not least, throughout my doctorial training, love and support were consistently provided by my family and friends. My Mom and Dad, Jane and Steve Collar, have been inspirational role models; the value that they place on education is just as high as the morals and values that they have instilled in me, my brothers (Curt and Kelly), and my sister (Kristin). My parents and siblings have shown interest in my research endeavors and with this engaged in insightful discussions. Family and friends worldwide encouraged me to pursue my goals and approach life from an optimistic perspective, while my beautiful golden retriever Romeo provided companionship and stress relief. By staying positive, my livelihood while conducting my doctoral research was enjoyable as well as educational.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: NEGLECTED DISEASES

CHAPTER 2: PREDICTIVE COMPUTATIONAL MODELS OF SUBSTRATE BINDING BY A NUCLEOSIDE TRANSPORTER

APPENDICES

LIST OF ABBREVIATIONS

CPDD	consortium for parasitic drug development
NIH	national institutes of health
NSF	national science foundation
CoMFA	comparative molecular field analysis
CoMSIA	comparative molecular similarity indices analysis
QSAR	quantitative structure-activity relationship
PLS	partial least squares
SEE	standard error of estimate
CD	Chagas disease
DA	diamidine
AIA	arylimidamide
Nfx	nifurtimox
Bz	benznidazole
DNA	deoxyribonucleic acid
A	adenine
T	thymine
G	guanine
C	cytosine
Py	*N*-methylpyrrole
Im	*N*-methylimidazole

f	formamido
NMR	nuclear magnetic resonance
ASA	accessible solvent area

CHAPTER 1: NEGLECTED DISEASES

NEGLECTED DISEASES

Worldwide, more than a billion individuals suffer from neglected diseases; yet very few drugs have been approved as therapeutics for these illnesses.[1-3] The lack of therapeutic agents and the adverse effects of those available necessitates drug discovery efforts. Studies addressed in this compilation of manuscripts are for neglected diseases caused by parasites of the Order Kinetoplastida: (1) trypanosomiasis, caused by *Trypanosoma brucei*, (2) Chagas disease, caused by *Trypanosoma cruzi*, and (3) leishmaniasis, caused by species of *Leishmania*.

Therapeutics

Trypanosomiasis. The type of treatment for trypanosomiasis depends on the stage of infection, first or second, and subspecies of parasite.[4-7] Suramin is used to treat *T. brucei rhodesiense* infections, while pentamidine is employed for *T. brucei gambiense*. Side effects of suramin treatment include nausea, vomiting, urticarial rash and lack of consciousness, whereas pentamidine's side effects include hypotension, abdominal pain, hypersalivation, vertigo, nausea, and chest pain. The second stage treatments for both subspecies calls for melarsoprol, a drug that is highly toxic and consists of the following side effects: convulsions, fever, loss of consciousness, rashes, bloody stool, nausea, and vomiting, as well as myocardial damage, albuminuria, and hypertension. Eflornithine can also be employed to specifically treat *T. brucei gambiense*. The side effects associated with this compound include diarrhea, suppression of bone marrow, anaemia, and leukopenia.

Chagas Disease. Accepted clinical treatments for Chagas disease are Nifurtimox (Nfx) and Benznidazole (Bz); these compounds are not FDA approved.[7-9] The most common side effects of Nfx are abdominal pain, dizziness, headache, loss of appetite, nausea, vomiting, and weight loss, whereas the most common side effects of Bz include gastrointestinal symptoms such as nausea and peripheral neuropathy.

Leishmaniasis. Infections of *Leishmania spp.* result in three forms of the leishmaniasis disease: cutaneous, mucosal, and visceral.[3, 4, 7, 10, 11] The preferred treatments are sodium stibogluconate for cutaneous and mucosal leishmaniasis and liposomal amphotericin B for visceral leishmaniasis. However, due primarily to the high cost of liposomal amphotericin B, sodium stibogluconate is commonly used to treat all three leishmaniasis disease forms. The most common side effect of sodium stibogluconate includes thrombophlebitis, abdominal pain, nausea, vomiting, anorexia, myalgia, arthralgia, and headache. The most predominant side effect of amphotericin B is nephrotoxicity.

Research Approach

Biological testing data were acquired through collaborations with world-renowned scientists in the Consortium for Parasitic Drug Development (CPDD). The compounds and their respective activities were employed for computational studies that examine: (1) quantitative structure-activity relationships (QSAR) through the development of predictive models and (2) explore structural relationships through docking.

Predictive Models. The QSAR of the structural and biological data acquired was assessed through partial least squares (PLS) regression modeling employing the biologically obtained activities and computationally calculated comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) molecular descriptors. In general, QSAR-PLS studies follow these steps of progression: (1) compound input, (2) compound minimization, (3) compound alignment, (4) molecular descriptor calculation, and (5) regression model formation.

Before QSAR-PLS predictive models can be formed, an extensive dataset of compound structures with biological activities must be acquired; it is important that biological activities are gained by the same biological assay for each compound of the dataset. Compounds employed for QSAR-PLS predictive modeling can consist of several diverse backbones. More diversity in a molecular modeling system leads to a greater range of structures applicable for prediction.

When employing the SYBYL[12] software environment to a dataset of compounds with biological activities, the Sketch Molecule menu can be opened and compounds may be drawn. Upon completion of a compound the Sketch Molecule menu needs to be exited and the compound ought to be named *via* the Name Molecule menu. The molecules should then undergo an initial minimization which can be done using the Minimize Molecule option. Examples of constructed and minimized structures may be viewed in Figure 1. Each named structure can then be placed into a constructed database through the Database Put Molecule option.

Subsequently to the input of all compounds into the database, possible conformations of structures should be assessed. This can be done through several methods including but not limited to: Systematic Conformational Search, Grid Search, Random Conformational Search, MultiSearch, and GA Conformational Search. The lowest energy conformations of compounds obtained ought to be further studied. To insure that compounds are in their lowest energy conformations these compounds may be re-minimized and moved to new databases. Figure 2 displays three low energy structures of an arylimidamide compound.

Compounds of similar low energy structural conformations ought to then be aligned; each alignment should consist of only one structural representation for each compound of the dataset. Alignment can be acquired in several ways including but not limited to: Fit Atoms, Match Atoms, Superimpose Atoms, Multifit, GALAHAD, and GASP. Examples are displayed in Figure 3. Optimal compound alignment is essential to the construction of employable QSAR-PLS models.[13]

A molecular spreadsheet ought to be constructed following alignment; this can be done by opening the database through the Open menu. Biological activities can then be input into the spreadsheet and molecular descriptors may be calculated by using the AutoFill menu of the spreadsheet. CoMFA and CoMSIA molecular descriptors can be calculated for QSAR-PLS modeling.[14] CoMFA has become a model system for QSAR modeling methods and CoMSIA was developed to overcome limitations of CoMFA.[14, 15] For CoMFA, each compound of a dataset is assigned interaction energies with respect to a probe atom and steric and electrostatic

molecular descriptors are calculated with a particular potential function; Lennard-Jones and Coulomb potentials, respectively.[16] To keep the calculation energies in reasonable boundaries cut-off values are fixed: 5 kcal/mol for the Lennard-Jones potential and ±30 kcal/mol for the Coulomb potential. For CoMSIA, similarity indices are compiled for the compounds of a dataset at the intersections of a regularly spaced lattice.[13, 16, 17] This is conducted with a grid and probe method, similar to CoMFA. In CoMSIA, a common probe is employed in a distance dependent approach that scans the entirety of the lattice and embeds each compound; the lattice points inside and outside the molecule are employed and cut-offs are not needed. Steric, electrostatic, hydrophobic, donor, and acceptor molecular descriptors are calculated using positive and negative fields acquired through similarity indices. The CoMSIA method indirectly evaluates the similarities of each molecule in the dataset, whereas the CoMFA method evaluates the compounds of the dataset through relative interaction energies dependant on molecular positions.

PLS can then be employed to compare the biological activities of compounds to their respective calculated molecular descriptors; the PLS regression technique solves the linear model in a stepwise approach that includes every predictor variable in the model.[12] A separate QSAR equation is prepared for each target property when multiple dependent variables are employed. The resulting coefficients are interrelated and usually differ from those that would be obtained by examining biological properties individually. An illustration of this regression technique can be viewed in Figure 4. With high-quality biological data and compound alignments as described above, predictive QSAR-PLS models can be acquired.

Docking. When receptor structures are available, useful information can be obtained through the docking of compounds into a binding site. Figure 5 displays the general scheme of FlexiDock, a genetic algorithm-based flexible docking method. Geometry optimization produces an initial population of compounds in complex with a receptor. Each complex consists of parameters that will be optimized: torsional angles, translation, and rotational angles. Reproduction takes place when complex populations swap coordinates, crossover, and/or exhibit random changes within the complex, mutation. Duplicate checking ensures that each complex is unique; this increases the complex population diversity. Conformational modifications are then made to the reproduced compounds and an evaluation function for scoring the resulting interaction is applied to the complex. The FlexiDock scoring function is based on the Tripos force field and estimates the energy of the compound, the receptor, and the complex energy. The score is evaluated with van der Waals and the user-selected energy terms, including electrostatic, torsional, constraint, and hydrogen bonding energies; lower energy in the complex state suggests better binding. The crossover options that can be implemented when using Flexidock include: (1) successive generations, (2) the creation of new members, created *via* crossover and mutation, and (3) parents that can be selected for crossover. Fitness scores can be scaled to aid in selection.

Manuscripts

The published and unpublished manuscripts presented in this dissertation are a result of a series of studies examining neglected diseases through the employment of biological data and computational tools to examine respective parasites and relevant druggable targets.[18-21] Chapters

two through five represent four independent studies. Chapter two examines a highly diverse dataset of inhibitors for *T. brucei* P2 transporters. A QSAR-PLS model was acquired through this study and the compounds of the model were examined to gain an understanding of inhibitory compound structural importance for P2 transporter inhibition. Chapter three examines arylimidamides and their inhibitory activity against two species of *Leishmania*. This research endeavor resulted in a conservative predictive method acquired *via* predictive models employing both rigid and flexible compound alignments. Compound structural importance to activity was then assessed. Chapter four examines a dataset of diamidines and arylimidamides with respect to inhibitory activity against *T. cruzi* at two different temperatures. A pharmacophore was obtained and used to construct a predictive model. Inhibitory compound importance was then extrapolated from the model and assessed with respect to the pharmacophore at each temperature. Chapter five examines dimer polyamide compounds bound by DNA with respect to their cognate DNA sequences. Structural importance and mechanisms of binding were evaluated through docking analyses. This study provides insight into DNA-compound interactions that may be applicable for targeting parasites of the Order Kinetoplastida, since the DNA of these parasites has been identified as druggable targets.[22-24]

References

1. Boutayeb, A., Developing countries and neglected diseases: challenges and perspectives. *Int J Equity Health* **2007,** 6, 20.

2. Hopkins, A. L.; Witty, M. J.; Nwaka, S., Mission possible. *Nature* **2007,** 449, (7159), 166-9.
3. Ouellette, M.; Drummelsmith, J.; Papadopoulou, B., Leishmaniasis: drugs in the clinic, resistance and new developments. *Drug Resist Updat* **2004,** 7, (4-5), 257-66.
4. Croft, S. L., In vitro screens in the experimental chemotherapy of leishmaniasis and trypanosomiasis. *Parasitol Today* **1986,** 2, (3), 64-9.
5. Kennedy, P. G., The continuing problem of human African trypanosomiasis (sleeping sickness). *Ann Neurol* **2008,** 64, (2), 116-26.
6. *Control and surveillance of African trypanosomiasis: report of a WHO Expert Committee*; World Health Organization: Geneva, Switzerland, 1998; 30-37.
7. Bartlett, J. G.; Auwaeter, P. G.; Pham, P. A., *The Johns Hopkins ABX Guide: Diagnosis & Treatment of Infectious Diseases*. Second Edition ed.; Jones and Bartlett Publishers, Inc: 2010; 1-860.
8. Cerecetto, H.; Gonzalez, M., Anti-T. cruzi agents: our experience in the evaluation of more than five hundred compounds. *Mini Rev Med Chem* **2008,** 8, (13), 1355-83.
9. Coura, J. R., Present situation and new strategies for Chagas disease chemotherapy: a proposal. *Mem Inst Oswaldo Cruz* **2009,** 104, (4), 549-54.
10. Herwaldt, B. L., Leishmaniasis. *Lancet* **1999,** 354, (9185), 1191-9.
11. Singh, S.; Sivakumar, R., Challenges and new discoveries in the treatment of leishmaniasis. *J Infect Chemother* **2004,** 10, (6), 307-15.
12. *SYBYL Molecular Modeling Software, 8.1 ed.*, Tripos Inc.: St. Louis, MO, 2008.

13. Cramer, R. D., III; Patterson, D. E.; Bunce, J. D., Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **1988,** 110, 5959.
14. Verma, J.; Khedkar, V. M.; Coutinho, E. C., 3D-QSAR in drug design--a review. *Curr Top Med Chem* **2010,** 10, (1), 95-115.
15. Podlogar, B. L.; Ferguson, D. M., QSAR and CoMFA: a perspective on the practical application to drug discovery. *Drug Des Discov* **2000,** 17, (1), 4-12.
16. Klebe, G.; Abraham, U.; Mietzner, T., Molecular similarity indices in a comparative analysis (CoMSIA) of drug molecules to correlate and predict their biological activity. *J Med Chem* **1994,** 37, (24), 4130-46.
17. Klebe, G.; Abraham, U., Comparative molecular similarity index analysis (CoMSIA) to study hydrogen-bonding properties and to score combinatorial libraries. *J Comput Aided Mol Des* **1999,** 13, (1), 1-10.
18. Collar, C. J.; Al-Salabi, M. I.; Stewart, M. L.; Barrett, M. P.; Wilson, W. D.; de Koning, H. P., Predictive computational models of substrate binding by a nucleoside transporter. *J Biol Chem* **2009,** 284, (49), 34028-35.
19. Collar, C. J.; Lee, M.; Wilson, W. D., Setting anchor in the minor groove: in silico investigation into formamido N-methylpyrrole and N-methylimidazole polyamides bound by cognate DNA sequences *J. Chem. Inf. Model.* **2010,** Submitted.

20. Collar, C. J.; Zhu, X.; Werbovetz, K.; Boykin, D. W.; Wilson, W. D., Governing inhibition of arylimidamides against leishmaniasis: Conservative computational modeling to improve chemotherapies. In Preparation.
21. Collar, C. J.; de Souza, E. M.; Batista, D. d. G. J.; da Silva, C. F.; Daliry, A.; Wiggins, M.; Soeiro, M. d. N. C.; Tidwell, R. R.; Boykin, D. W.; Wilson, W. D., Screening for affinity: Pharmacophore and QSAR-PLS modeling of biological inhibitory data for compounds active against *Trypanosoma cruzi*. In Preparation.
22. Wilson, W. D.; Tanious, F. A.; Mathis, A.; Tevis, D.; Hall, J. E.; Boykin, D. W., Antiparasitic compounds that target DNA. *Biochimie* **2008,** 90, (7), 999-1014.
23. Hu, L.; Arafa, R. K.; Ismail, M. A.; Wenzler, T.; Brun, R.; Munde, M.; Wilson, W. D.; Nzimiro, S.; Samyesudhas, S.; Werbovetz, K. A.; Boykin, D. W., Azaterphenyl diamidines as antileishmanial agents. *Bioorg Med Chem Lett* **2008,** 18, (1), 247-51.
24. da Silva, C. F.; da Silva, P. B.; Batista, M. M.; Daliry, A.; Tidwell, R. R.; Soeiro Mde, N., The biological in vitro effect and selectivity of aromatic dicationic compounds on Trypanosoma cruzi. *Mem Inst Oswaldo Cruz* **2010,** 105, (3), 239-45.

Figures

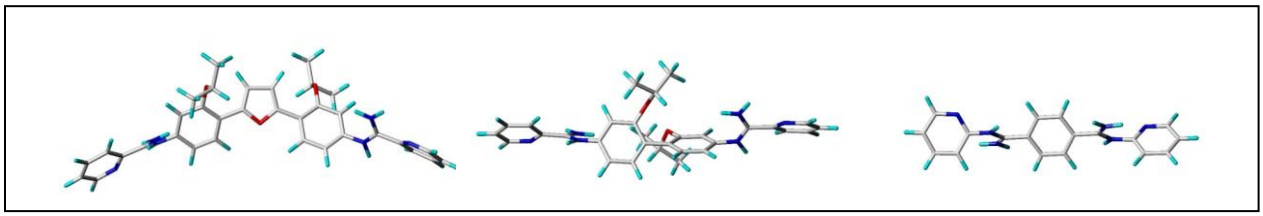


Figure 1. Examples of compounds constructed and minimized within the molecular modeling software. Minimization should include an assigned Force Field, such as Tripos, and Charges, such as Gasteiger-Huckel.

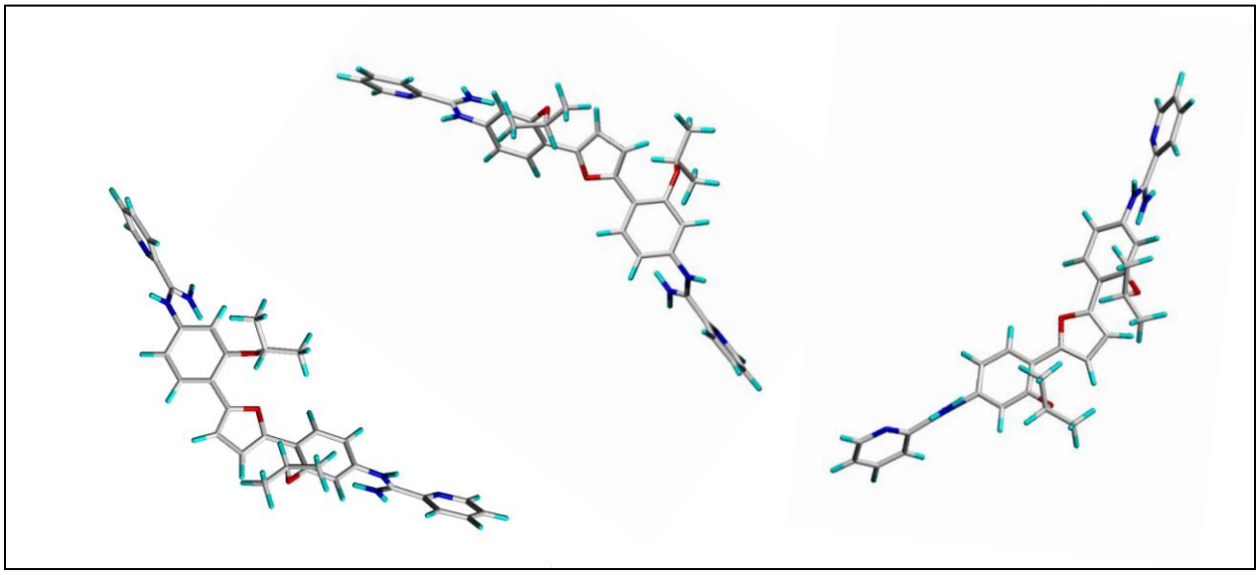


Figure 2. Possible conformations of structures can be explored through various methods.

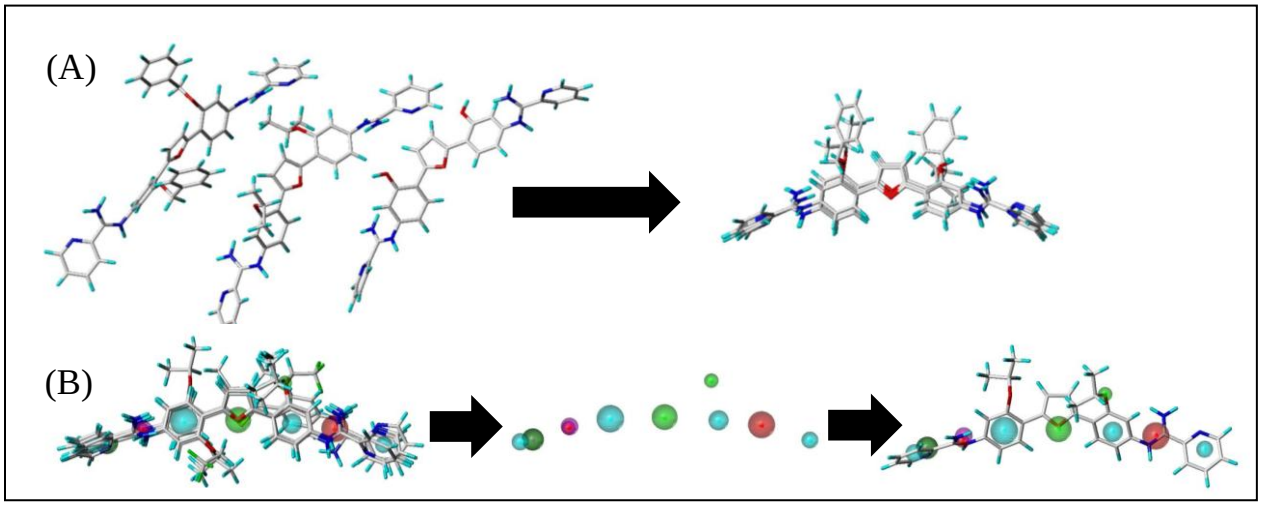


Figure 3. The three-dimensional molecular structures aligned within Cartesian space. (A) The QSAR module of SYBYL can be employed to overlay rigid low energy structures *via* individual molecule translations and/or rotations. (B) The GALAHAD module of SYBYL can be used to overlay flexible or rigid molecular structures in torsional space. The identified features are color coded: cyan for hydrophobes, magenta for donor atoms, green for acceptor atoms and red for positive nitrogens.

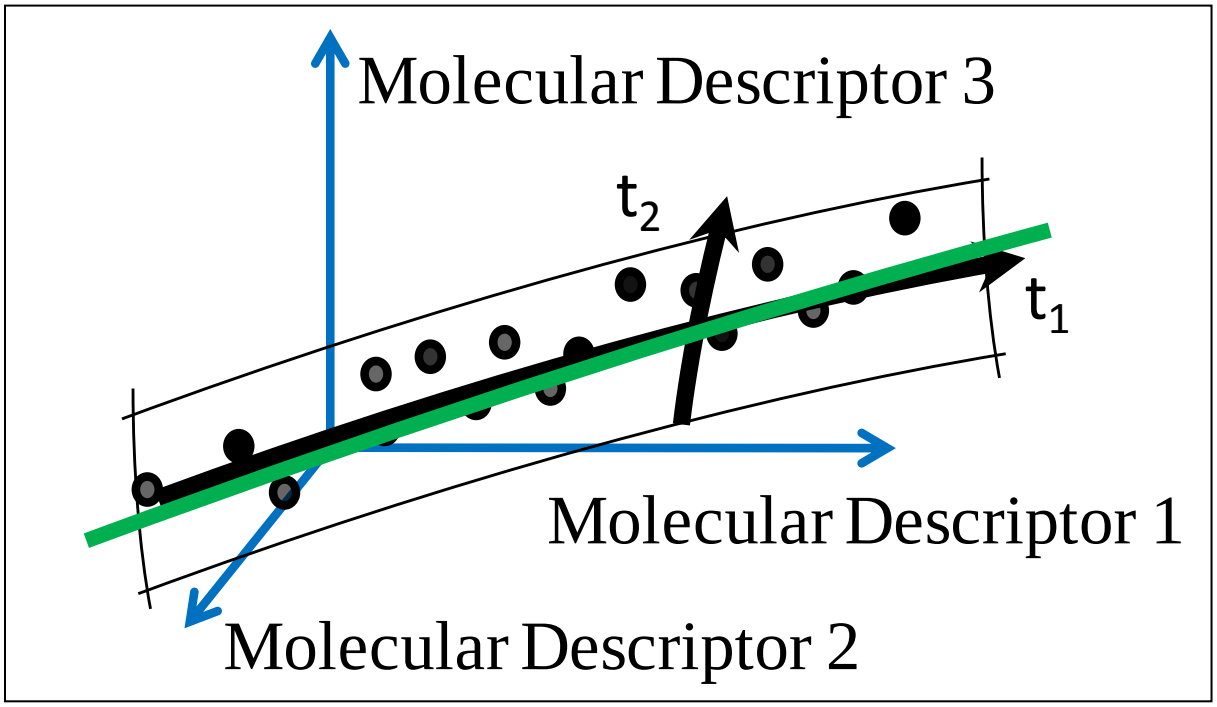


Figure 4. Partial least squares (PLS) is a regression technique that is employed to compare experimentally obtained activity values to compound molecular descriptors acquired from respective compounds. PLS results in a linear model.

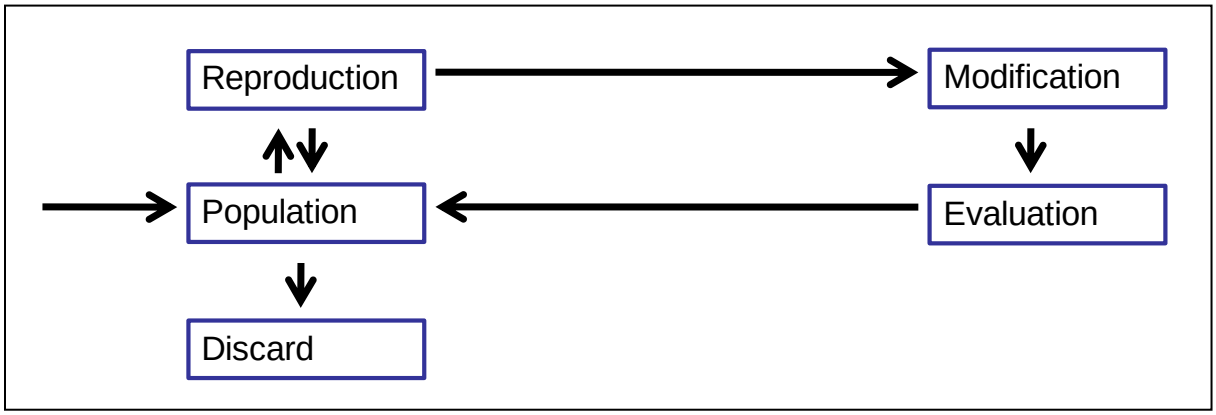


Figure 5. FlexiDock employs genetic algorithms as global optimizers to apply methods of biological evolution.

CHAPTER 2: PREDICTIVE COMPUTATIONAL MODELS OF SUBSTRATE BINDING BY A NUCLEOSIDE TRANSPORTER

PREDICTIVE COMPUTATIONAL MODELS OF SUBSTRATE BINDING BY A NUCLEOSIDE TRANSPORTER

Catharine J. Collar[1], Mohammed I. Al-Salabi[2], Mhairi L. Stewart[2], Michael P. Barrett[2], W. David Wilson[1], and Harry P. de Koning[2]

From Department of Chemistry, Georgia State University, Atlanta, Georgia, 30303[1] and the Department of Infection and Immunity, Institute of Biomedical and Life Sciences, University of Glasgow, Glasgow G12 8TA, Scotland, United Kingdom.

Running head: Modeling substrate binding by a nucleoside transporter

Address correspondence to Harry P. de Koning, Division of Infection & Immunity, Glasgow Biomedical Research Centre, 120 University Place, University of Glasgow, Glasgow G12 8TA, Scotland, United Kingdom.

The abbreviations used are: CoMFA, Comparative Molecular Field Analysis; CoMSIA, Comparative Molecular Similarity Indices Analysis; QSAR, Quantitative Structure-Activity Relationship; PLS, Partial Least Squares; SEE, Standard Error of Estimate; CPDD, Consortium for Parasitic Drug Development.

Transporters play a vital role in both the resistance mechanisms of existing drugs and effective targeting of their replacements. Melarsoprol and diamidine compounds similar to pentamidine and furamidine are primarily taken up by trypanosomes of the genus *Trypanosoma brucei* through the P2 aminopurine transporter. In standardized competition experiments with [^{3}H]adenosine, P2 transporter inhibition constants (K_i) have been determined for a diverse dataset of adenosine analogs, diamidines, Food and Drug Administration-approved compounds and analogs thereof, and custom-designed trypanocidal compounds. Computational biology has been employed to investigate compound structure diversity in relation to P2 transporter interaction. These explorations have led to models for inhibition predictions of known and novel compounds to obtain information about the molecular basis for P2 transporter inhibition. A common pharmacophore for P2 transporter inhibition has been identified along with other key structural charisteristics. Our model provides insight into P2 transporter interactions with known compounds and contributes to strategies for the design of novel antiparasitic compounds. This approach offers a quantitative and predictive tool for molecular recognition by specific transporters without the need for structural or even primary sequence information of the transport protein.

Introduction

Trypanosoma brucei are unicellular trypanosomal parasites that cause African sleeping sickness in humans and nagana in livestock. These trypanosomes are auxotrophic for purines and thus rely entirely on purine supplies salvaged from the host environment. As such, *T. brucei brucei* expresses a multitude of purine nucleoside and nucleobase transporters.[1] One of these, the *T. brucei* aminopurine P2 transporter, is unusual as a genuine nucleoside-nucleobase transporter in that it equally transports the nucleoside adenosine and the nucleobase adenine but has virtually no affinity for any other natural purines or pyrimidines.[1-3] Yet, despite this apparent high level of selectivity, it has been shown that P2 also mediates cellular uptake of the Food and Drug Administration-approved drugs melarsoprol and pentamidine,[2, 4, 5] the main veterinary trypanocides diminazene aceturate[6] and possibly isometamidium,[7] and various nucleoside drugs.[8]

The unusual nature of this transporter has led to efforts to exploit it as an efficient conduit for novel trypanocides,[9, 10] but this requires the identification of the exact pharmacophore as well as the physical limitations on size and charge distribution of the extracellular binding site of the transporter. From the structural similarities between known P2 substrates, it could be concluded early on that the so-called amidine motif of adenine, *i.e.* N(1)=C(6)−NH_2 (see Figure 1), was very likely to play a major role in the high affinity interaction with the transporter.[3, 11] However, quantitative information or three-dimensional models explaining the high affinity binding, by one transporter, of such diverse molecules as adenosine (Figure 1A),[2, 3] stilbamidine (Figure

1C),[12] melarsoprol (Figure 1F),[2, 3] and even isometamidium (Figure 1G),[7] have not been available. The apparent broad selectivity has been all the more intriguing for the highly similar transport efficiencies of P2 for adenosine and adenine, a most unusual feature for nucleoside transporters.[1]

To construct a predictive and quantitative model of P2-substrate interactions, we determined the K_i values of a large number of highly diverse potential inhibitors, with affinities ranging over several orders of magnitude, through competition experiments with radiolabeled adenosine. These values and structures were then employed for a computational modeling approach to gain more information about the molecular basis for P2 transporter inhibition. The resulting model can be used to evaluate the affinity of the P2 transporter for existing and novel compounds *in silico*, potentially aiding in the development of novel and selectively targeted trypanocides. More important yet, this strategy allows robust three-dimensional insights into transporter-ligand binding while not requiring knowledge of the structure, or indeed the sequence, of a transporter and can be applied to any solute transport mechanism for which uptake or binding experiments can be routinely performed.

Experimental Procedures

Transport of [^{3}H]Adenosine by Bloodstream Forms of T. brucei. Bloodstream forms of *T. brucei brucei* strain 427 were taken from stocks in liquid nitrogen and injected in adult female Wistar rats, from which they were harvested by exsanguination by cardiac puncture at peak

parasitaemia. Parasites were isolated from the blood by elution over a DE52 column (Whatman)[13] and washed twice in assay buffer (AB: 33 mM HEPES, 98 mM NaCl, 4.6 mM KCl, 0.3 mM $CaCl_2$, 0.07 mM $MgSO_4$, 5.8 mM NaH_2PO_4, and 14 mM glucose, pH 7.3). Cells were resuspended in this buffer at approximately 10^8 cells/ml prior to use in transport experiments. Cell counts were performed using a haemocytometer. Transport of [^{3}H]adenosine (20-40 Ci/mmol; Amersham Biosciences) was performed exactly as described previously,[14] in the presence of 250 μM inosine to block the P1 adenosine uptake system. Briefly, 100 μl of 50 nM [^{3}H]adenosine, mixed with various concentrations of nonradiolabeled test compounds, was added to 100 μl AB containing 10^7 trypanosomes and incubated at room temperature for 30 s, within the linear phase of uptake.[3] Uptake was terminated by the addition of 1 ml of ice-cold assay buffer containing 1 mM adenosine followed by immediate centrifugation through an oil layer to separate cells from external radiolabel. The amount of radiolabeled adenosine inside the cell was then determined using a scintillation counter and corrected for externally associated label as described previously.[14] A plot of inhibitor concentration *versus* adenosine uptake rate (expressed as pmol(10^7 cells^{-1}s^{-1})) yielded sigmoidal curves with Hill coefficients of approximately −1, consistent with monophasic competitive inhibition (Prism 4.0; GraphPad). Inhibition constants were calculated from the EC_{50} values, using the Cheng-Prusoff equation as described previously.[12]

Inhibitor Dataset. Compounds were acquired from several academic laboratories as well as purchased from various commercial sources. Their respective *in vitro* transport activities

along with the compound names and sources are shown in Supplemental Table 1 (Appendix A). Employing the formula p*Ki* = −log(*Ki*), the *Ki* μM values for the 112 compounds were converted to corresponding p*Ki* values. The p*Ki* values for this training set span more than 4 log units.

Software. All 112 compounds were constructed *in silico* with the SYBYL 8.1[15] software package on a Fedora Core 5 Linux workstation. Compound structures were minimized to convergence using a conjugate gradient of 0.01 kcal/(mol Å) and a maximum of 10^4 iterations employing the Tripos force field with Gasteiger-Hückel charges. A three-dimensional cubic lattice with 2 Å grid spacing in all directions was created to analyze compounds that were aligned as described below. No improvement was seen in the models when the grid spacing was reduced to 1 Å.[16]

Initial Alignment. Through the implementation of the SYBYL software alignment modules, the compounds were three-dimensionally arranged by an initial analysis of structurally and chemically related atoms. Algorithm generated alignment was performed using the align database command, whereas the atom-to-atom alignment implemented the match feature of the alignment tools. The algorithm alignment took place first by employing similar backbone structures so that the majority of similar compounds were overlaid in the same molecular space. Structurally related compounds were then moved into separate databases. The compounds that belonged to the same structural classes, but which varied in atom types or had slight structural differences, were placed into respective databases and aligned to the most structurally related

compound using atom-to-atom alignment. Seven optimum databases of compounds resulted from initial alignment.

When more rigid compound structures, consisting of a larger number of atoms, were selected as scaffolds for alignments a greater number of databases were created. These databases lacked the variation necessary to form Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) models for predictability. Also, when the databases were aligned by less rigid scaffolds, consisting of a smaller number of atoms, fewer models resulted, and the models produced were not statistically significant in terms of q^2_{cv}. The best models were obtained when compounds were aligned by the carbons of common compound backbones. These scaffolds for alignment were obtained from the compounds displayed in Figure 1: dataset A, adenine; dataset B, furamidine; dataset C, stilbamidine; dataset D, pentamidine; dataset E, 1,1'-(nonane-1,9-diyl)diguanidine; dataset F, melarsoprol; and dataset G, isometamidium. Datasets E-G are comprised of four, seven and four compounds, respectively. The alignment for these last three datasets can be viewed in Supplemental Figure 1 (Appendix A). These databases together consist of less than 8% of the total compounds. Because the purpose of the initial alignment was to determine the pharmacophore for the final alignment, only initial datasets A-D were evaluated through statistics and contour maps. All 112 compounds were included in the final pharmacophore models.

Multiple Regression Analysis. CoMFA and CoMSIA Quantitative Structure-Activity Relationship (QSAR) models were generated for molecular databases through a Partial Least

Squares (PLS) multiple regression analysis with molecular descriptors as independent variables and the p*Ki* values as dependent variables. Statistical significance in the form of q^2_{cv} was assessed through the leave-one-out cross-validation method. The number of components (n) was determined by the smallest predicted error sum of squares, a value that does not always correspond to the highest correlation coefficient (q^2) value. Further statistical significance assessment was preformed for the final model using 10-fold cross-validation. The values obtained from the 10-fold cross-validation assessment are averages of ten trials implementing random compound selection. Column filtering did not improve the signal to noise ratio.[15]

Molecular Descriptors. There are two CoMFA molecular descriptors. The steric van der Waals interaction and the electrostatic Coulombic interaction descriptors were calculated at each lattice intersection using a probe, an sp^3 carbon atom with a formal +1 charge. Standard scaling and default energy cutoffs were employed. There are five CoMSIA molecular descriptors. Steric, electrostatic, hydrophobic, hydrogen bond donor and hydrogen bond acceptor descriptors were calculated using a standard probe: 1 Å radius, +1 charge, +1 hydrophobicity, +1 hydrogen bond donor, and +1 hydrogen bond acceptor. Steric descriptors are related to the third power of the atomic radii. Electrostatic descriptors are derived from partial atomic charges. Hydrophobic descriptors are derived from atom-based parameters. Hydrogen bond donor and acceptor atoms are derived from experimental values.

Three-Dimensional Contour Analysis. The interactions of CoMFA and CoMSIA descriptors were visualized through the mapping of the product standard deviation with respect

to molecular descriptor values and coefficients (StDev*Coeff) at each lattice point. For the initial models, the default levels of contour by contribution were employed as follows: 80% for a favored region and 20% for a disfavored region. Data were analyzed, and a common pharmacophore was identified. The compounds of the final pharmacophore model were further analyzed through a contour by actual analysis, where the software output assisted in the determination of proper ranges for assigned values of favored and disfavored contour regions.

Pharmacophore Model. Common contours for the initial QSAR models were identified through the analysis of favored and disfavored contour regions. The alignment of such contours aided in the identification of a final pharmacophore. All compounds were realigned, and the final models were constructed.

Results

As seen in Supplemental Table 1 (Appendix A), this study employs 112 compounds acquired from several academic and industry locations. These compounds all exhibit some level of inhibitory activity for the *T .brucei brucei* P2 transporter. For large datasets of compounds with known activity values, it is possible to employ computational biology to investigate the molecular basis of their activity in terms of structural contributions to *Ki* values. Predictive models can then be constructed, and important interactions can be identified. Because a large number of diverse compounds are in our database, a two-step procedure was used to establish a final model.

Initial QSAR Models. As a first step, compounds were obtained in their minimal energy conformation by using standard molecular mechanics energy minimization methods with the Tripos force field. Compound alignment by similar atoms of backbone structures initially separated the 112 compounds into seven databases, although the majority of the compounds resided in four of the sets. The datasets with the majority of compounds were used for initial PLS modeling. Table 1 displays the total number of compounds in each dataset, the n used in PLS, and the statistics for each model as follows: cross-validated q^2 (q^2_{cv}), the standard error of estimate (SEE), the coefficient of determination (r^2) and the F statistic. When q^2 is greater than 0.5, a model is said to have predictability better than chance; however, it is also important that the r^2 value is near one, the SEE is small, and the F statistic is large.[15] The r^2 is a positive value between zero and one; with one being the best correlation and zero being no correlation. The SEE is a measure of the accuracy of the predictions. The F statistic is used in comparing the variance between the experimental and predicted values; a larger value indicates a more statistically significant model.

The average statistics for the initial four models with CoMFA molecular descriptors are as follows: q^2_{cv} equal to 0.64; SEE equal to 0.23; r^2 equal to 0.95; and F statistic equal to 123. Similarly, the average statistics for the four models with CoMSIA molecular descriptors were as follows: q^2_{cv} equal to 0.58; SEE equal to 0.26; r^2 equal to 0.92; and F statistic equal to 130. Although the models with CoMFA and CoMSIA molecular descriptors were comparable, the ones with CoMFA molecular descriptors display better overall potential for analysis of molecular

descriptor contribution by contour maps. This is primarily due to the simplicity of two *versus* five molecular descriptors.

Contour maps of CoMFA molecular descriptor contribution were generated for each model (Figure 2). The electrostatic interactions are shown as red and blue contours, and the steric interactions are displayed as green and yellow contours. Increasing partial positive charge is favored in blue regions, and increasing partial negative charge is favored in red regions, whereas increasing bulk in substituents is favored in green regions and disfavored in yellow regions.

The red, blue, yellow, and green regions were then analyzed to find common alignment features of structures that are of importance for the final, combined pharmacophore alignment. Red regions of dataset A are in the areas above C6, below N9, and beside the imidazole ring of adenine, while those of datasets B-D were localized to a single location most often than not on the backbone structure. The red contours of datasets A-D can be aligned in several ways to one another; thus, this descriptor alone is not enough to find the final pharmacophore for alignment. The blue regions were most commonly found in areas of $N(R_1)=C(R_2)-NH(R_3)$, where R_3 is usually H. The alignment was much improved with the inclusion of both the red and blue regions and further enhanced by the addition of the yellow and green regions. Yellow contour regions can be reduced by realignment of compounds into green regions. The yellow regions for dataset A are small in relation to all other contours, and reside near the 2'- and 3'- hydroxy groups of the ribose moiety. Dataset B exhibited yellow contours on both ends of the furamidine backbone, whereas dataset C displayed a yellow contour only at one end of the stilbamidine backbone. The

areas of yellow contour appear most at regions that consist of several compounds with substituents that are not precisely aligned, either because they differ largely in structure or because the backbone allows for deviations in the alignment. Dataset D consisted of yellow regions in the areas consisting of compounds that were longer than pentamidine and/or that did not align fully to the pentamidine backbone. Green regions of dataset A were shown above C6 and next to bond C8/N9 of the adenine backbone, whereas the green contours of dataset B appear near and encompassing the phenyl with the most precise alignment. Datasets C consists of green contour near the most precise alignment of the compounds. For dataset D, green contours were located in areas that were not precisely aligned to the pentamidine structure. The green and yellow contours of dataset D both reside in areas of structural deviation; however, the green appears nearest the aromatic linking oxygen and the unaligned amidines.

The identification of important structural features, described above, made it possible to realign all 112 compounds, primarily by the common $N(R_1)=C(R_2)-NH(R_3)$ structure found in the blue contour regions and secondarily by the other contour regions. The red regions of the four main datasets overlapped strongly, whereas the yellow regions of datasets B-D can be aligned to green regions of dataset A. The large compounds of dataset A also had to be realigned. Figure 3A displays the alignment of all 112 compounds with adenine displayed in purple and Figure 3B zooms in on the location of the adenine now with the purple displayed as transparent, and this clearly shows the pharmacophore alignment.

Final Pharmacophore Model. Aligned by the $N(R_1)=C(R_2)-NH(R_3)$ structure with respect to contour regions, as described above, compounds were then employed for PLS modeling. As before, CoMFA and CoMSIA models were generated and examined for statistical significance. The two models each consisted of 112 compounds but use different molecular descriptors and a different n. Although the q^2_{cv} values are similar, the remaining statistics are not; the model with CoMFA molecular descriptors achieved a higher level of confidence than the model with CoMSIA molecular descriptors (Table 2). To further validate these models, 10-fold cross-validation was performed. The $q^2_{10\text{-Fold}}$ values for the models with CoMFA and CoMSIA molecular descriptors were 0.56 and 0.54, respectively. These values, along with the rest of the statistics, indicate statistical significance within each model.

The calculated predictions of the models formed from the dataset with 112 compounds exhibit linear relationships with the experimental *Ki* values (Figure 4). Predictions from the model with CoMSIA molecular descriptors are somewhat scattered, especially at high affinity, whereas the model with CoMFA molecular descriptors produces more linear p*Ki* predictions, especially for compounds with high affinity for the P2 adenosine transporter (Figure 4). The r^2 values for the linear relationships are 0.95 for the model with CoMFA molecular descriptors and 0.86 for the model with CoMSIA molecular descriptors.

These models can be further evaluated through examination of the final contour maps. Although it is useful to analyze models as a whole to gain information about a possible pharmacophore, once a pharmacophore model is obtained, much more information can be

gathered by evaluating the contour regions of individual compounds within the model. Because the model with CoMFA molecular descriptors is outperforming the model with CoMSIA molecular descriptors, the focus of this analysis will remain on the contours of the model with CoMFA molecular descriptors. As before, the steric contributions are displayed in yellow and green while the electrostatic contributions are shown in red and blue.

The overall contour regions from the initial model have changed significantly with realignment and incorporation of all 112 compounds. These changes appear most dramatic when looking at individual compounds. In the initial models, each compound contributed roughly 2.8-6.3 percent. This was due to similar compounds being aligned by a common backbone scaffold and their being only 16-36 compounds in each dataset; 1 in 36 is approximately 2.8 percent and 1 in 16 is about 6.3 percent. This percent of contribution is much larger than the final model, where 1 in 112 compounds is roughly 0.89 percent. It is also important to note that a larger quantity of compounds with similar backbones will have a significant effect on the contribution. Hence, based on initial models, the compounds with the adenosine scaffold structure should contribute the most. There are 36 of these compounds. Those with the pentamidine and stilbamidine scaffolds are similar and align to one another well within the final model. There are 32 of these compounds, whereas there are 29 compounds related to furamidine.

From close observations of compound structure relationships in the form of contour maps, it is possible to determine where partial charge addition or subtraction to substituents could improve compound interactions with the P2 adenosine transporter. The evolutionary

process by which this model calculates predictions can be viewed through the evaluation of contour regions and experimentally determined *Ki* values (Figure 5). The *Ki* of 2-aminopyridine is 14 μM. When an amino group is added into the favorable steric and positive electrostatic contour regions to form 4,6-diaminopyrimidine the *Ki* becomes 3.2 μM. Note that the amino group has a partial positive charge. This amino group addition thus results in improved affinity. When the additional groups, which reside in even more favorable contour regions, are added to the compound structure, the *Ki* value becomes even smaller. Adenine is an example of a compound with groups residing in favorable contour regions. This compound has a *Ki* of 0.30 μM. When a compound interacts with both positive and negative contour regions, the *Ki* increases; the *Ki* value for adenosine, for example, increases three-fold relative to adenine as a result of the bulky ribose group. The evolutional process taken when using these potentials to design compounds for synthesis is quite similar to the progression shown in Figure 5. It is important to make small changes and evaluate how the designed compound will fit within the steric and electrostatic potentials assigned by the model.

Other important compounds to evaluate with this model are the pentamidine-, furamidine-, and melarsoprol-like compounds (Figure 6). Pentamidine, furamidine, and melarsoprol all have good affinity for the P2 transporter with respective *Ki* values of 0.37, 1.19 and 0.54 μM. Contour regions of pentamidine, furamidine, and melarsoprol are displayed in Figure 6. These regions display several areas where some steric bulk and partial positive charge can be added to improve affinity for the P2 adenosine transporter. A loss of affinity will occur if

bulky substituents interact with the unfavorable yellow contour regions and/or if positive charge interacts with the red contour regions.

With pentamidine, which is a very flexible compound, the final pharmacophore model yellow contours display most central atoms to be suitable for substituent addition; however, the area nearest the pharmacophore should not be modified. Melarsoprol is a more rigid structure, though rotation can occur throughout the compound. There can be rotation between the melamine ring and the phenyl and between the phenyl and the dithiarsolan ring. For this compound, the yellow contours reside near the melamine and the phenyl. This suggests that a loss of affinity may result from substituent addition to the atoms in these regions. Furamidine is a much more rigid and curved structure. For this structure, the yellow contours are much more abundant near the phenyls and yet away from the furan and the amidines. This is even clearer when the compound and its contour are viewed in three-dimensional space. The areas where yellow contours do not exist are optimum for substituent modification.

The red contours encompass both pentamidine and furamidine, whereas blue contours surround melarsoprol. The blue contours appear to be based on the partial charge distribution. For the diamidine compounds the partial charge distribution is strongly localized at the amidines. This appears beneficial for binding to the transporter; however, it is evident that more charge to an amidine location will not improve binding. Instead a partial charge distribution that is shared within a ring structure appears to be more advantageous. This is seen in the melamine-like

structure of melarsoprol. Findings suggest that additional charge, which is less localized, may be able to improve binding of diamidine compounds.

Discussion

The efficacy of many drugs is determined to a large extent by the processes that govern their uptake into the cell or into the cellular compartment that is the site of action.[7, 17-19] These processes obviously include transporters for water-soluble drugs but even rates of diffusion for lipophilic drugs. An example of the latter is chloroquin, which as a weak base diffuses across several membranes before it reaches the *Plasmodium falciparum* food vacuole where it is trapped by protonation and fatally inhibits heme polymerization.[20, 21] Equally, efflux systems such as ATP-binding cassette transporters and the *P. falciparum* CRT1 channel-like protein have been implicated in resistance to drugs ranging from antibiotics and antiparasitics to antineoplastic drugs.[22, 23] As such, detailed insights into the processes that determine drug flux across the (plasma) membranes of target cells are vital for the rational optimization of drug activity and both the prevention and bypassing of drug resistance.

It is of pivotal importance that we gain insight into the molecular mechanisms by which transporters bind and thus select their substrates as this would allow us to construct models with predictive value, which would allow us to optimize substrate design. Although *in silico* screening of virtual libraries and predictions of substrate affinity are now possible for proteins with known or computable structure,[24-26] this is not ordinarily possible for transporters as very few structures

have been obtained, and the protein structures, with usually 10–12 transmembrane domains, are highly complex and extremely difficult to crystallize, although there have recently been some notable successes, mostly with prokaryotic membrane proteins.[27-29] One approach is to use the few known transporter structures as scaffolds for other transporters, by a computational process called fold recognition or threading. We recently obtained a model for the *T. brucei brucei* nucleobase transporter NBT1 by this process and validated it by site-directed and random mutagenesis.[30] The creation of a structural model of the closely related *Leishmania donovani* LdNT1.1 nucleoside transporter by *ab initio* calculation was also very recently reported.[31] Although these approaches did produce approximate models for the overall structure of the transporters and identified key amino acid residues, they allow at best limited prediction of substrate selection, and only if the amino acids involved in binding have been separately identified. Thus, with the current technologies, it is exceedingly difficult to obtain the required functional insights with the protein structure as a starting point.

A radically different approach was pioneered some time ago to study purine transport in *T. brucei brucei* by systematically altering the substrate and calculating inhibition constants, *Ki*, and from there binding energy ΔG^0.[1, 18] This method was used to explain substrate preferences of purine and pyrimidine transporters in *T. brucei brucei*,[32] *Leishmania major*,[33, 34] *Toxoplasma gondii*,[35] *Leishmania mexicana*,[36] as well as the human NBT1 nucleobase transporter,[14] human concentrative nucleoside transporters,[37] and human equilibrative nucleoside transporters[38] with semi-quantitative models of substrate binding that did not require any structural or genetic

information about the transport protein. However, this method still did not allow genuinely quantitative or three-dimensional predictions nor was it suitable for screening virtual libraries.

In this study, we have adapted the method to address the above issues; energy-minimized three-dimensional structures of 112 compounds with experimentally obtained binding affinities for the TbAT1/P2 transporter were employed through the use of CoMFA and CoMSIA molecular descriptors for PLS model regression construction and analysis. The various molecules were preliminarily aligned by their common structural and chemical features, resulting in four datasets of compounds, Figure 2, A-D, large enough for individual model formation and analysis. This was followed by optimized alignment of all 112 compounds using four molecular descriptor contour potentials, negative and positive steric and electrostatic, as a guide. This has generated an *in silico* computational model into which new molecules can be entered to arrive at a reliable estimate of binding energy. This constitutes a first computational approach to the design of novel ligands for the TbAT1/P2 transporter and allows for *in silico* evaluation of large numbers of known and novel compounds as substrates. The computational analysis was validated to be statistically significant using leave-one-out cross-validation and 10-fold cross-validation, as well as by other statistics and the internal predictability of this model, as displayed in Figure 4.

The contour profiles of steric and electrostatic factors also allow fundamental insights into how various ligands interact with the transporter binding pocket. The P2 transporter, with its highly unusual substrate profile and involvement in drug transport and resistance,[2-5, 11, 39] was

chosen for this study to gain insight into how a transporter that is on the one hand completely selective for adenine and adenosine only (out of all nucleosides and nucleobases) can also bind molecules as diverse as isometamidium, melarsoprol, and furamidine with similar affinity. Previous studies already identified the "amidine" motif formed by R_1−N1=C6(R_2)−NH_2 of adenine as the main motif responsible for P2 binding,[3, 11] and it was further argued that the positive charge on N9 of adenine and adenosine, as well as the aromaticity of the purine, also makes important contributions to the high substrate affinity.[3, 18]

The calculated substrate-transporter interaction contours for adenine and adenosine in Figure 5 now allow us to evaluate these earlier conclusions against the advanced modeling approach employed in this study. Figure 5 identifies four substrates that have a partial positive charge on the position of the amino group of 2-aminopyridine/adenine/adenosine as essential for optimal binding. Similarly, a partial negative charge is strongly favored at position 1, along with a positive charge at positions 8 and 9, whereas there is no clear electrostatic preference at positions 3 and 7 or most of the ribose moiety, except perhaps a preference for a positive charge at the 2'-position. Large substitutions are indicated as unfavorable in positions 1, 2, 8 and 2', and at the 6-amino group of adenosine (Figure 5, yellow indicators), but the position of the ribose group does not appear to be restricted with respect to further expansion/elongation, in line with the positioning and high affinity of the long diamidines.

The above interpretation of the CoMFA and CoMSIA models is entirely consistent with the experimentally obtained ΔG^0 values listed in Supplemental Table 1 (Appendix A). For

instance the importance of the partial negative charge on position 1, presumably as hydrogen bond acceptor, is demonstrated by the reduced affinity of 1-deazaadenosine *versus* adenosine ($\delta(\Delta G^0)$ = 9.7 kJ/mol) and of 1-deazapurine *versus* purine ($\delta(\Delta G^0)$ = 4.9 kJ/mol). Similarly, the positive charge provided by the 6-position amine is quantified by comparison of purine riboside with adenosine ($\delta(\Delta G^0)$ = 7.3 kJ/mol), purine with adenine ($\delta(\Delta G^0)$ = 10.2 kJ/mol) and 6-chloropurine riboside with adenosine ($\delta(\Delta G^0)$ = 7.0 kJ/mol). As shown in Figure 7, this gives estimates of contributions of 9.7 and 8.2 kJ/mol for the N1 and 6-amino groups, respectively. The loss of both these groups should thus result in a loss of binding energy of approximately 16 kJ/mol and this was demonstrated by comparing 2'-deoxyinosine with 2'-deoxyadenosine ($\delta(\Delta G^0)$ = 16.3 kJ/mol) and 1-deazapurine with adenine ($\delta(\Delta G^0)$ = 15.1 kJ/mol). The strong contribution from N9 likewise follows from comparing 9-deazaadenosine with adenosine and 4,6-diaminopyrimindine with 2-aminopyridine ($\delta(\Delta G^0)$ = 6.4 and 5.7 kJ/mol, respectively). The relative unimportance of positions N3 and N7 was demonstrated using 3-deazaadenosine and 7-deazaadenosine, respectively, as catalogued in Supplemental Table 2 (Appendix A), which also lists relative affinities for compounds with substitutions at positions 2 and 8.

Finally, a substantial contribution to binding is made through interactions between the aromatic purine or benzamidine moieties and amino acids in the transporter binding pocket, through π-π stacking with aromatic residues, cation-π bonding or amino-aromatic interactions.[40] Although this cannot be directly demonstrated by the use of "nonaromatic purines", which would have a completely different three-dimensional structure, uniquely for P2 this can be shown and

quantified by comparing aromatic and nonaromatic diamidines (Supplemental Table 2, Appendix A). The diagram in Figure 7 summarizes these data in the form of an interaction diagram between P2 transporter and adenosine. This figure, gained from experimental data and using a previously validated approach,[1, 18] is in close agreement with data presented in Figure 5 based on the predictive PLS regression model. It is important however to be clear that both modeling approaches (Figures 5 and 7) are predictive with respect to substrate binding rather than translocation, *i.e.* it does not predict transport efficiency for any individual substrate. This limitation is not inherent to the computational approach, rather it is the result of using *Ki* values (transport inhibition through extracellular binding) instead of Michaelis-Menten constants (K_m and V_{max} values, determined from measurement of transport) as input for the models. A similar approach as followed here could predict transport, but it would have required radiolabeled analogues of all the compounds used in the study, and this was not feasible. We also would not wish to suggest that efficient uptake by a pathogen is sufficient to ensure efficacy of a potential therapeutic agent, as this requires optimal interaction with the intended intracellular target as well. In summary, we have developed and validated a novel computational approach to analyze, explain, and predict the interactions between transporters and their substrates that does not require prior knowledge of transporter structure or indeed primary sequence.

Acknowledgments

This work was supported by the Bill and Melinda Gates Foundation through the Consortium for Parasitic Drug Development (CPDD) (to WDW, MB and HdK) and by the

Georgia State University Molecular Basis of Disease Fellowship and the David W. Boykin Graduate Fellowship in Medicinal Chemistry (to CJC). The authors are greatly indebted to the following persons who generously contributed compounds to this study: Dr Philip Blower (University of Kent at Canterbury, UK), Professor David Boykin (Georgia State University, GA, USA), Professor Bernard Bouteille (Limoges, France), Dr Christophe Dardonville, (Instituto de Química Médica; Madrid, Spain), Professor Alan Fairlamb (University of Dundee, UK), Professor Ian Gilbert (University of Dundee, UK), Professor Achiel Haemers (University of Antwerp, Belgium), Professor Simon Jarvis (University of Westminster; London, UK), Professor Gerrit-Jan Koomen, University of Amsterdam, The Netherlands), Professor Mahmoud el Kouni (University of Alabama at Birmingham; Birmingham, AL, USA), Dr Paul O'Neil (University of Liverpool, UK), Professor Katherine Radtke-Seley (University of Maryland, Baltimore Co, MA, USA), Professor Richard Tidwell (University of North Carolina Chapel Hill, NC, USA).

References

1. de Koning, H. P.; Bridges, D. J.; Burchmore, R. J., Purine and pyrimidine transport in pathogenic protozoa: from biology to therapy. *FEMS Microbiol Rev* **2005,** 29, (5), 987-1020.
2. Carter, N. S.; Fairlamb, A. H., Arsenical-resistant trypanosomes lack an unusual adenosine transporter. *Nature* **1993,** 361, (6408), 173-6.

3. de Koning, H. P.; Jarvis, S. M., Adenosine transporters in bloodstream forms of Trypanosoma brucei brucei: substrate recognition motifs and affinity for trypanocidal drugs. *Mol Pharmacol* **1999,** 56, (6), 1162-70.
4. Carter, N. S.; Berger, B. J.; Fairlamb, A. H., Uptake of diamidine drugs by the P2 nucleoside transporter in melarsen-sensitive and -resistant Trypanosoma brucei brucei. *J Biol Chem* **1995,** 270, (47), 28153-7.
5. Matovu, E.; Stewart, M. L.; Geiser, F.; Brun, R.; Maser, P.; Wallace, L. J.; Burchmore, R. J.; Enyaru, J. C.; Barrett, M. P.; Kaminsky, R.; Seebeck, T.; de Koning, H. P., Mechanisms of arsenical and diamidine uptake and resistance in Trypanosoma brucei. *Eukaryot Cell* **2003,** 2, (5), 1003-8.
6. de Koning, H. P.; Anderson, L. F.; Stewart, M.; Burchmore, R. J.; Wallace, L. J.; Barrett, M. P., The trypanocide diminazene aceturate is accumulated predominantly through the TbAT1 purine transporter: additional insights on diamidine resistance in african trypanosomes. *Antimicrob Agents Chemother* **2004,** 48, (5), 1515-9.
7. de Koning, H. P., Transporters in African trypanosomes: role in drug action and resistance. *Int J Parasitol* **2001,** 31, (5-6), 512-22.
8. Geiser, F.; Luscher, A.; de Koning, H. P.; Seebeck, T.; Maser, P., Molecular pharmacology of adenosine transport in Trypanosoma brucei: P1/P2 revisited. *Mol Pharmacol* **2005,** 68, (3), 589-95.

9. Baliani, A.; Bueno, G. J.; Stewart, M. L.; Yardley, V.; Brun, R.; Barrett, M. P.; Gilbert, I. H., Design and synthesis of a series of melamine-based nitroheterocycles with activity against Trypanosomatid parasites. *J Med Chem* **2005,** 48, (17), 5570-9.
10. Stewart, M. L.; Boussard, C.; Brun, R.; Gilbert, I. H.; Barrett, M. P., Interaction of monobenzamidine-linked trypanocides with the Trypanosoma brucei P2 aminopurine transporter. *Antimicrob Agents Chemother* **2005,** 49, (12), 5169-71.
11. Barrett, M. P.; Fairlamb, A. H., The biochemical basis of arsenical-diamidine crossresistance in African trypanosomes. *Parasitol Today* **1999,** 15, (4), 136-40.
12. De Koning, H. P., Uptake of pentamidine in Trypanosoma brucei brucei is mediated by three distinct transporters: implications for cross-resistance with arsenicals. *Mol Pharmacol* **2001,** 59, (3), 586-92.
13. Lanham, S. M., Separation of trypanosomes from the blood of infected rats and mice by anion-exchangers. *Nature* **1968,** 218, (5148), 1273-4.
14. Wallace, L. J.; Candlish, D.; De Koning, H. P., Different substrate recognition motifs of human and trypanosome nucleobase transporters. Selective uptake of purine antimetabolites. *J Biol Chem* **2002,** 277, (29), 26149-56.
15. *SYBYL Molecular Modeling Software, 8.1 ed.*, Tripos Inc.: St. Louis, MO, 2008.
16. Kimand, S. K.; Jacobson, K. A., Three-dimensional quantitative structure-activity relationship of nucleosides acting at the A3 adenosine receptor: analysis of binding and relative efficacy. *J Chem Inf Model* **2007,** 47, (3), 1225-33.

17. Cascorbi, I., Role of pharmacogenetics of ATP-binding cassette transporters in the pharmacokinetics of drugs. *Pharmacol Ther* **2006,** 112, (2), 457-73.
18. Luscher, A.; de Koning, H. P.; Maser, P., Chemotherapeutic strategies against Trypanosoma brucei: drug targets vs. drug targeting. *Curr Pharm Des* **2007,** 13, (6), 555-67.
19. Molina-Arcas, M.; Trigueros-Motos, L.; Casado, F. J.; Pastor-Anglada, M., Physiological and pharmacological roles of nucleoside transporter proteins. *Nucleosides Nucleotides Nucleic Acids* **2008,** 27, (6), 769-78.
20. Bray, P. G.; Janneh, O.; Raynes, K. J.; Mungthin, M.; Ginsburg, H.; Ward, S. A., Cellular uptake of chloroquine is dependent on binding to ferriprotoporphyrin IX and is independent of NHE activity in Plasmodium falciparum. *J Cell Biol* **1999,** 145, (2), 363-76.
21. Fitch, C. D., Ferriprotoporphyrin IX, phospholipids, and the antimalarial actions of quinoline drugs. *Life Sci* **2004,** 74, (16), 1957-72.
22. Borst, P.; Elferink, R. O., Mammalian ABC transporters in health and disease. *Annu Rev Biochem* **2002,** 71, 537-92.
23. Bray, P. G.; Mungthin, M.; Hastings, I. M.; Biagini, G. A.; Saidu, D. K.; Lakshmanan, V.; Johnson, D. J.; Hughes, R. H.; Stocks, P. A.; O'Neill, P. M.; Fidock, D. A.; Warhurst, D. C.; Ward, S. A., PfCRT and the trans-vacuolar proton electrochemical gradient: regulating the access of chloroquine to ferriprotoporphyrin IX. *Mol Microbiol* **2006,** 62, (1), 238-51.

24. Congreve, M.; Murray, C. W.; Blundell, T. L., Structural biology and drug discovery. *Drug Discov Today* **2005,** 10, (13), 895-907.

25. Muegge, I.; Oloff, S., Advances in virtual screening. *Drug Discov Today Tech* **2006,** 3, 405-411.

26. Song, C. M.; Lim, S. J.; Tong, J. C., Recent advances in computer-aided drug design. *Brief Bioinform* **2009,** 10, (5), 579-91.

27. Abramson, J.; Smirnova, I.; Kasho, V.; Verner, G.; Kaback, H. R.; Iwata, S., Structure and mechanism of the lactose permease of Escherichia coli. *Science* **2003,** 301, (5633), 610-5.

28. Padan, E.; Kozachkov, L.; Herz, K.; Rimon, A., NhaA crystal structure: functional-structural insights. *J Exp Biol* **2009,** 212, (Pt 11), 1593-603.

29. Yamashita, A.; Singh, S. K.; Kawate, T.; Jin, Y.; Gouaux, E., Crystal structure of a bacterial homologue of Na+/Cl--dependent neurotransmitter transporters. *Nature* **2005,** 437, (7056), 215-23.

30. Papageorgiou, I.; De Koning, H. P.; Soteriadou, K.; Diallinas, G., Kinetic and mutational analysis of the Trypanosoma brucei NBT1 nucleobase transporter expressed in Saccharomyces cerevisiae reveals structural similarities between ENT and MFS transporters. *Int J Parasitol* **2008,** 38, (6), 641-53.

31. Valdes, R.; Arastu-Kapur, S.; Landfear, S. M.; Shinde, U., An ab Initio structural model of a nucleoside permease predicts functionally important residues. *J Biol Chem* **2009,** 284, (28), 19067-76.

32. Al-Salabi, M. I.; Wallace, L. J.; Luscher, A.; Maser, P.; Candlish, D.; Rodenko, B.; Gould, M. K.; Jabeen, I.; Ajith, S. N.; de Koning, H. P., Molecular interactions underlying the unusually high adenosine affinity of a novel Trypanosoma brucei nucleoside transporter. *Mol Pharmacol* **2007,** 71, (3), 921-9.
33. Al-Salabi, M. I.; Wallace, L. J.; De Koning, H. P., A Leishmania major nucleobase transporter responsible for allopurinol uptake is a functional homolog of the Trypanosoma brucei H2 transporter. *Mol Pharmacol* **2003,** 63, (4), 814-20.
34. Papageorgiou, I. G.; Yakob, L.; Al Salabi, M. I.; Diallinas, G.; Soteriadou, K. P.; De Koning, H. P., Identification of the first pyrimidine nucleobase transporter in Leishmania: similarities with the Trypanosoma brucei U1 transporter and antileishmanial activity of uracil analogues. *Parasitology* **2005,** 130, (Pt 3), 275-83.
35. De Koning, H. P.; Al-Salabi, M. I.; Cohen, A. M.; Coombs, G. H.; Wastling, J. M., Identification and characterisation of high affinity nucleoside and nucleobase transporters in Toxoplasma gondii. *Int J Parasitol* **2003,** 33, (8), 821-31.
36. Al-Salabi, M. I.; de Koning, H. P., Purine nucleobase transport in amastigotes of Leishmania mexicana: involvement in allopurinol uptake. *Antimicrob Agents Chemother* **2005,** 49, (9), 3682-9.
37. Zhang, J.; Visser, F.; Vickers, M. F.; Lang, T.; Robins, M. J.; Nielsen, L. P.; Nowak, I.; Baldwin, S. A.; Young, J. D.; Cass, C. E., Uridine binding motifs of human concentrative nucleoside transporters 1 and 3 produced in Saccharomyces cerevisiae. *Mol Pharmacol* **2003,** 64, (6), 1512-20.

38. Vickers, M. F.; Zhang, J.; Visser, F.; Tackaberry, T.; Robins, M. J.; Nielsen, L. P.; Nowak, I.; Baldwin, S. A.; Young, J. D.; Cass, C. E., Uridine recognition motifs of human equilibrative nucleoside transporters 1 and 2 produced in Saccharomyces cerevisiae. *Nucleosides Nucleotides Nucleic Acids* **2004,** 23, (1-2), 361-73.
39. Matovu, E.; Geiser, F.; Schneider, V.; Maser, P.; Enyaru, J. C.; Kaminsky, R.; Gallati, S.; Seebeck, T., Genetic variants of the TbAT1 adenosine transporter from African trypanosomes in relapse infections following melarsoprol therapy. *Mol Biochem Parasitol* **2001,** 117, (1), 73-81.
40. Scrutton, N. S.; Raine, A. R., Cation-pi bonding and amino-aromatic interactions in the biomolecular recognition of substituted ammonium ligands. *Biochem J* **1996,** 319 (Pt 1), 1-8.

Tables and Figures

Table 1. CoMFA and CoMSIA model statistics for the datasets A-D of Figure 2.

CoMFA

	A	B	C	D
Total Compounds	36	29	16	16
n	7	4	5	2
q^2_{cv}	0.65	0.55	0.57	0.79
SEE	0.29	0.32	0.17	0.12
r^2	0.93	0.89	0.98	0.99
F	55.9	50.5	74.4	311

CoMSIA

	A	B	C	D
Total Compounds	36	29	16	16
n	5	3	3	2
q^2_{cv}	0.50	0.55	0.65	0.61
SEE	0.32	0.34	0.26	0.11
r^2	0.91	0.86	0.93	0.99
F	62.5	51.5	47.4	358

Table 2. CoMFA and CoMSIA model statistics for the 112 compound database.

	CoMFA	CoMSIA
Total Compounds	112	112
n	11	6
q^2_{cv}	0.55	0.54
$q^2_{10\text{-Fold}}$	0.56	0.54
SEE	0.22	0.37
r^2	0.95	0.86
F	190	109

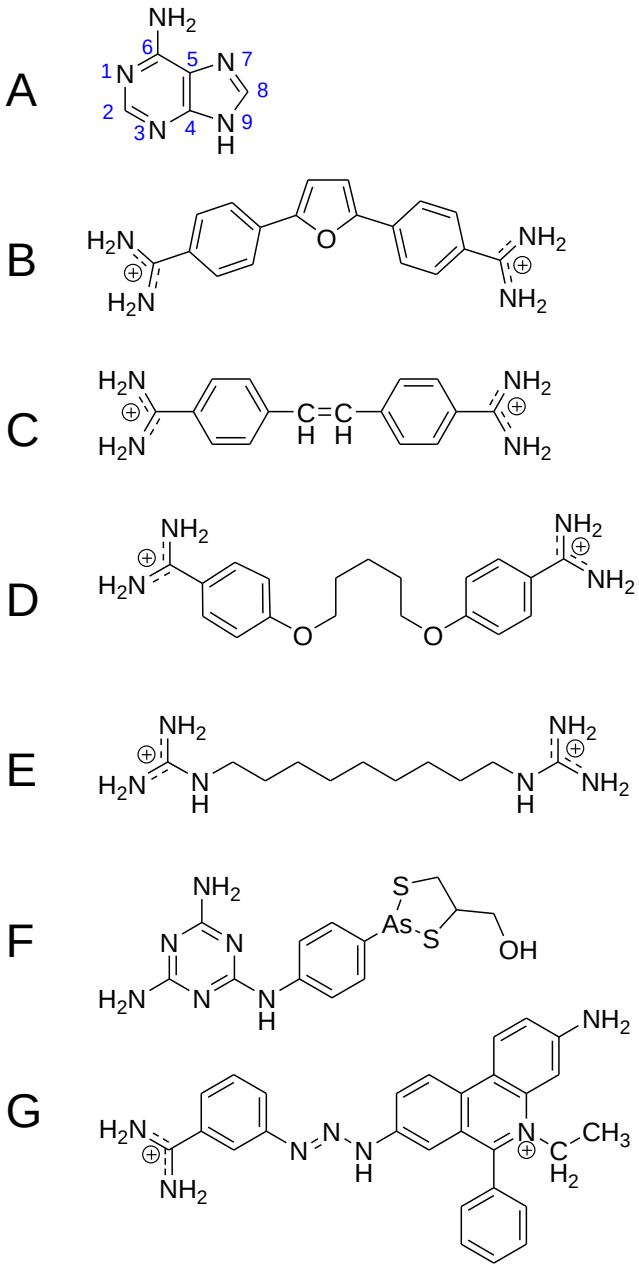


Figure 1. Scaffolds for initial alignment: A, adenine; B, furamidine; C, stilbamidine; D, pentamidine; E, 1,1’-(nonane-1,9-diyl)diguanidine; F, melarsoprol; G, isometamidium. All 112 compounds could be aligned to one of these scaffolds. Most compounds were in A-D.

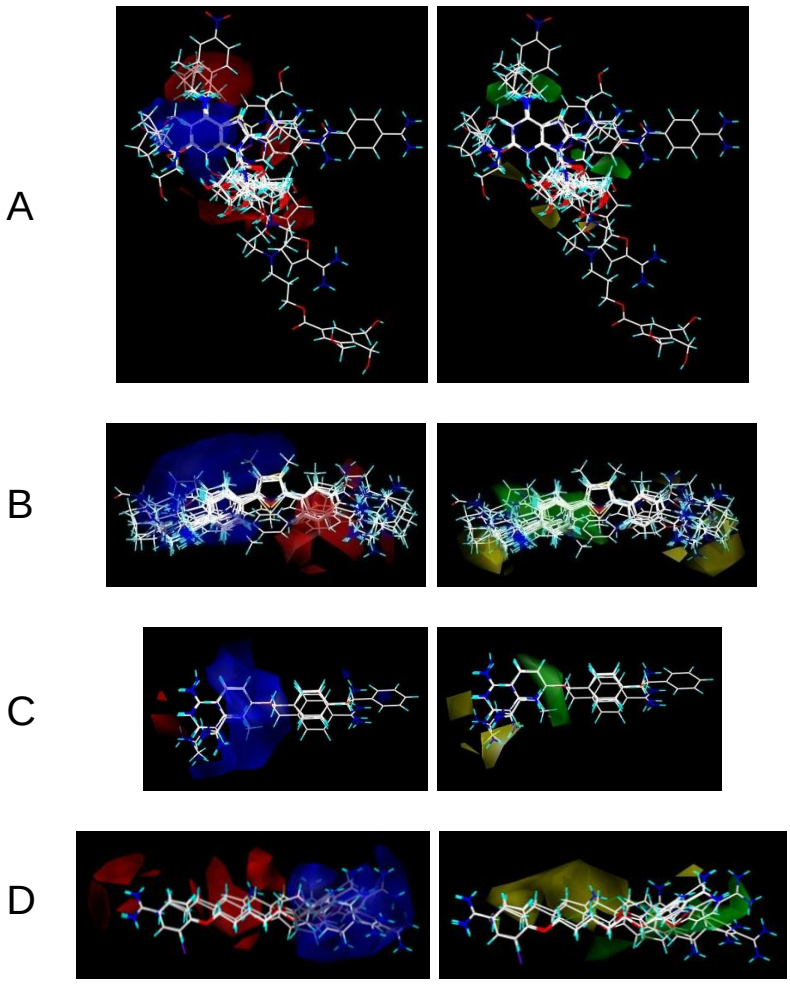


Figure 2. First alignment processes produced seven different databases for the 112 compounds. The compounds of the larger datasets, A-D, were employed for QSAR CoMFA and CoMSIA studies. Resulting three-dimensional CoMFA molecular surfaces are shown for datasets A-D, which are labeled A-D, respectively. Steric contributions are shown in green (favors bulky substituents) and yellow (bulky substituents impact negatively on binding), and the electrostatic contributions are displayed in blue (favoring a positive charge) and red (favoring a negative charge).

A

B

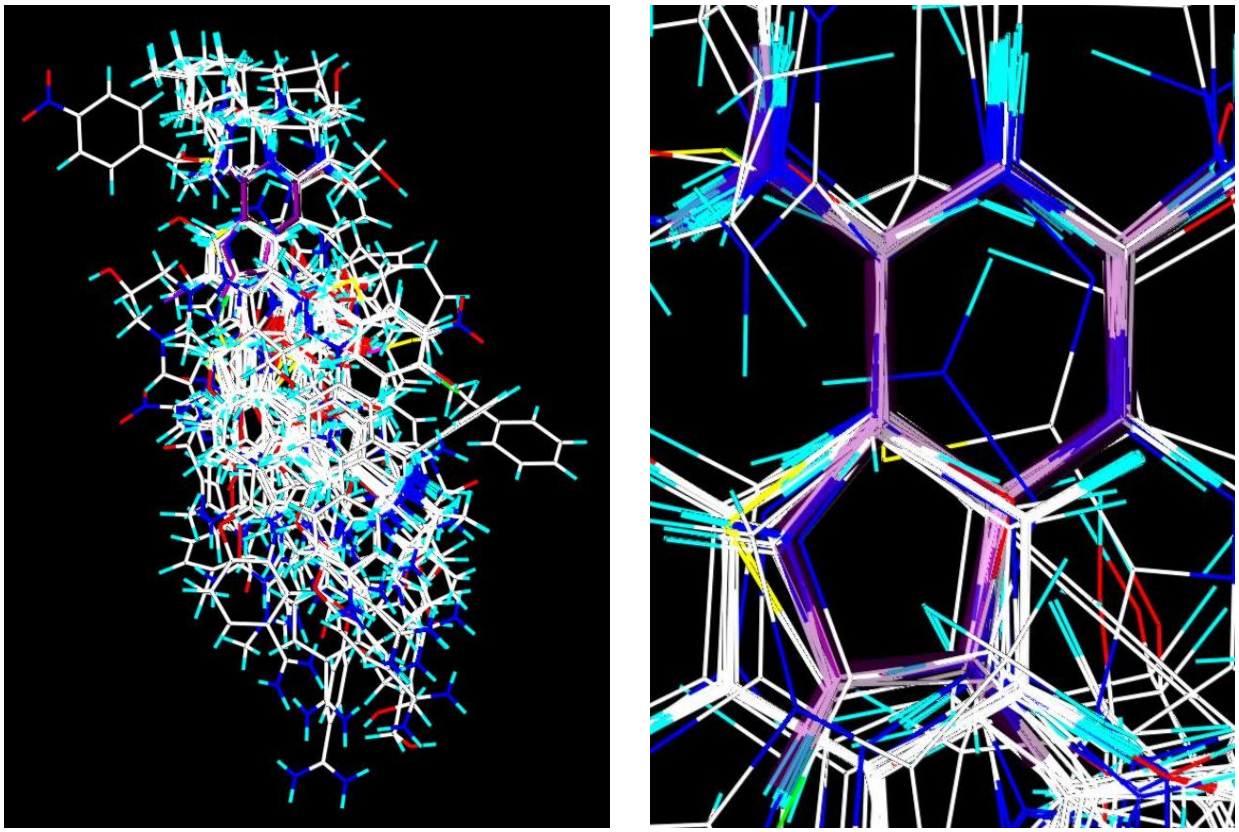


Figure 3. Final alignment of 112 compounds, with adenine displayed in purple.

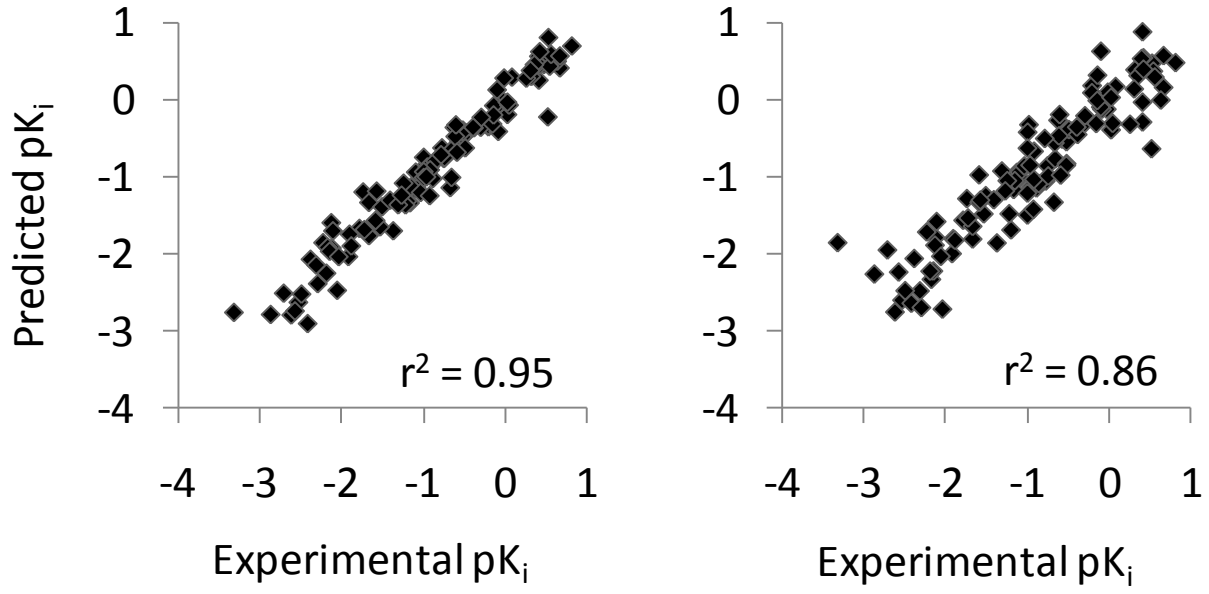


Figure 4. Actual *versus* predicted results from PLS models employing CoMFA (left) and CoMSIA (right) molecular descriptors.

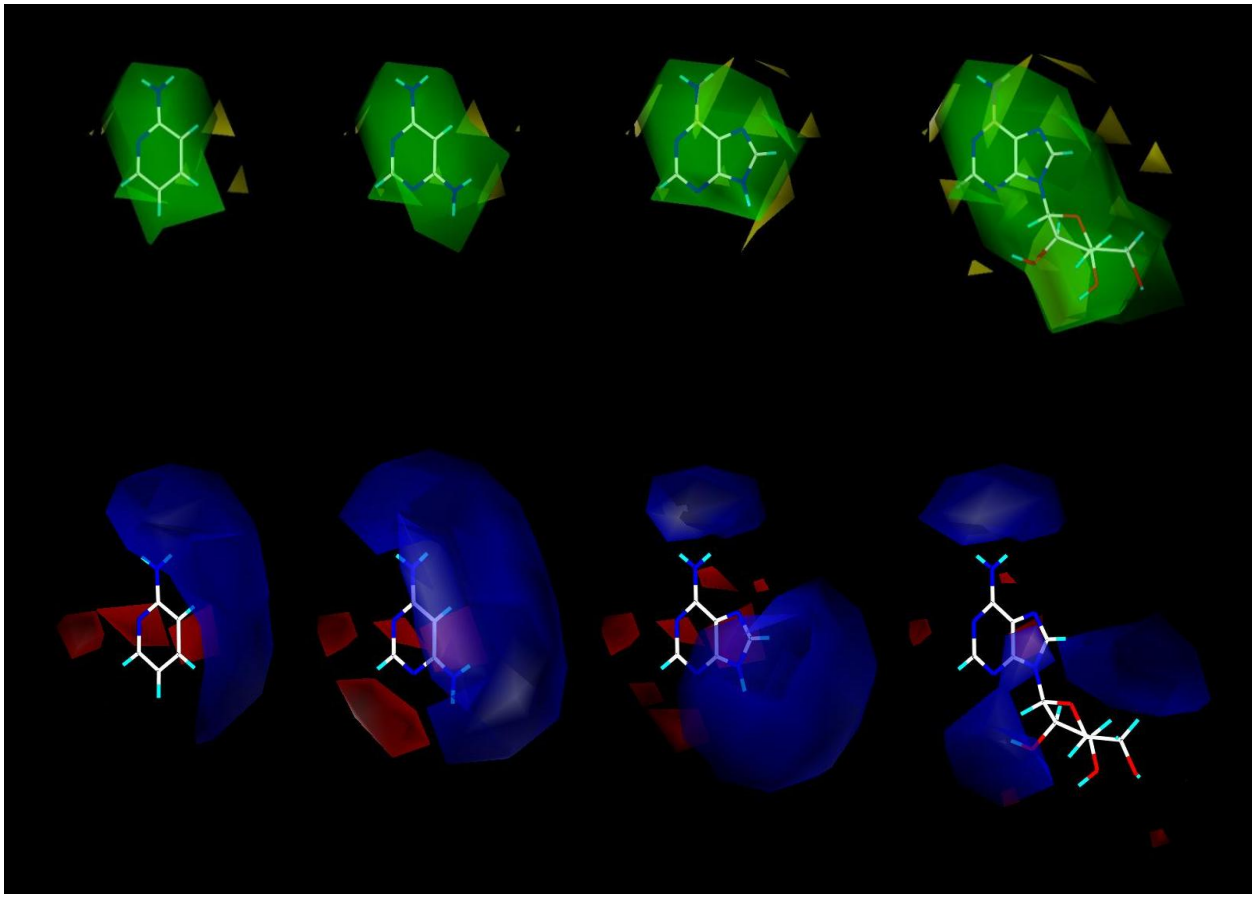


Figure 5. Calculated three-dimensional molecular surfaces for analyses of compound structural relationships with P2 transporter inhibition. From left to right, the compounds shown above are 2-aminopyridine, 4,6-diaminopyrimidine, adenine, and adenosine. Colors are as in Figure 2.

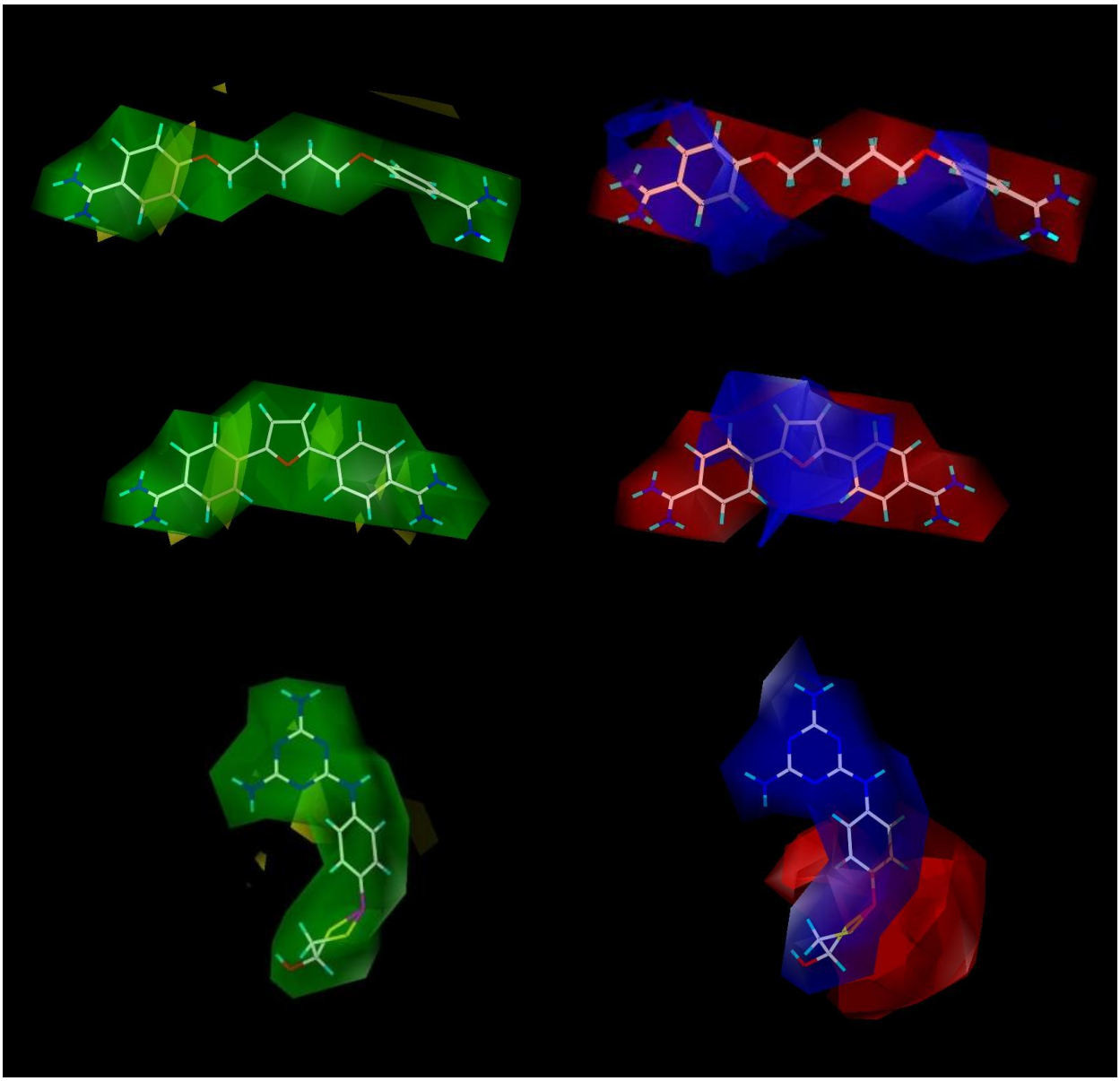


Figure 6. Three-dimensional molecular surfaces for pentamidine (top), furamidine (middle), and melarsoprol (bottom). Colors are as in Figure 2.

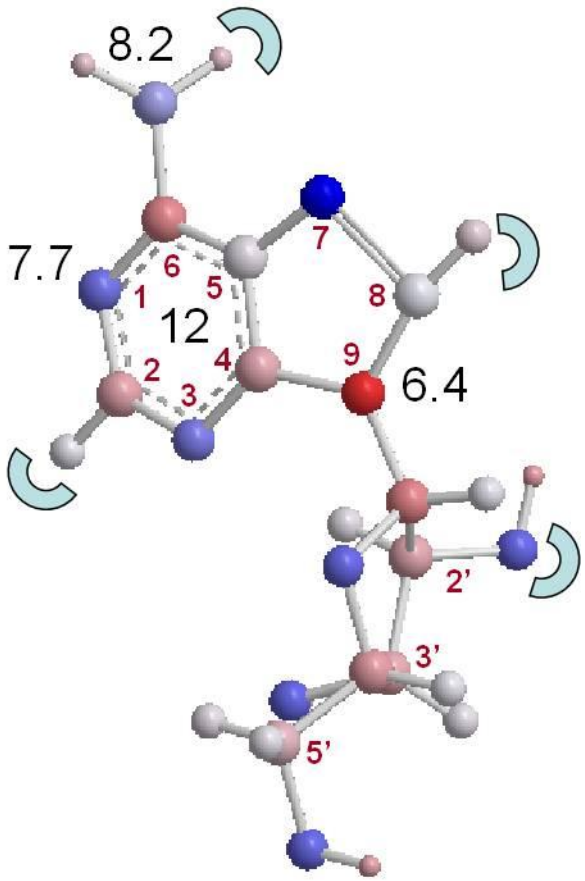


Figure 7. Model of adenosine, giving estimates of the contributions to the total binding energy of 34 kJ/mol in the black numbers, with the red numbers indicating the position on the purine or ribose rings. The half-circles indicate positions where substitutions reduced the adenosine binding affinity. The aromatic rings are estimated to contribute approximately 12 kJ/mol to the binding energy, although this could not be verified directly, as a nonaromatic adenosine analog would have a completely different three-dimensional structure. However, comparisons between aromatic diamidines and nonaromatic diamidines (Supplemental Table 2, Appendix A) are consistent with this estimate.

CHAPTER 3: GOVERNING INHIBITION OF ARYLIMIDAMIDES AGAINST LEISHMANIA: CONSERVATIVE COMPUTATIONAL MODELING TO IMPROVE CHEMOTHERAPIES

GOVERNING INHIBITION OF ARYLIMIDAMIDES AGAINST LEISHMANIA: CONSERVATIVE COMPUTATIONAL MODELING TO IMPROVE CHEMOTHERAPIES

Catharine J. Collar[1], Xiaohua Zhu[2], Karl Werbovetz[2], David W. Boykin[1], and W. David Wilson[1]

From Department of Chemistry, Georgia State University, Atlanta, Georgia 30303[1] and Division of Medicinal Chemistry and Pharmacognosy, The Ohio State University, Columbus, Ohio 43210[2].

Running head: Conservative modeling of inhibition against *Leishmania*

Address correspondence to W. David Wilson. Telephone: +1-404-413-5503. Fax: +1-404-413-5551. E-mail: wdw@gsu.edu.

The abbreviations used are: CoMFA, Comparative Molecular Field Analysis; CoMSIA, Comparative Molecular Similarity Indices Analysis; QSAR, Quantitative Structure-Activity Relationship; PLS, Partial Least Squares; SEE, Standard Error of Estimate; CPDD, Consortium for Parasitic Drug Development.

A dataset of 55 compounds with inhibitory activity against *L. donovani* axenic amastigotes and *L. amazonensis* intracellular parasites was examined through three-dimensional quantitative structure-activity relationship modeling employing molecular descriptors from both rigid and flexible compounds. For training and testing purposes, the compounds were divided into two datasets of 45 and 10 compounds, respectively. Statistically significant models were constructed and validated *via* the internal and external predictions. For all models employing steric, electrostatic, hydrophobic, H-donor and H-acceptor molecular descriptors, the R^2 values were greater than 0.90 and the SEE values were less than 0.22. The models obtained from rigid and flexible compounds were employed together to obtain a conservative method for predictions. This method minimized under predictions. Molecular descriptors from the models were then extrapolated, for the overall predictive devices and the individual compounds, and examined with regard to inhibitory activity. Information gained from the molecular descriptors is useful to the design of novel compounds. The models obtained can be employed to predict activities of the compounds designed and/or form predictions for compounds that exist and have not yet been examined with biological inhibitory assays.

Introduction

Leishmania species cause leishmaniasis, which is an endemic disease found in tropic and subtropic regions riddled with poverty and neglect.[1-3] This infection is most often in the form of cutaneous leishmaniasis, visible skin sores, or visceral leishaniasis, affected internal organs. Primarily transported through the bite of a female phlebotomine sandfly, millions of new cases are reported annually.[1, 4] When untreated, tens of thousands of these parasitic infections result in death.

Primary treatments for leishmaniasis include sodium stibogluconate and *N*-methylglucamine, while secondary therapies, which are often toxic, include pentamidine isethionate, amphotericin B and paromomycin sulfate.[1, 5] These classic treatments are costly and embedded with implications of high toxicity, resistance, pain, nausea, and diarrhea. Possible new therapies for the treatment of leishmaniasis have been examined and these include the implementation of liposomes, natural products, synthetic compounds and vaccines.[1, 6] Most methods employing liposomes are costly and hence not feasible, while other methods of therapy improvement have showed promice.[1, 7] Several natural products and synthetic compounds possess better therapeutic profiles with regard to activity and toxicity than the compounds currently used in treatment, while vaccine development has been too specific to *Leishmania* species and thus unsuccessful.[1]

Through several research endeavors activities and toxicities of series of compounds have been gathered and such data have been implemented in rational drug design.[8-11] These studies employ biological data of natural and/or synthetic compounds and computational tools to examine compounds with activity against *Leishmania* species. Examination of such compounds has led to the formation of predictive devices and from these devices the importance of some molecular structures has been ascertained. Although specific receptor interaction studies are important, especially when studying mechanisms, intact parasite studies of inhibition and toxicity are crucial for identifying compounds that will eradicate the parasite from hosts.[1] Such studies of synthetic chalcones and phospholipids display effective antileishmanial activity for compounds with: (1) a long alkyl chain, (2) bulky group's terminal the alkyl chain, and (3) an electron deficient group.[9, 12]

Our studies examine a biological dataset of synthetic arylimidamides which possess activities against *L. donovani* axenic amastigotes and *L. amazonensis* intracellular parasites. Inhibitory data, in the form of IC_{50} values, and Comparative Molecular Field Analysis (CoMFA) and Comparative Molecular Similarity Indices Analysis (CoMSIA) molecular descriptors were employed for partial least squares (PLS) regression. Predictive models and resulting molecular descriptor potentials contribute to the identification and understanding of important molecular features that govern the inhibitory actives of arylimidamides against species of *Leishmania*.

Experimental Procedures

Inhibitory Data. Briefly, IC_{50} (μM) values were gathered for compounds of interest using two assays. The first assay screened against axenic amastigote-like *L. donovani*, while the second screened against *L. amazonensis* intracellular parasites. Screening against *L. donovani* was conducted by: (1) culturing Ld1s parasites in potassium-based medium at pH 5.5, 37 °C, (2) incubating for three days with compounds in a 96-well plate, and (3) adding tetrazolium dye and quantifying the assay spectrophotometrically. While screening against *L. amazonensis* intracellular parasites was conducted by: (1) plating macrophages and allowing adhering overnight, (2) adding *L. amazonensis* promastigotes transfected with β-Lactamase gene (MOI: 5:1) and incubating overnight, (3) adding compounds of interest and incubating for 72 hours at the temperature of interest, (4) adding nitrocefin in lysis buffer and incubating an additional 3 to 5 hours, and (5) reading the plate at 490 nm.[13] Experimental IC_{50} values for *L. donovani* axenic amastigotes and *L. amazonensis* intracellular parasites were obtained for 55 compounds.

Preparation of Compounds for Computational Studies. SYBYL 8.1[14] software was employed to construct all compounds in three-dimensional space. Compounds were then divided into training and testing datasets. These datasets consisted of 45 and 10 compounds, respectively. The compounds of the training dataset then underwent a short molecular dynamics simulation of 1 ns. This system employed SYBYL 8.1 default settings at a constant temperature and volume (NTV). Briefly, (1) the system temperature was 300K with a coupling constant of 100 fs, (2)

Maxwell-Boltzmann distribution was employed for initial atom velocities, (3) the non-bonded pair list was updated every 25 fs, (4) and the duration of the molecular dynamics simulations *in vacuo* was 1ns with a time step of 100 fs and a snapshot every 1000 fs. This displayed several low energy structures. Torsional angles of all training dataset compounds were modified to explore the low energy conformations and modified compounds were minimized to convergence using the Tripos force field, conjugate gradient algorithm, and Gaseiger-Huckel charges. The termination gradient was 0.01 kcal/(mol Å) and the maximum iterations were 10^4.

Rigid Alignment of Compounds and Resulting Models. Each training dataset of compounds with modified torsional angles was aligned using the "Align Database" option of the QSAR module in SYBYL. Aligned structures were then analyzed through the use of molecular descriptors. CoMFA (steric and electrostatic) and CoMSIA (steric, electrostatic, hydrophobic, H-donors and H-acceptors) molecular descriptors were calculated and PLS regression was employed to compare the molecular descriptors of compounds to obtained average IC_{50} values. The number of components was determined by the smallest predicted error sum of squares. Optimum models employing CoMFA molecular descriptors consisted of three components, whereas the ones with CoMSIA molecular descriptors employed six.

Flexible Alignment of Compounds and Resulting Models. Five compounds with low IC_{50} values for the *L. amazonensis* intracellular parasite assay were employed for flexible compound alignment using the "Align Pharmacophore" option of the GALAHAD module. Parameters were acquired through the "Suggest from Data" option and the best 20 models were gained. The

highest scoring model with respect to maximized pharmacophore consensus, maximized steric consensus, and minimized energy was employed as a template for individual compound alignment of the entire training dataset. The "Align Molecules to Template Individually" option was selected and parameters were acquired once more through the "Suggest from Data" option; the "Keep Best N Models" option was reset to 20. Molecular descriptors were calculated for the highest scoring model and PLS regression was implemented in the same manner as for the rigid compounds. The optimum numbers of components were determined as previously described; models with CoMFA molecular descriptors consisted of three components, whereas models with CoMSIA molecular descriptors employed six.

Statistical Analyses. The statistics calculated from PLS regression included: a cross-validated correlation coefficient (Q^2), the coefficient of determination (R^2), the standard error of estimate (SEE), the F statistic, a bootstrap R^2 (R^2_{bs}), and a bootstrap SEE (SEE_{bs}). The bootstrap analysis was used to check the stability of the models through cross-validation into two, five, and ten groups. The average values of the bootstrap analysis are displayed with the rest of the statistics.

Testing Datasets and the Conservative Model Method. The models constructed from the rigid and flexible alignments were employed to examine testing datasets that were aligned *via* rigid and flexible methods. Of the pIC_{50} values predicted, for both training and testing datasets, the more negative pIC_{50} prediction was considered the most viable. This method favors over prediction rather than under prediction.

Molecular Descriptor Potentials. Molecular descriptor potentials acquired through the mapping of the product standard deviation with respect to molecular descriptor values and coefficients at each lattice point were extrapolated from the models. Default levels of contour by contribution were employed to gather favored and disfavored potentials for overall models. The individual compounds of the models were analyzed *via* the contour by actual analysis method. Software output was used to determine the proper ranges of assigned favored and disfavored contour regions for individual compounds.

Results

The entirety of the dataset, 45 training and 10 testing compounds, can be represented *via* the scaffold structure displayed in Figure 1. At each of the five positions labeled in this figure there are differing atoms or groups: positions one and four display single atom changes in the form of carbon, oxygen, sulfur, and nitrogen, whereas positions two, three and five display larger group substituent modifications.

Biological IC_{50} values were acquired for each compound of the training and testing datasets through two assays targeting *L. donovani* axenic amastigotes and *L. amazonensis* intracellular parasites. These inhibitory values were averaged and standard deviations were acquired (Supplemental Table 3, Appendix B and Table 3). For modeling purposes, the IC_{50} values were log transformed into pIC_{50} values ($pIC_{50} = -\log(IC_{50})$). Figure 2 displays the pIC_{50} data; experimental values against the *L. donovani* axenic amastigotes are shown in green,

whereas the values against the *L. amazonensis* intracellular parasites are displayed in blue. The standard deviations of the data are represented in general by trend lines and the averaged values are shown as triangles and squares, respectively. Notice that the slopes are very similar with values between 0.96 and 1.0, and R^2 values are 0.95 or higher. This displays the relative range of inhibitory activity and respective deviations from the average inhibitory values associated with each synthetic compound in the training and testing datasets. The pIC_{50} distribution of data are also shown in this figure; the inhibitory activity of arylimidamides against *L. donovani* axenic amastigotes ranges between approximately -2.5 and 0.5, whereas those active against *L. amazonensis* intracellular parasites range between about -1.5 and 1.5.

Compounds examined through biological assays were aligned *in silico* in three-dimensional conformations using two methods: (1) rigid alignments of compounds were obtained through the implementation of the SYBYL "Align Database" option of the QSAR module, and (2) flexible alignments of compounds were acquired through the use of the "Pharmacophore Alignment" option of the GALAHAD module. Rigid alignments were preformed on low energy conformations of compounds. Molecular descriptors were then calculated and PLS regression was employed to construct predictive models implementing the descriptors and respective biological inhibitory data. The best computational models formed consisted of compounds in their most linear conformation with an overall plus one charge.

Flexible alignment of compounds was also implemented. This process employed the five most active compounds against the *L. amazonensis* intracellular parasites from the training

dataset (Figure 3). Figure 4 displays the outcome of pharmacophore simulations that lead to PLS regression models employing flexible compounds. The rotation of the compounds allows for visualization of alignment and positioning of identified feature potentials. The observed features governing structure alignment are: (1) four aromatic rings (cyan); (2) N=C−N groups, two positive nitrogen (red) and a H-donor (magenta); (3) atoms at the one and a two position of Figure 1, two H-acceptor (green); (4) atoms at a five position, a H-donor or H-acceptor (overlaid magenta and green equates to dark green). Then, all of the training and testing compounds were flexibly aligned to the pharmacophore. These alignments can be viewed in relation to rigid alignments (Figure 5). Rigid compounds were aligned by N=C−N groups. Notice that there is a difference in the spatial relationships of the compounds.

Inhibitory and compound structural data were employed to construct predictive models through PLS regression methods. The statistics for these models indicate that models employing CoMSIA molecular descriptors should outperform those constructed with CoMFA molecular descriptors (Table 1). This is shown in higher Q^2, R^2, and F statistics and lower SEE statistics for the models constructed with rigid compounds. Similarly, the models of flexible compounds displayed higher R^2 and F statistics and lower SEE statistics. The low Q^2 values for models of flexible compounds were attributed to: (1) torsional variability, (2) differences in optimal low energy structural conformations, and (3) contributions of compound inhibitory activities. Based on statistics, models with CoMSIA molecular descriptors were examined further with regard to molecular descriptors (Table 2). The factors governing these models were dominated by

hydrophobic potentials followed by H-donor potentials. Smaller contributions were made by H-acceptor, steric, and electrostatic potentials, respectively.

The internal (training dataset) and external (testing dataset) predictions of these models are displayed in Figure 6 through the plotting of predicted pIC_{50} values in relation to experimental pIC_{50} values. The training dataset of this figure is colored in accordance to Figure 2, whereas all testing dataset predictions are in red. Although internal predictions were linear, some testing dataset compounds were more difficult to predict for than others. The variance in compound prediction differed between the models for compounds of rigid and flexible alignments; hence, by taking the most negative prediction of each compound regardless of rigid or flexible alignment and plotting these values against respective experimental data a conservative method for prediction can be obtained. The combination of the models reduces under prediction. Table 3 displays the testing dataset along with experimental average IC_{50} values, plus or minus respective standard deviations, along with the conservative model IC_{50} prediction; note that model error for each compound is not shown.

From the models employing CoMSIA molecular descriptors, potentials were extrapolated and viewed in relation to the overall models (Figure 7) and individual compounds thereof (Figure 8). Figure 7 displays the overall CoMSIA molecular descriptors for the rigid and flexible models. It is evident that each overall model displays different molecular descriptor potential contributions, for all molecular descriptors (steric, electrostatic, hydrophobic, H-donor and H-acceptor). This indicates that each model is constructed somewhat differently; although, there are

similarities between the potentials obtained. Using the positions of Figure 1 as a reference: (1) steric bulk is favored (green) at positions three and five and perhaps not symmetrically, whereas disfavored steric bulk (yellow) regions are just outside those favored, (2) positive electrostatic charge is favored (blue) at one, if not both, of the N=C−N groups near position five, whereas negative charge is favored (red) predominantly at or near position one and outside one of the N=C−N groups, (3) hydrophobic interactions are favored (yellow) at positions two and five, whereas disfavored hydrophobic interactions (gray) are near three positions and outside five positions, (4) H-donor atoms are favored (cyan) predominantly at or below the five position and disfavored (purple) in regions beyond favored regions and on three positions, and (5) H-acceptors are favored (magenta) near the terminal N=C−N groups, and disfavored (red) below the four position(s) and outside favored N=C−N groups of the comparison molecule DB766.

With regard to the scaffold structure of Figure 1, Figure 8 displays the molecular descriptor potentials of individual compounds extrapolated from respective models employing CoMSIA molecular descriptors. The molecular descriptor potential regions of individual molecules appear to be more consistent within their respective rigid and flexible models than they were in the overall models of Figure 7. However, the molecular potentials that resulted were also fewer. These included favored and disfavored hydrophobic, favored H-donor and favored H-acceptor potentials. With X of Figure 8 representing positions one through five of the Figure 1 scaffold structure and biological inhibitory data in Supplemental Table 3 (Appendix B), it is possible to examine not only the molecular descriptor potentials with regard to model

contribution but also the contribution of substructures to biological inhibitory activity. To most effectively describe these findings, it is important that comparisons are made to a compound that is active in both datasets. DB766 was selected for analyses.

With respect to Figure 1, the molecular descriptor potentials for DB766 include: (1) favored hydrophobic potentials near the aromatic rings consisting of the four position, opposite position three, and on the flanking five position aromatic rings, (2) favored H-donor potentials are displayed below the left five position N=C−N group, and (3) favored H-acceptor potentials are on the N=C−N group opposite the side of the favored H-donor and extended to the outer aromatic ring. The IC_{50} values for this compound against *L. amazonensis* intracellular parasites and *L. donovani* axenic amastigotes are 0.09 and 0.50 μM, respectively. The general structure of DB1867, compared to DB766, differs only by a sulfur atom at position one and with this change the compound becomes more linear and favored hydrophobic interactions are spread to positions one and two. Potentials for favored H-acceptors are near N=C−N groups and the IC_{50} values are 0.05 and 0.68 μM, respectively. DB946 is the only compound in the training dataset to differ from DB766 at position two; this compound also differs at position three. The methyl groups at position two fill similar special areas as substituents in position three. Favored hydrophobic potentials reside in position two, three, and five locations. This compound's IC_{50} values are 0.11 and 0.37 μM, respectively. DB667 and DB1876 differ from DB766 at position three. DB667 consists of hydrogen atoms at position three and molecular descriptor potentials similar to those of DB946 (hydrophobic) and DB766 (H-donor and H-acceptor); although, the favored

hydrophobic potentials span a greater length for DB667. The IC_{50} values for this compound are 0.53 and 1.6 μM, respectively. DB1876 displays large disfavored hydrophobic molecular descriptor potentials at the three positions. The remaining potentials are favored hydrophobic potentials near the aromatic rings and favored H-donor and H-acceptor potentials near the N=C−N group(s). This compound has IC_{50} values of 2.1 and 28 μM, respectively. DB1851 differed from DB766 at position four. This resulted in favored hydrophobic interactions that span more of the molecule than previous compounds discussed and H-donor and H-acceptor potentials similar to those of DB1876. The IC_{50} values for these compounds are poor, greater than 10 and 50 μM, respectively. DB1921, DB1942, and DB1906 all differ from DB766 at the five positions. DB1921 is flanked at the five positions and has different substituents at the three positions. This compound consists of potentials similar to DB1876; however, it is also missing most of the favored hydrophobic and H-acceptor potentials. The IC_{50} values for this compound are 4.7 and 41 μM, respectively. DB1942 consists of a longer more flexible ring structure than DB766 and consists of disfavored hydrophobic molecular descriptor potentials primarily at the five positions. Positive hydrophobic potentials are on the inner aromatic rings or the outer rings near the five position, whereas H-donors are favored on one side of a N=C−N group and H-acceptors are favored at both N=C−N group(s). This compound has IC_{50} values of 0.81 and 3.6 μM, respectively. The five positions of DB1906 consist of more rings than DB766. The molecular descriptor potentials for this compound were similar to those of DB766 (hydrophobic and H-donor) and DB946 (H-acceptor), IC_{50} values are 0.27 and 1.9 μM, respectively.

Discussion

Studies examined chalcones and phospholipids and found that high inhibitory activity occurred when compounds possessed a long alkyl chain, bulky groups terminal the alkyl chain, and an electron deficient group.[9, 12] The structures of chalcones and phospholipids are quite different from each other, and these compounds differ substantially from the arylimidamides examined in this study (Figure 1 and Supplemental Table 3, Appendix B).

The numbered locations of Figure 1 aid in the explanation of inhibitory data displayed in Figure 2 through the interpretation of pharmacophore consensus potentials (Figure 4) and molecular descriptor contribution potentials (Figures 7 and 8). The pharmacophore alignment of Figure 4 is calculated using the compounds of Figure 3. By only employing the most active compounds, the pharmacophore is strictly for compounds of similar structure and inhibitory activity. The pharmacophore results suggest importance of aromatic and positively charged N=C−N groups. These can be related to the bulky groups flanking the alkyl chain and the electron deficient group, respectively. The long alkyl chain may be related to the long carbon backbone that is in part aromatic groups and/or the carbons of the furan that link the rings.

PLS regression of calculated molecular descriptor potentials and respective biological inhibitory values, for both the rigid and flexible alignments of compounds presented in Figure 5, produced statistically significant models; yet, those employing the CoMSIA molecular descriptors from rigid structure alignment were the only ones with Q^2 values greater than 0.5

(Table 1). It has been stated for years that if Q^2 values are greater than 0.5 then the model has predictability better than chance.[14] What we realize from our models, especially those aligned by flexible conformations, is that each compound contributes to the entirety of the model and that the models constructed from molecular descriptors of flexible compounds may be predicting just as well, if not better, than those constructed from molecular descriptors of rigid compounds (Figure 6). The rigid models of Figure 6 produce a greater amount of under prediction than the flexible models. For example, for the rigid model, one of the compounds active against *L. donovani* has an experimentally determined pIC_{50} value of -1.7 and a predicted value of 0, these IC_{50} values are 50 and 1, respectively, whereas for the flexible model the same compound has a predicted pIC_{50} value of -1.7, the same as the experimental value.

Under prediction is a problem that needs to be addressed since predictive models such as the ones constructed in this study can be employed to scan potential candidates for synthetic drug design. Often synthetic measures are costly and timely; hence, it is better to synthesize only compounds expected to have the best inhibitory activity and disregard those expected to have the worse. To minimize under prediction, the minimum pIC_{50} predictions from the rigid or flexible models were plotted against the average experimental values. These data are shown as the conservative predictions of Figure 6. In this column we see that under predictions are no longer occurring for these models; yet, there are still over predictions. Over predictions, as long as they are few, are not as problematic since these values are larger and synthetic measures will most likely result in biologically obtained inhibitory activity better than calculated.

From models, such as the ones constructed during this study, molecular descriptor contributions can be obtained and observed. A previous study that employed CoMSIA molecular descriptors found that steric and hydrophobic interactions governed the model.[9] This study was for synthetic phospholipids. Similarly, our model was governed by hydrophobic interactions; yet, this contribution of molecular descriptor interactions was followed by H-donor, H-acceptor, steric and then electrostatic (Table 2). When the overall CoMSIA potentials for each molecular descriptor are visualized, compound structures appear applicable for modifications. Figure 7 allows for comparison between the models and overall analyses. It is important to realize that molecular descriptor potentials are unique to each model; hence, no two models are the same. As was found important previously, positive blue electrostatic potentials display the importance of the N=C−N groups, whereas the steric and hydrophobic potentials show the importance of the rings and substituents. To fully understand molecular descriptor contribution in relation to biological inhibitory data, it is important to analyze the potentials of individual compounds; a select set of which are shown in Figure 8. These potentials display much more consistency than those for the overall models of Figure 7.

New compounds can be designed by employing the data acquired from the pharmacophore (Figure 4) and extrapolated molecular descriptor potentials (Figures 7 and 8). To do this, the basic pharmacophore must remain intact and the potentials of the overall models and those of individual compounds should be used for guidance. Since the importance of hydrophobic, H-donor and H-acceptor atoms are clearly displayed as essential potentials for the

individual compounds in Figure 8; this is a good place to begin. The favored hydrophobic potentials of the four aromatic rings exhibit significance (Figures 7 and 8); these are also seen as important in the pharmacophore (Figure 4). Hence, it appears imperative that the four rings remain a constant in our initial modeling efforts. H-donors appear to be important to regions near the N=C−N groups (Figures 7 and 8). One of the N=C−N groups is shown as essential in the pharmacophore (Figure 4). Likewise, H-acceptors appear to be significant to the region including and between the N=C−N groups and the N of the outer most aromatic rings (Figures 7 and 8). One such region was identified in the pharmacophore (Figure 4). Based on these observations, new compounds have been designed and predictions have been obtained (Figure 9). The ranges include the smallest and largest prediction obtained *via* the models constructed of rigid and flexible compound structures.

In summary, by employing such findings it is possible to scan for potentially active compounds both efficiently and conservatively through the use of predictive models. The governing of inhibition results as models are employed and new compounds are designed, activities are predicted, compounds are synthesized, and biological assays provide experimental data for analyses.

Acknowledgements

This work was supported by the Bill and Melinda Gates Foundation through the Consortium for Parasitic Drug Development (CPDD) (to WDW, DWB, and KW) and by the

Georgia State University Molecular Basis of Disease Fellowship and the David W. Boykin Graduate Fellowship in Medicinal Chemistry (to CJC).

References

1. Santos, D. O.; Coutinho, C. E.; Madeira, M. F.; Bottino, C. G.; Vieira, R. T.; Nascimento, S. B.; Bernardino, A.; Bourguignon, S. C.; Corte-Real, S.; Pinho, R. T.; Rodrigues, C. R.; Castro, H. C., Leishmaniasis treatment--a challenge that remains: a review. *Parasitol Res* **2008,** 103, (1), 1-10.
2. Herwaldt, B. L., Leishmaniasis. *Lancet* **1999,** 354, (9185), 1191-9.
3. Alvar, J.; Yactayo, S.; Bern, C., Leishmaniasis and poverty. *Trends Parasitol* **2006,** 22, (12), 552-7.
4. Killick-Kendrick, R., The biology and control of phlebotomine sand flies. *Clin Dermatol* **1999,** 17, (3), 279-89.
5. Murray, H. W.; Berman, J. D.; Davies, C. R.; Saravia, N. G., Advances in leishmaniasis. *Lancet* **2005,** 366, (9496), 1561-77.
6. Singh, S.; Sivakumar, R., Challenges and new discoveries in the treatment of leishmaniasis. *J Infect Chemother* **2004,** 10, (6), 307-15.
7. Golenser, J.; Frankenburg, S.; Ehrenfreund, T.; Domb, A. J., Efficacious treatment of experimental leishmaniasis with amphotericin B-arabinogalactan water-soluble derivatives. *Antimicrob Agents Chemother* **1999,** 43, (9), 2209-14.

8. Avery, M. A.; Muraleedharan, K. M.; Desai, P. V.; Bandyopadhyaya, A. K.; Furtado, M. M.; Tekwani, B. L., Structure-activity relationships of the antimalarial agent artemisinin. 8. design, synthesis, and CoMFA studies toward the development of artemisinin-based drugs against leishmaniasis and malaria. *J Med Chem* **2003,** 46, (20), 4244-58.
9. Kapou, A.; Benetis, N. P.; Avlonitis, N.; Calogeropoulou, T.; Koufaki, M.; Scoulica, E.; Nikolaropoulos, S. S.; Mavromoustakos, T., 3D-Quantitative structure-activity relationships of synthetic antileishmanial ring-substituted ether phospholipids. *Bioorg Med Chem* **2007,** 15, (3), 1252-65.
10. Ryu, C. K.; Lee, Y.; Park, S. G.; You, H. J.; Lee, R. Y.; Lee, S. Y.; Choi, S., 3D-QSAR studies of heterocyclic quinones with inhibitory activity on vascular smooth muscle cell proliferation using pharmacophore-based alignment. *Bioorg Med Chem* **2008,** 16, (22), 9772-9.
11. Sanders, J. M.; Gomez, A. O.; Mao, J.; Meints, G. A.; Van Brussel, E. M.; Burzynska, A.; Kafarski, P.; Gonzalez-Pacanowska, D.; Oldfield, E., 3-D QSAR investigations of the inhibition of Leishmania major farnesyl pyrophosphate synthase by bisphosphonates. *J Med Chem* **2003,** 46, (24), 5171-83.
12. Liu, M.; Wilairat, P.; Croft, S. L.; Tan, A. L.; Go, M. L., Structure-activity relationships of antileishmanial and antimalarial chalcones. *Bioorg Med Chem* **2003,** 11, (13), 2729-38.

13. Buckner, F. S.; Wilson, A. J., Colorimetric assay for screening compounds against Leishmania amastigotes grown in macrophages. *Am J Trop Med Hyg* **2005,** 72, (5), 600-5.
14. *SYBYL Molecular Modeling Software, 8.1 ed.*, Tripos Inc.: St. Louis, MO, 2008.

Tables and Figures

Table 1. Statistics of partial least squares predictive models for a biological dataset of synthetic arylimidamides with activities against *L. donovani* axenic amastigotes (LD) and *L. amazonensis* intracellular parasites (LA).

	Rigid Alignment				Flexible Alignment			
	CoMFA		CoMSIA		CoMFA		CoMSIA	
	LA	LD	LA	LD	LA	LD	LA	LD
Q^2	0.23	0.25	0.47	0.59	0.16	0.60	0.07	0.22
SEE	0.45	0.41	0.25	0.18	0.24	0.33	0.16	0.14
R^2	0.68	0.77	0.91	0.96	0.91	0.85	0.96	0.97
F	29.1	44.5	61.4	137	131	78.0	159	229
SEE_{bs}	0.35	0.37	0.21	0.16	0.21	0.26	0.12	0.12
R^2_{bs}	0.80	0.82	0.94	0.97	0.93	0.90	0.98	0.98

Table 2. Contribution of CoMSIA molecular descriptors for rigid and flexible models employing structures of training dataset compounds and respective biological activities.

	Rigid Alignment		Flexible Alignment	
	L. amazonensis	*L. donovani*	*L. amazonensis*	*L. donovani*
Steric	0.15	0.14	0.13	0.13
Electrostatic	0.11	0.08	0.14	0.15
Hydrophobic	0.47	0.43	0.33	0.34
H-Donor	0.15	0.21	0.20	0.21
H-Acceptor	0.12	0.14	0.20	0.17

Table 3. Predictions in terms of IC_{50}. Experimental values for compound inhibitory activity against *L. donovani* axenic amastigotes (LD) and *L. amazonensis* intracellular parasites (LA) are displayed in columns LD Calc and LA Calc. The predicted values are those from the conservative predictions of Figure 6.

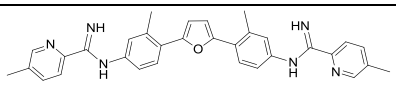
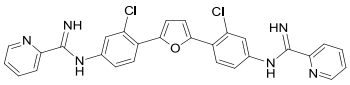
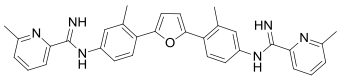
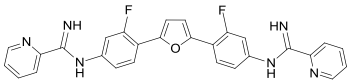
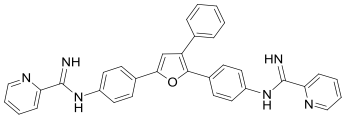
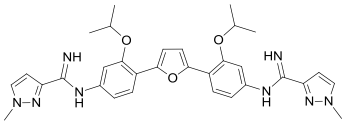
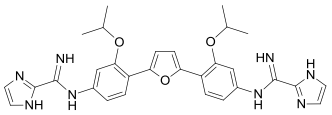
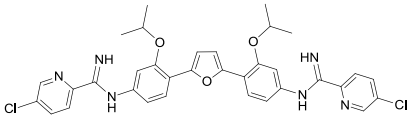
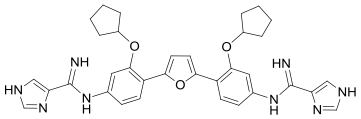
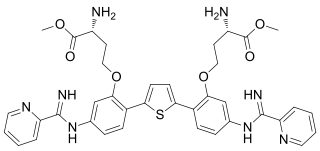
Name	Structure	LA Exp	LA Calc	LD Exp	LD Calc
DB710		0.16 ± 0.04	1.4	0.84 ± 0.2	4.0
DB712		0.56 ± 0.08	0.65	2.0 ± 0.6	4.6
DB749		3.1 ± 0.7	1.4	>50	50
DB874		1.4 ± 0.3	0.66	4.2 ± 1.3	3.2
DB889		0.11 ± 0.02	2.3	1.7 ± 0.5	10
DB1856		0.74 ± 0.3	3.2	5.8 ± 0.7	10
DB1857		0.37 ± 0.2	0.69	1.0 ± 0.2	10
DB1864		>10	9.3	5.6 ± 1.8	16

Table 3 (continued)

Name	Structure	LA Exp	LA Calc	LD Exp	LD Calc
DB1908		>10	4.7	14 ± 1	16
DB1930		5.5 ± 0.2	9.3	>100	63

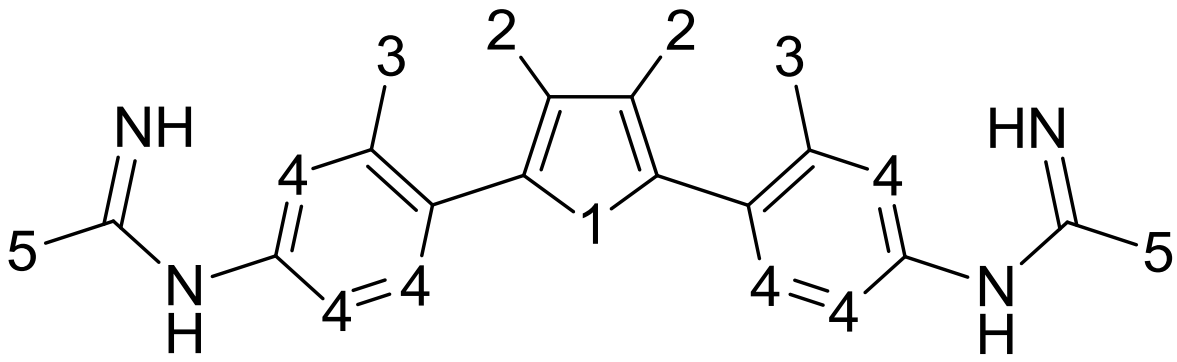


Figure 1. Scaffold structure for compounds being employed to examine biological inhibitory data through quantitative structure-activity relationships of *L. donovani* axenic amastigotes and *L. amazonensis* intracellular parasites. All training dataset structures and respective inhibitory data can be viewed in Supplemental Table 3 (Appendix B).

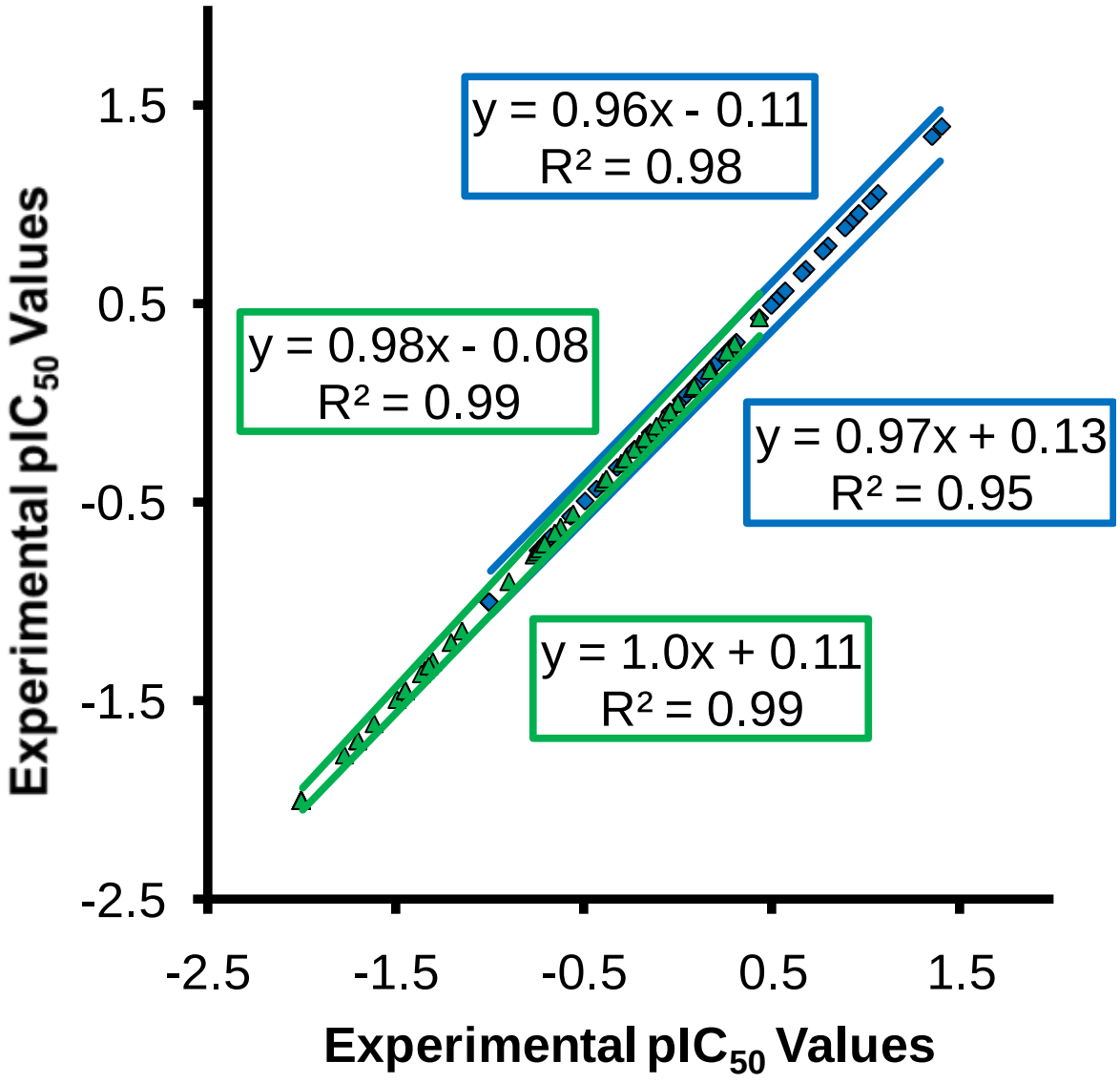


Figure 2. Biological pIC_{50} data of synthetic arylimidamides active against *L. donovani* axenic amastigotes (green) and *L. amazonensis* intracellular parasites (blue). The negative log values of average experimentally obtained IC_{50} data, displayed in Supplemental Table 3 (Appendix B) and Table 3, and these values plus and minus respective standard deviations are all plotted against the negative log value of average experimentally obtained IC_{50} data.

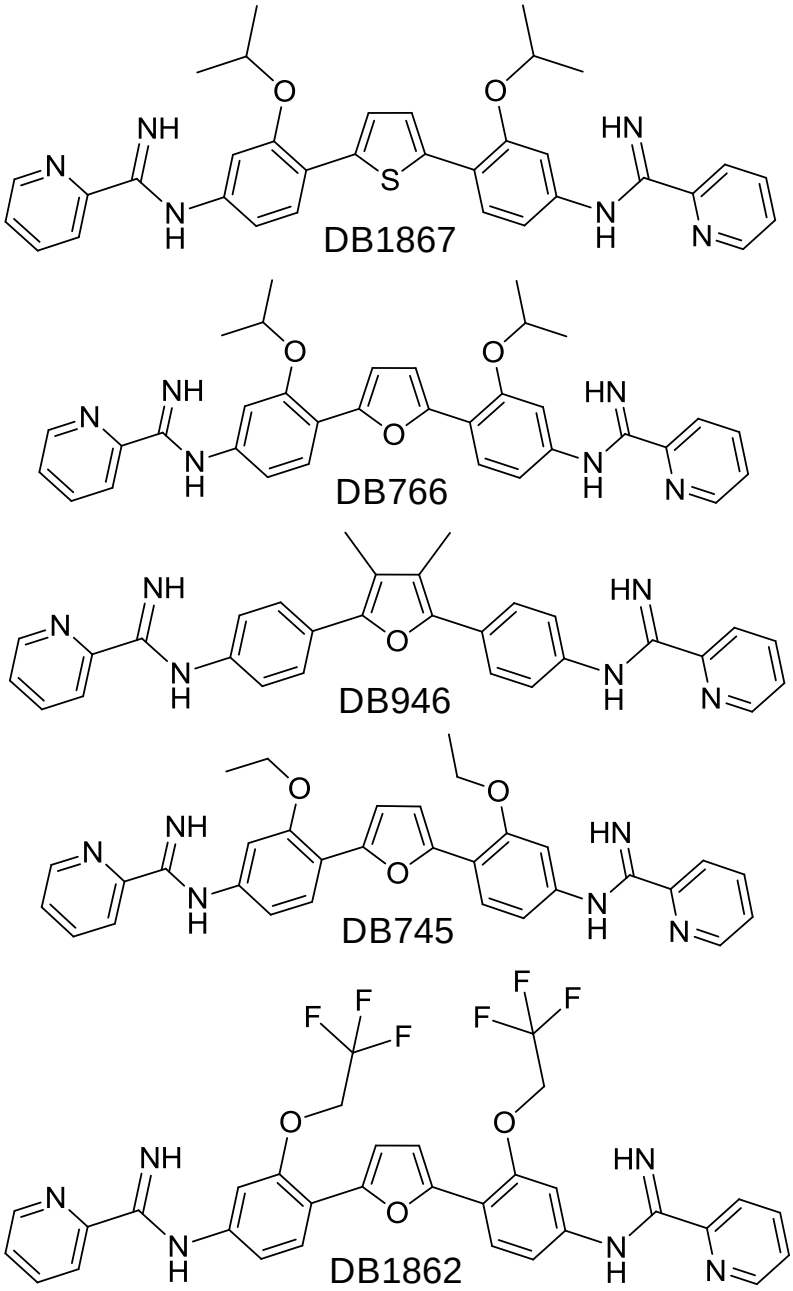


Figure 3. Five of the most active compounds against *L. amazonensis* intracellular parasites and *L. donovani* axenic amastigotes.

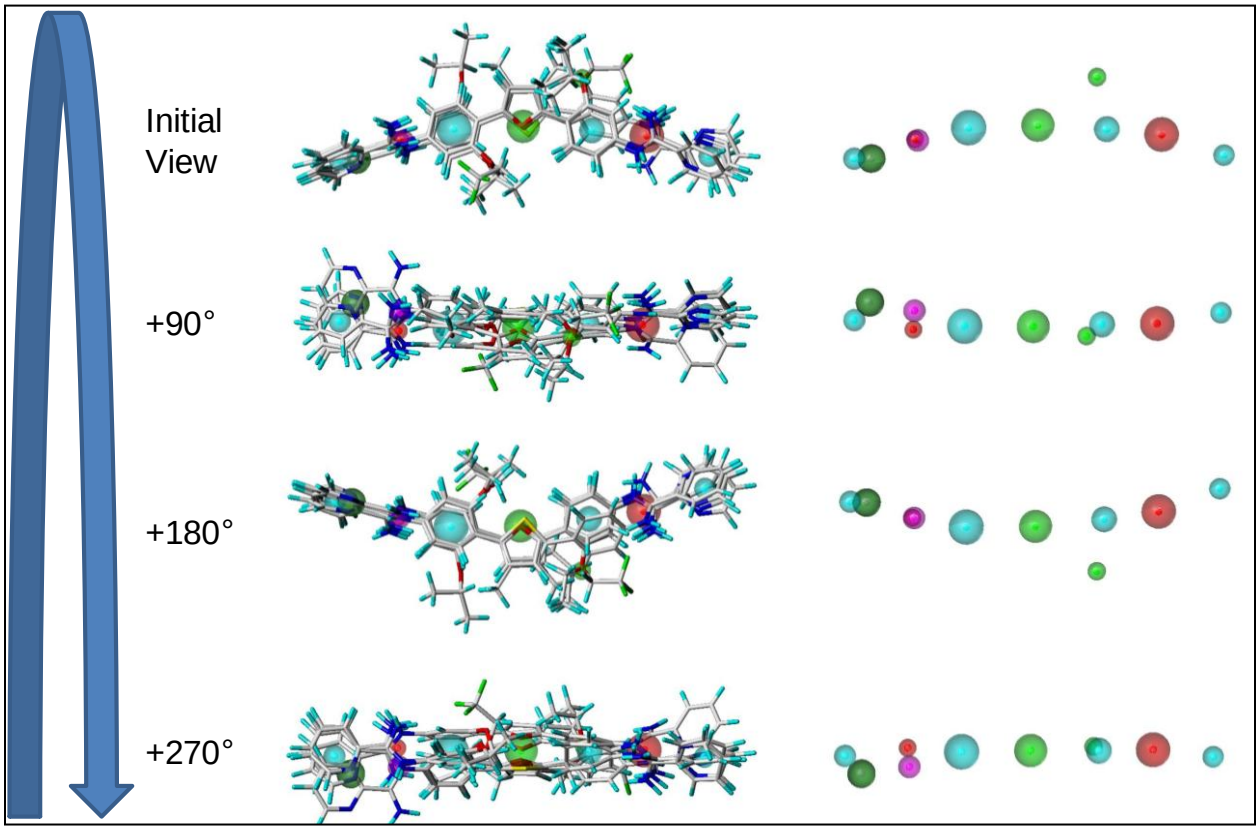


Figure 4. GALAHAD potentials as identified by simulations employing the compounds of Figure 3. The identified features are color coded: cyan, hydrophobes; magenta, donor atoms; green, acceptor atoms; red, positive nitrogens.

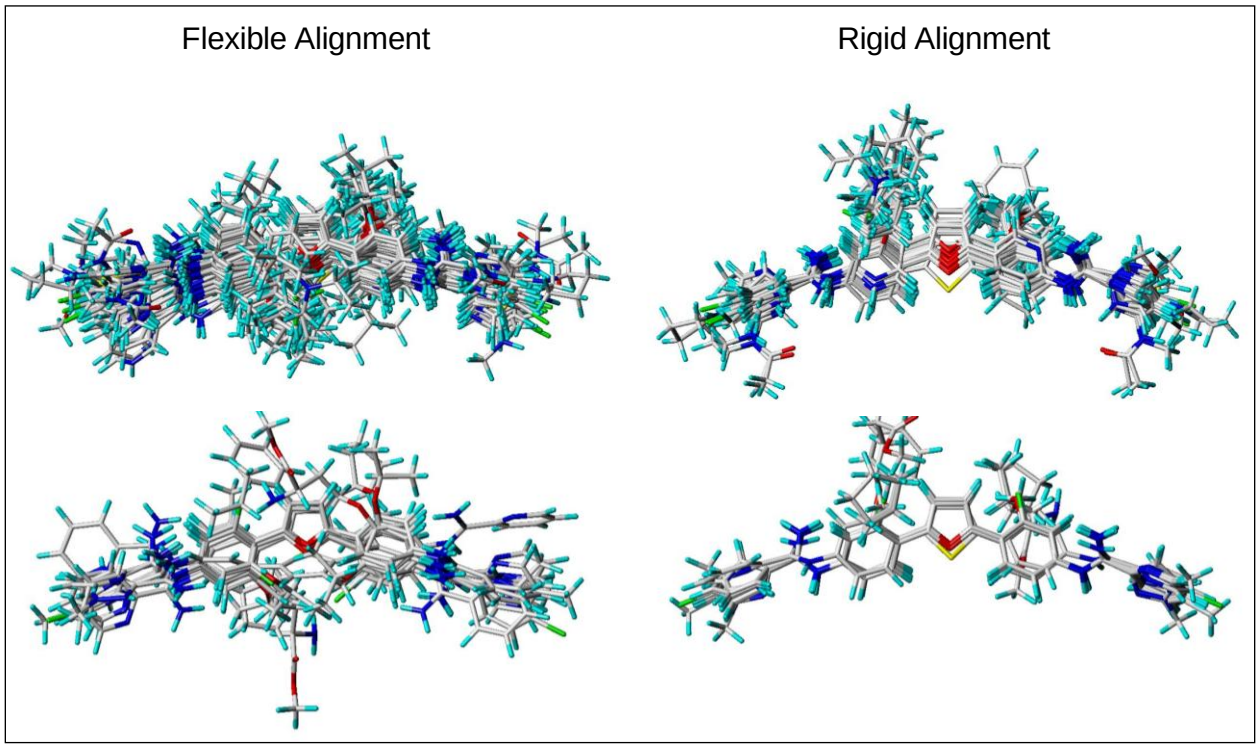


Figure 5. Final training (top) and testing (bottom) datasets: flexible alignments (left) and rigid alignments (right).

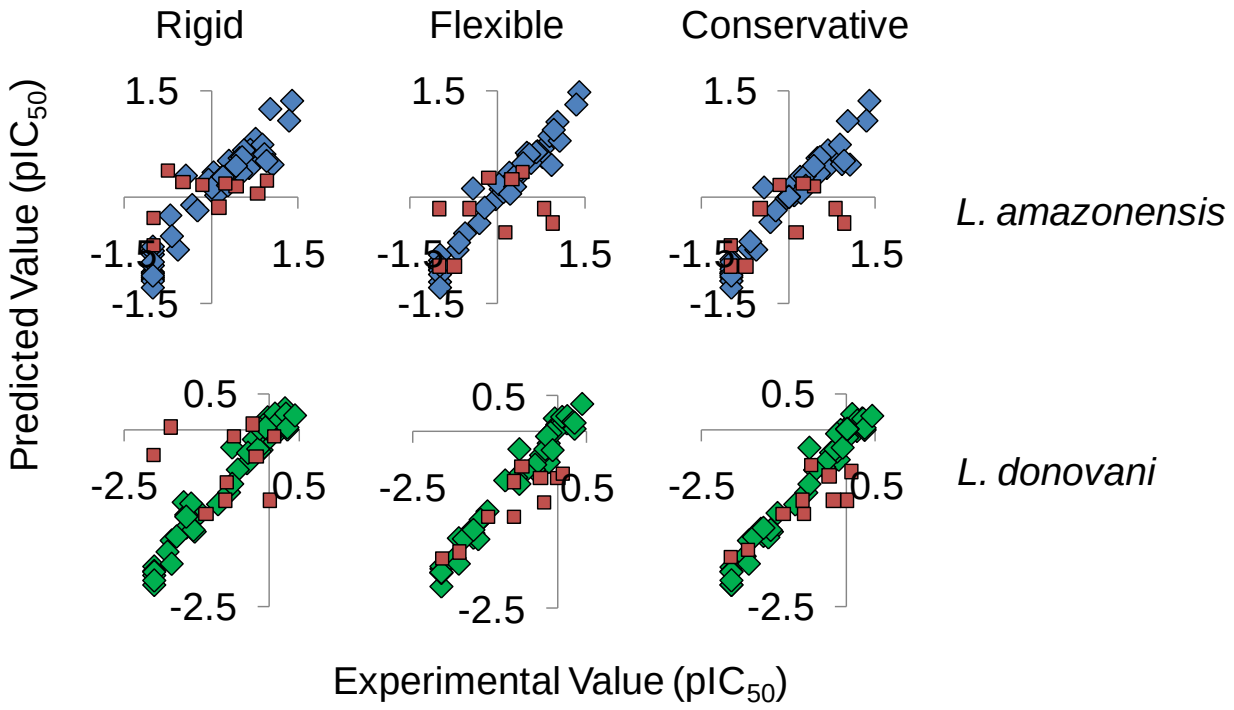


Figure 6. Internal (blue and green) and external (red) predictions. The internal predictions are those for the training dataset compounds implementing the model constructed, whereas external predictions are those for the testing dataset compounds. The models have never seen the testing datasets. The *L. amazonensis* experimental *versus* predicted results are shown in blue above those for *L. donovani* in green. The experimental *versus* predicted results from left to right are predictions from implementing rigid (left) and flexible (center) compounds. The conservative predictions (right) are essentially the more negative of the two pIC_{50} predictions resulting from the models with rigid and flexible compounds. Since the scale observed is the negative log of the IC_{50}, this method reduces under prediction.

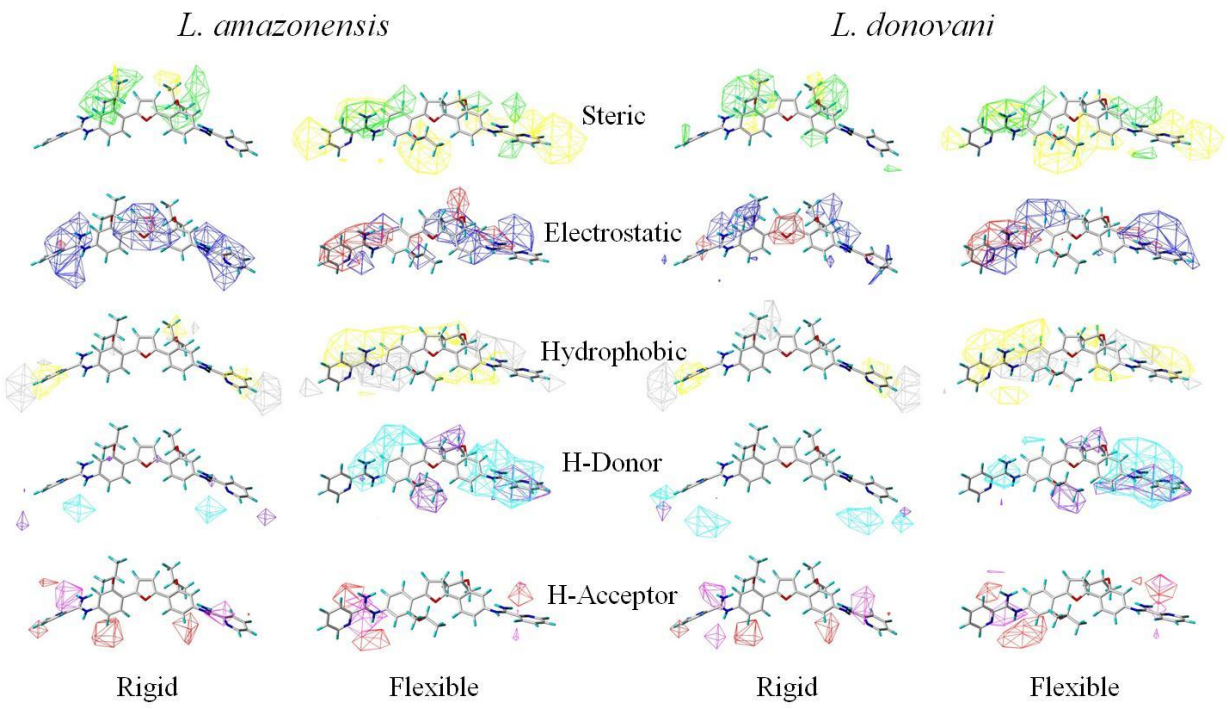


Figure 7. Overall models with CoMSIA molecular descriptors for both rigid and flexible compound alignments. DB766 is displayed as a reference compound for each molecular descriptor potential. Favored potentials from steric to H-acceptor molecular descriptors are green, blue, yellow, cyan, and magenta, whereas disfavored potentials from steric to H-acceptor molecular descriptors are yellow, red, gray, purple, and red.

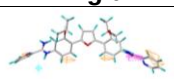
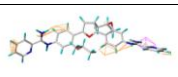
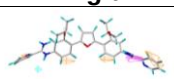
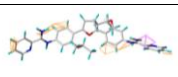
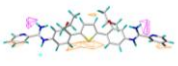
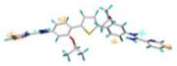
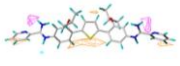
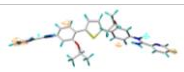
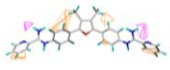
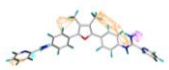
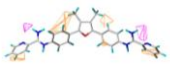
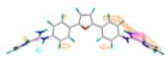
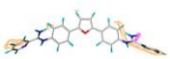
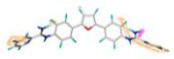
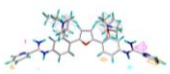
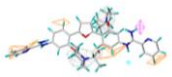
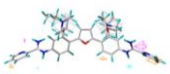
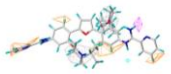
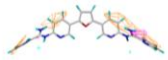
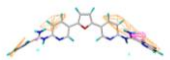
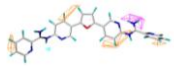
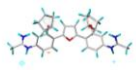
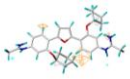
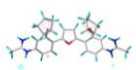
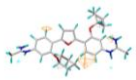
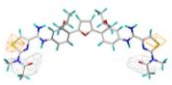
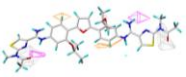
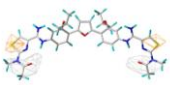
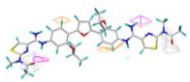
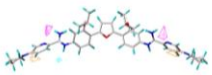
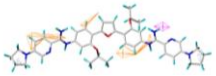
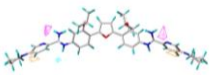
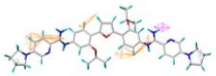
X	Compound	*L. amazonensis*		*L. donovani*	
		Rigid	Flexible	Rigid	Flexible
0	DB766				
1	DB1867				
2	DB946				
3a	DB667				
3b	DB1876				
4	DB1851				
5a	DB1921				
5b	DB1942				
5c	DB1906				

Figure 8. CoMSIA findings with respect to Figure 1 and molecular descriptor potentials of Figure 7. The favored hydrophobic potentials have been changed to orange to improve visualization and insure that steric potentials were not displayed. The left most column consists of numbers correlated to positions of Figure 1. The column to the right consists of the compounds name. This is followed by the compounds and their respective molecular descriptor potentials for each the final models.

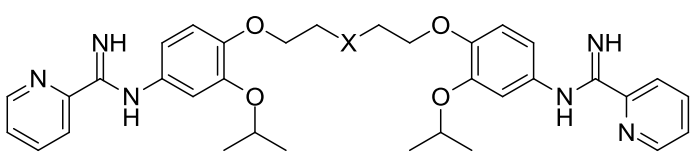
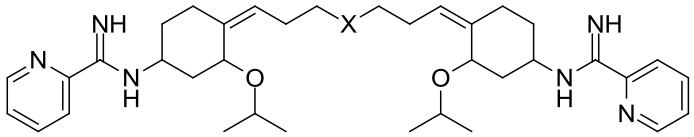
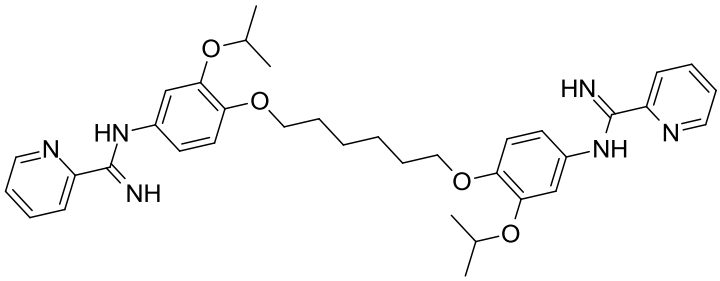
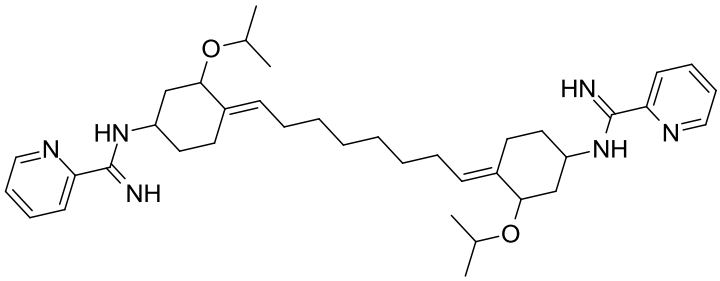
Compound	*L. donovani*	*L. amazonensis*
	X / Predict. C / (0.39,0.96) O / (0.72, 1.8) S / (0.50, 1.3)	X / Predict. C / (0.08, 0.41) O / (0.12, 0.70) S / (0.10, 0.53)
	X / Predict. C / (0.40, 2.0) O / (1.6, 4.5) S / (1.2, 6.1)	X / Predict. C / (0.14, 1.7) O / (0.21, 3.0) S / (0.14, 4.5)
	(0.81, 2.0)	(0.08, 0.28)
	(1.0, 7.3)	(0.16, 3.5)

Figure 9. Compounds designed using the pharmacophore data of Figure 4 and the CoMSIA molecular descriptor fields of Figures 7 and 8. The structures of the compounds are to the left of the predicted IC_{50} ranges. Ranges of predicted IC_{50} values were acquired through the use of both models constructed of rigid and flexible compounds.

CHAPTER 4: SCREENING FOR AFFINITY: PHARMACOPHORE AND QSAR-PLS MODELING OF BIOLOGICAL INHIBITORY DATA FOR COMPOUNDS ACTIVE AGAINST *TRYPANOSOMA CRUZI*

SCREENING FOR AFFINITY: PHARMACOPHORE AND QSAR-PLS MODELING OF BIOLOGICAL INHIBITORY DATA FOR COMPOUNDS ACTIVE AGAINST *TRYPANOSOMA CRUZI*

Catharine J. Collar[1], Elen Mello de Souza[2], Denise da Gama Jaen Batista[2], Cristiane França da Silva[2], Anissa Daliry[2], Melanie Wiggins[1], Maria de Nazaré Correia Soeiro[2], Richard R. Tidwell[3], David W. Boykin[1], and W. David Wilson[1]

From Department of Chemistry, Georgia State University, Atlanta, GA 30303, United States[1], Laboratório de Biologia Celular, Instituto Oswaldo Cruz, Fundação Oswaldo Cruz, Rio de Janeiro, RJ, 21040-900, Brazil[2], Department of Pathology and Laboratory Medicine, School of Medicine, The University of North Carolina, Chapel Hill, North Carolina[3].

Running head: Pharmacophore and QSAR-PLS

Address correspondence to W. David Wilson. Telephone: +1-404-413-5503. Fax: +1-404-413-5551. E-mail: wdw@gsu.edu.

The abbreviations used are: CD, Chagas Disease; Nfx, Nifurtimox; Bz, Benznidazole; DA, Diamidine; AIA, Arylimidamide; CoMFA, Comparative Molecular Field Analysis; CoMSIA, Comparative Molecular Similarity Indices Analysis; QSAR, Quantitative Structure-Activity Relationship; PLS, Partial Least Squares; SEE, Standard Error of Estimate; NIH, National Institutes of Health; CPDD, Consortium for Parasitic Drug Development.

Trypanosoma cruzi, which affects millions of people in endemic areas of Latin America, is the etiological agent of Chagas disease. The available therapy is not ideal since it presents limited efficacy, especially in chronic patients, and displays considerable side effects; thus the need for new trypanocidal compounds is indisputable. *In vitro* assays have been used to determine the biological activities of diamidines and arylimidamides against *T. cruzi* at two tempertures relating to that of blood stored at blood banks (4°C) and that of the human body (37°C). Our studies employ the corresponding biological IC_{50} values acquired to examine compound structural importance through computational biology. Hence, a pharmacophore was identified and implemented to assess quantitative structure-activity relationships (QSAR) through partial least squares (PLS) regression modeling employing the biologically obtained IC_{50} values and computationally calculated comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) molecular descriptors. Statistically significant models were acquired; these models have Q^2 values greater than 0.51 and R^2 values greater than 0.94. Models were internally and externally validated and the molecular descriptor potentials were extrapolated for overall models. The computational data acquired can be used to screen for compounds with inhibitory activities against *T. cruzi* and design novel therapeutic agents.

Introduction

Trypanosoma cruzi is the protozoan parasite that causes Chagas disease (CD), a tropical illness that affects 12-14 million people in many developing countries of Latin America and puts about 50 million at risk of infection.[1] The occurrence of CD in nonendemic regions such as the United States and Europe is mainly due to the migration of infected people but also represents an important concern in these areas.[2-4]

CD has two successive phases: a short acute phase characterized by patent parasitemia followed by a long, progressive chronic phase. The acute phase starts shortly after the infection and is often nonsymptomatic but may manifest as flu-like symptoms with a self-limited febrile illness that lasts a few weeks. If untreated, the symptomatic chronic disease develops in about 20-40% of the infected individuals after a long latent period (several months and even decades), while the majority of the patients remain in the indeterminate state.[5, 6] Due to the long, asymptomatic state, CD is considered a "silent killer," impairing early specific diagnosis and treatment.[7] The main clinical chronic manifestations of CD include cardiac and/or digestive alterations.[8] CD is responsible for considerable rates of mortality and morbidity, however, in the centennial of its discovery by Carlos Chagas (1909), no prophylactic or efficacious treatment is available.[9]

Although they provide limited efficacy and activity upon different parasite stocks, especially for the chronic sufferers, and cause deleterious side effects, the currently accepted

clinical treatments for Chagas are Nifurtimox (Nfx) and Benznidazole (Bz); these two nitrogen-based compounds were introduced more than four decades ago.[10, 11] Thus, the limitation of the current treatment for CD justifies the search and screening of other drugs that could replace Nfx and Bz and/or could be used in cases of therapeutic failure.[12]

Aromatic synthetic diamidine (DA) and arylimidamide (AIA) compounds have been identified as potential therapeutics for CD.[9, 12-22] Our biological assays for DA and AIA compounds have resulted in inhibitory activity assessment for 47 compounds at temperatures of 4°C, the temperature of blood stored in blood banks, and 37°C, the temperature of the human body. In this study, the data from our biological assays were employed to develop predictive models and examine the structural importance of these compounds at the two temperatures. A pharmacophore for active compounds was acquired and implemented in quantitative structure-activity relationship (QSAR) examination through partial least squares (PLS) regression modeling employing biologically obtained inhibitory values, in the form of IC_{50} data, and computationally calculated comparative molecular field analysis (CoMFA) and comparative molecular similarity indices analysis (CoMSIA) molecular descriptors. Subsequent to acquiring statistically significant and validated predictive models, molecular descriptor potentials were extrapolated from final models and employed along with the pharmacophore and predictive models to gain insight into compound structural importance for the inhibition of *T. cruzi*.

Experimental Procedures

Compounds. DA and AIA compounds were synthesized, and stock solutions were prepared in dimethyl sulfoxide (DMSO) with the final concentration not exceeding 0.6%, which did not exert any toxicity towards the parasite or mammalian host cells (data not shown).

Parasites. At peak parasitaemia, bloodstream trypomastigotes (BT; Y strain) were harvested by heart puncture from *T. cruzi* infected Swiss mice as previously reported. All procedures were carried out in accordance with the guidelines established by the FIOCRUZ Committee of Ethics for the Use of Animals (approved protocol number: CEUA L-028/09).

Trypanocidal Analysis. IC_{50} values of 47 compounds were previously published.[13-16, 18-21] These data were acquired through bloodstream trypomastigotes (BT) assays. Treatment entailed different protocols as follows: BT (5 X 10^6 per mL) were incubated for 24 h at 37ºC in RPMI 1640 medium (Roswell Park Memorial Institute, Sigma Aldrich, USA) supplemented with 5% FBS, in the presence or absence of serial dilutions of each compound (0 to 32 μM). Alternately, experiments were performed at 4ºC with BT maintained in freshly isolated mouse blood (96%) in the presence or absence of serial dilutions of the compound (0 to 200 μM). After drug incubation, the parasite death rates were determined by light microscopy through the direct quantification of the number of live parasites using a Neubauer chamber, and the IC_{50} values were then calculated. The IC_{50} values were averaged for at least three determinations done in duplicate.

Computational Biology. All 47 compounds (41 training compounds and 6 testing compounds) with IC_{50} values were constructed in SYBYL 8.1[23] on a Fedora Core 5 Linux workstation. Structures for each compound were energy minimized to convergence using the conjugate gradient method, Tripos force field, and Gasteiger-Hückel charges. The termination gradient was 0.01 kcal/(mol Å) and the maximum iterations were 10^4. Compounds were then semi-randomly separated into training and testing datasets of 41 and 6 compounds, respectively. Compounds selected for the testing dataset represented the entirety of the dataset; they had diverse backbones and biological activities.

The GALAHAD module in SYBYL was employed to gain a pharmacophore for inhibitory compounds of the training dataset; four compounds (DB1831, DB1853, DB1868 and DB766) with low IC_{50} values (Supplemental Table 4, Appendix C) were employed and the parameters were acquired through the "Suggest from Data" option. The best model resulted in maximized pharmacophore consensus, maximized steric consensus, and minimized energy. Due to the structural diversity of compounds in the training and testing datasets, compounds were aligned by the atoms of key features using the "Align Database" option of the QSAR module in SYBYL. Identified important structural atoms were used as a template; DB766 was employed for Cartesian coordinates. Compounds without the common structure were aligned by central atoms that were similar.

PLS analyses employed the training datasets. Computationally calculated CoMFA and CoMSIA molecular descriptors were calculated and mathematically modeled with respect to log

transformed biologically acquired IC_{50} values; $pIC_{50} = -\log(IC_{50})$. Predictive models were gained and implemented to predict pIC_{50} values for compounds of the training and testing datasets. These models were then used to examine molecular descriptor contributions through model extrapolated molecular potentials; visualization was attained by contribution through the mapping of the product standard deviation with respect to molecular descriptor values and coefficients (S.D.*Coeff.) at each lattice point. The default levels of the contour by contribution were employed as follows: 80% for a favored region and 20% for a disfavored region.

Results

Inhibitory experimental assays against *T. cruzi* at 4°C and 37°C provided IC_{50} values for 47 compounds (Supplemental Table 4, Appendix C). All compounds with inhibitory data were constructed in SYBYL and four compounds with low IC_{50} values, for both assays, were employed for pharmacophore identification. It is important to note that the best inhibitory compounds of this dataset are all AIAs. Given the structural variation of this dataset, only a select set of features are applicable for the alignment of all compounds.

Figure 1 displays the pharmacophore potentials; structural importance appears to be attributed to the central furan and its flanking aromatic rings. These rings are identified by hydrophobe potentials. Using DB766 as a reference compound (Figure 2), Figure 1 also displays: (1) a fourth hydrophobe potential residing on one of the isopropyl substituents, (2) donor potentials at the N=C−N groups, (3) an acceptor potential on the furan O, as well as on both

isopropyl O, and (4) a positive N potential at one N=C−N group. Perhaps the importance of the one identified positive N comes from the net compound +1 charge of most AIAs. The pharmacophore identification appeared to highlight the rigid central structure of the compounds; the central structure is similar for AIAs and DAs.

Atoms representing the hydrophobe and acceptor potentials were identified and implemented for training and testing dataset alignment; the atoms used in alignment are those identified in Figure 2. An effective overlay of the 41 training and 6 testing compounds allowed for three-dimensional compound comparison. The identified pharmacophore regions that were not used for alignment were still maintained by AIAs (Figure 3). With compounds aligned, CoMFA and CoMSIA molecular descriptors were calculated and these were employed along with the experimentally obtained inhibitory values for QSAR studies with PLS modeling. The models constructed have correlation coefficients (Q^2) values greater than 0.51, standard error of estimate (SEE) values lower than 0.29, coefficient of determination (R^2) values greater than 0.94, and large F statistics (Table 1).

Both CoMFA models consisted of approximately 70% steric and 30% electrostatic molecular descriptor contributions, while the CoMSIA models exhibited more variation. The model employing experimental inhibitory data at 4°C consisted of roughly 13% steric, 13% electrostatic, 23% hydrophobic, 26% H-donor, and 25% H-acceptor contributions, whereas the model with experimental inhibitory data at 37°C consisted of roughly 14% steric, 13% electrostatic, 26% hydrophobic, 24% H-donor, and 23% H-acceptor contributions. The molecular

descriptor contributions for both CoMSIA models were within 3% of each other; this difference is not statistically significant. However, molecular descriptor contributions are related to the weights that are employed by PLS to relate structural importance to inhibitory activity. This relationship results in predictions. These molecular descriptor contributions suggest that predictions from models employing CoMFA molecular descriptors will be more similar than those from models employing CoMSIA molecular descriptors.

Figure 4 displays the training and testing dataset predictions for the PLS models; trendlines are displayed for the training dataset predictions. Training data are displayed in blue and green for the assays at 4°C and 37°C, respectively. Notice that the slopes of the trendlines are all near one as expected for a valid model. Also, there are very few noticeable outliers. Hence, the models should all be able to provide useful predictions for the compounds of the testing dataset. This holds true according to the testing dataset predictions, the CoMFA model at 4°C and the CoMSIA model at 37°C are outperforming the other two models. The compounds of the testing dataset can be viewed along with their experimental and predicted data in Table 2. Notice that each of the testing compounds is quite different; they vary in size, shape, and conformation. This allows for a full predictability range inspection based on compound structure.

The phenyl-pyridine DA DB1627 consists of accurate predictions, especially since the inhibitory values for this compound are at an experimental cut-off. 10SAB031 is a DA that consists of a triazole center ring and flanking aromatic rings, the amidines are in the *m*-position. The models constructed of CoMFA molecular descriptors predicted better than those with

CoMSIA molecular descriptors. DB1362 consists of a 3-bromo-4-methylthiophene central ring and flanking aromatic rings. The amidines for this DA compound are in the *p*-position. The model constructed of experimental data at 4°C and CoMFA molecular descriptors predicted best, as did the model with experimental data at 37°C and CoMSIA molecular descriptors. 14SMB013 is a diamidine compound that has an aliphatic linker instead of a central ring. Again, the model constructed of experimental data at 4°C and CoMFA molecular descriptors predicted best, as did the model with experimental data at 37°C and CoMSIA molecular descriptors. AIA 613A, has a scaffold structure similar to DB766; yet, this compound lacks isopropyl substituents and N in the outer aromatic rings. As before, the model constructed of experimental data at 4°C and CoMFA molecular descriptors predicted best, as did the model with experimental data at 37°C and CoMSIA molecular descriptors. DB1868 is the most active compound of the testing dataset. This compound is similar in structure to DB766; it has additional O-Me at the *p*-positions of the outer aromatic rings. Both models constructed from experimental inhibitory data at 4°C predict well, while those for data at 37°C are over predicting significantly. It is also important to note that the largest deviations in predictions occurred when the experimental values between the two assays were the greatest. These deviations are seen in the predictions for 10SAB031, 14SMB013, and DB1868. The residuals between the two biological assays are 0.32, 0.53, and 0.65, respectively.

Molecular descriptor potentials were extrapolated from all four models (Figures 5 and 6). Figure 5 displays the steric and electrostatic potentials for the models constructed with CoMFA molecular descriptors and inhibitory data at 4°C and 37°C, respectively. The model with CoMFA

molecular descriptors and inhibitory data at 4°C show that steric bulk is favored (green) on both ends of reference compound DB766, as well as above the furan and an isopropyl group. Yet, disfavored steric potentials (yellow) suggest a size limit to regions near isopropyl substituents. The electrostatic potentials display the N=C−N groups to exhibit favored (blue) positive charge. There are very few areas which call for negative charge (red). The model with CoMFA molecular descriptors and inhibitory data at 37°C display steric potentials similar to those of the model constructed from the data collected at 4°C. The electrostatic potentials again show importance of an N=C−N group and suggests more regions where negative charge is favored.

Figure 6 displays the molecular descriptor potentials for the models employing CoMSIA molecular descriptors. These models displayed smaller more defined regions of importance than those previously discussed. The model constructed of CoMSIA molecular descriptors and inhibitory data at 4°C, show similar steric findings to those from the CoMFA models. Steric bulk is favored at both ends of the compound and there is a size limit to regions of isopropyl substituents. Electrostatic potentials identify important positive charge regions as those consisting of isopropyl substituents. Hydrophobic regions are favored (yellow) near the outer aromatic rings and disfavored (white) by the central furan ring. H-donors are favored (cyan) near the N=C−N groups, as are H-acceptors (magenta). The model constructed of CoMSIA molecular descriptors and inhibitory data at 37°C displays similar results to the model with CoMSIA molecular descriptors and inhibitory data at 4°C. Slight differences in steric potentials were seen in a shift of positive steric bulk potential from one isopropyl substituent region to the other

isopropyl substituent. Favored positive electrostatic potentials were reduced and favored negative electrostatic potentials were increased. A greater amount of favored hydrophobic potential resulted along the backbone of the reference structure, whereas the amount of disfavored hydrophobic potential expanded into regions that were previously favored. Favored H-donor potentials shifted toward the outer aromatic rings, while disfavored potentials were increased near one of these rings. The favored potentials for H-acceptors surround the N=C−N groups, whereas disfavored H-acceptor areas are located below the interior aromatic and furan rings, as well as near one of the isopropyl substituents.

Discussion

Found historically in rural areas of Latin America among those living in close proximity with vectors and in poor housing conditions, Chagas disease is spread through vectors, transfusion, organ transplant, and from mother to infant.[4, 12] This disease has spread with migration to the United States and Europe. An estimated 100,000 people in the United States have this disease which is often unrecognized until the chronic phase is reached. The only accepted clinical treatments are Nfx and Bz, both of which are not approved for treatment in the United States. The only way to obtain these compounds is from the CDC, due to their adverse side effects that can be as devastating as the disease. Hence, the spread of Chagas disease and the limitations of current therapies necessitate the screening and development of therapeutics that could replace Nfx and Bz or be used in cases of therapeutic failure. Since several studies have examined the inhibitory activities of compounds that show inhibitory affinity for targeting *T.*

cruzi,[13-16, 18-21] this data is readily available and can be employed to develop effective screening devices and gain insight into the inhibition mechanisms of the compounds.

A pharmacophore was deduced from AIAs with high inhibitory affinities (Figure 1). The pharmacophore displays the importance of the rigid three ring system through the identification by hydrophobes. Three acceptor positions of these compounds also appear important; hence, the atoms surrounding the identified hydrophobes and acceptors of Figure 2 were used for the alignment of all compounds (Figure 3). PLS was then employed to construct predictive models that implemented experimental inhibitory data and respective three-dimensional compound conformations that were defined by the molecular descriptors used, CoMFA or CoMSIA. The four models developed were all statistically significant and validated accordingly (Tables 1 and 2 and Figure 4). Although all models were shown to be useful predictive devices the models that are most optimal for predicting inhibitory activity accurately are: (1) the model employing CoMFA molecular descriptors and experimental inhibitory data acquired at 4°C, and (2) the model employing CoMSIA molecular descriptors and experimental inhibitory data acquired at 37°C. This appeared to be counterintuitive until models were further examined; the other two models were much more rigid and this was reflected in predictions. Through the analysis of model predictions it was found that large deviations in predictions were acquired most often when the experimental values between the two assays were large.

Molecular descriptor potentials were extrapolated from the models and used to gain insight about important structural contributions that may be employed to improve compound

design (Figures 5 and 6). The most useful data appears to be that which comes from the most optimal models that were previously identified. Steric and electrostatic molecular descriptors for the model employing CoMFA molecular descriptors and experimental inhibitory data acquired at 4°C displayed that: (1) compounds can be elongated; (2) there is some room for modification of the isopropyl substituent region of reference compound DB766, although there appear to be some size limitations; and (3) positive charge is shown to be important to regions of N=C−N groups (Figure 5). This suggests that the pharmacophore identification of the positive nitrogen potential, and two donor potentials at the N=C−N groups, have some relevance to important inhibitory structure for treating blood stored at 4°C (Figure 1). Steric, electrostatic, hydrophobic, H-donor and H-acceptor molecular descriptors for the model employing CoMSIA molecular descriptors and experimental inhibitory data acquired at 37°C displayed that: (1) compounds can be elongated; (2) there is some room for modification of the isopropyl substituent region of reference compound DB766, although there appear to be some size limitations; (3) the outer hydrophobic rings appear to be of importance; (4) the addition of H-donors to *p*-position of the outer aromatic rings may lead to improved inhibitory compounds; and (5) H-acceptors are favored at the N=C−N groups of reference structure DB766 (Figure 6).

In summary, a pharmacophore for highly inhibitory compounds was identified, predictive devices for the screening of inhibitory activity at 4°C and 37°C were constructed for DA and AIA compounds, and molecular descriptor potentials have been extracted to determine structural importance as related to the design for novel therapeutics. Structural importance for inhibiting *T.*

cruzi was found to be an overall rigid conformation, similar to that of current AIAs, that has N=C−N groups. Areas for modifications have been identified as the *p*-position of the outer aromatic rings, isopropyl substituent regions of reference compound DB766, and central atoms of the AIAs.

Acknowledgments

Support was provided by the National Institutes of Health (NIH) AI064200, by the Bill and Melinda Gates Foundation through the Consortium for Parasitic Drug Development (CPDD), by the Georgia State University Molecular Basis of Disease Fellowship, by CNPq and PAPES V/Fiocruz, and by the David W. Boykin Graduate Fellowship in Medicinal Chemistry.

References

1. Dias, J. C., Elimination of Chagas disease transmission: perspectives. *Mem Inst Oswaldo Cruz* **2009,** 104 Suppl 1, 41-5.
2. Gascon, J.; Albajar, P.; Canas, E.; Flores, M.; Gomez i Prat, J.; Herrera, R. N.; Lafuente, C. A.; Luciardi, H. L.; Moncayo, A.; Molina, L.; Munoz, J.; Puente, S.; Sanz, G.; Trevino, B.; Sergio-Salles, X., [Diagnosis, management and treatment of chronic Chagas' heart disease in areas where Trypanosoma cruzi infection is not endemic]. *Rev Esp Cardiol* **2007,** 60, (3), 285-93.

3. Rodriguez-Morales, A. J.; Benitez, J. A.; Tellez, I.; Franco-Paredes, C., Chagas disease screening among Latin American immigrants in non-endemic settings. *Travel Med Infect Dis* **2008,** 6, (3), 162-3.
4. Bern, C.; Montgomery, S. P.; Herwaldt, B. L.; Rassi, A., Jr.; Marin-Neto, J. A.; Dantas, R. O.; Maguire, J. H.; Acquatella, H.; Morillo, C.; Kirchhoff, L. V.; Gilman, R. H.; Reyes, P. A.; Salvatella, R.; Moore, A. C., Evaluation and treatment of chagas disease in the United States: a systematic review. *Jama* **2007,** 298, (18), 2171-81.
5. Bilate, A. M.; Cunha-Neto, E., Chagas disease cardiomyopathy: current concepts of an old disease. *Rev Inst Med Trop Sao Paulo* **2008,** 50, (2), 67-74.
6. Rassi, A., Jr.; Rassi, A.; Marin-Neto, J. A., Chagas heart disease: pathophysiologic mechanisms, prognostic factors and risk stratification. *Mem Inst Oswaldo Cruz* **2009,** 104 Suppl 1, 152-8.
7. Tarleton, R. L.; Reithinger, R.; Urbina, J. A.; Kitron, U.; Gurtler, R. E., The challenges of Chagas Disease-- grim outlook or glimmer of hope. *PLoS Med* **2007,** 4, (12), e332.
8. Sosa-Estani, S.; Viotti, R.; Segura, E. L., Therapy, diagnosis and prognosis of chronic Chagas disease: insight gained in Argentina. *Mem Inst Oswaldo Cruz* **2009,** 104 Suppl 1, 167-80.
9. Soeiro, M. N.; de Castro, S. L., Trypanosoma cruzi targets for new chemotherapeutic approaches. *Expert Opin Ther Targets* **2009,** 13, (1), 105-21.
10. Coura, J. R.; Castro, S. L. d., A Critical Review on Chagas Disease Chemotherapy. *Memórias do Instituto Oswaldo Cruz* **2002,** 97, 3-24.

11. Urbina, J. A., Ergosterol biosynthesis and drug development for Chagas disease. *Mem Inst Oswaldo Cruz* **2009,** 104 Suppl 1, 311-8.
12. Soeiro, M. d. N. C.; Dantas, A. P.; Daliry, A.; Silva, C. F. d.; Batista, D. G.; Souza, E. M. d.; Oliveira, G. M.; Salomão, K.; Batista, M. M.; Pacheco, M. G.; Silva, P. B. d.; Santa-Rita, R. M.; Barreto, R. F. M.; Boykin, D. W.; Castro, S. L. d., Experimental chemotherapy for Chagas disease: 15 years of research contributions from in vivo and in vitro studies. *Memórias do Instituto Oswaldo Cruz* **2009,** 104, 301-310.
13. Batista Dda, G.; Batista, M. M.; de Oliveira, G. M.; do Amaral, P. B.; Lannes-Vieira, J.; Britto, C. C.; Junqueira, A.; Lima, M. M.; Romanha, A. J.; Sales Junior, P. A.; Stephens, C. E.; Boykin, D. W.; Soeiro Mde, N., Arylimidamide DB766, a potential chemotherapeutic candidate for Chagas' disease treatment. *Antimicrob Agents Chemother* **2010,** 54, (7), 2940-52.
14. Batista, D. G.; Pacheco, M. G.; Kumar, A.; Branowska, D.; Ismail, M. A.; Hu, L.; Boykin, D. W.; Soeiro, M. N., Biological, ultrastructural effect and subcellular localization of aromatic diamidines in Trypanosoma cruzi. *Parasitology* **2010,** 137, (2), 251-9.
15. da Silva, C. F.; Batista, M. M.; Batista, D. d. G. J.; de Souza, E. M.; da Silva, P. B.; de Oliveira, G. M.; Meuser, A. S.; Shareef, A.-R.; Boykin, D. W.; Soeiro, M. d. N. C., In Vitro and In Vivo Studies of the Trypanocidal Activity of a Diarylthiophene Diamidine against Trypanosoma cruzi. *Antimicrob. Agents Chemother.* **2008,** 52, (9), 3307-3314.

16. da Silva, C. F.; da Silva, P. B.; Batista, M. M.; Daliry, A.; Tidwell, R. R.; Soeiro Mde, N., The biological in vitro effect and selectivity of aromatic dicationic compounds on Trypanosoma cruzi. *Mem Inst Oswaldo Cruz* **2010,** 105, (3), 239-45.
17. Daliry, A.; Da Silva, P. B.; Da Silva, C. F.; Batista, M. M.; De Castro, S. L.; Tidwell, R. R.; Soeiro Mde, N., In vitro analyses of the effect of aromatic diamidines upon Trypanosoma cruzi. *J Antimicrob Chemother* **2009,** 64, (4), 747-50.
18. De Souza, E. M.; Lansiaux, A.; Bailly, C.; Wilson, W. D.; Hu, Q.; Boykin, D. W.; Batista, M. M.; Araujo-Jorge, T. C.; Soeiro, M. N., Phenyl substitution of furamidine markedly potentiates its anti-parasitic activity against Trypanosoma cruzi and Leishmania amazonensis. *Biochem Pharmacol* **2004,** 68, (4), 593-600.
19. Pacheco, M. G.; da Silva, C. F.; de Souza, E. M.; Batista, M. M.; da Silva, P. B.; Kumar, A.; Stephens, C. E.; Boykin, D. W.; Soeiro Mde, N., Trypanosoma cruzi: activity of heterocyclic cationic molecules in vitro. *Exp Parasitol* **2009,** 123, (1), 73-80.
20. Silva, C. F.; Batista, M. M.; Mota, R. A.; de Souza, E. M.; Stephens, C. E.; Som, P.; Boykin, D. W.; Soeiro Mde, N., Activity of "reversed" diamidines against Trypanosoma cruzi "in vitro". *Biochem Pharmacol* **2007,** 73, (12), 1939-46.
21. Silva, C. F.; Meuser, M. B.; De Souza, E. M.; Meirelles, M. N.; Stephens, C. E.; Som, P.; Boykin, D. W.; Soeiro, M. N., Cellular effects of reversed amidines on Trypanosoma cruzi. *Antimicrob Agents Chemother* **2007,** 51, (11), 3803-9.

22. Soeiro, M. N.; De Souza, E. M.; Stephens, C. E.; Boykin, D. W., Aromatic diamidines as antiparasitic agents. *Expert Opin Investig Drugs* **2005,** 14, (8), 957-72.

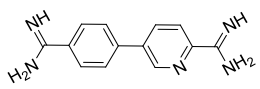
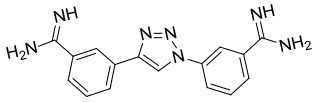
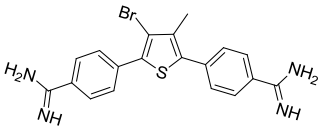
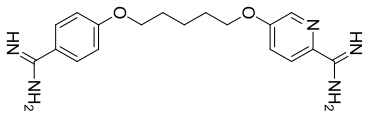
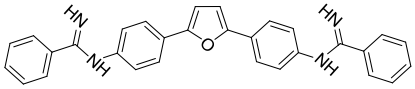
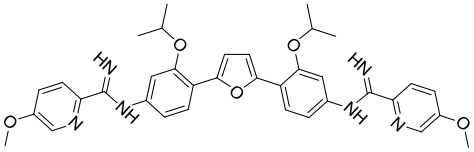
23. *SYBYL Molecular Modeling Software, 8.1 ed.*, Tripos Inc.: St. Louis, MO, 2008.

Tables and Figures

Table 1. Statistics of partial least squares predictive models for a biological dataset of synthetic diamidines and arylimidamides with activities against *Trypanosoma cruzi* at 4°C and 37°C. Models employ either CoMFA or CoMSIA molecular descriptors. The optimal N components were determined by the smallest predicted error sum of squares; N was determined to be optimal at 3, 4, 4, and 3 for models displayed from left to right, respectively.

	CoMFA		CoMSIA	
	4°C	37°C	4°C	37°C
Q^2	0.58	0.56	0.51	0.54
SEE	0.27	0.28	0.16	0.26
R^2	0.94	0.95	0.98	0.95
F	180	160	420	260

Table 2. Experimental and predicted pIC_{50} values for test set compounds. The name of the structure is displayed to the far left; this is followed by the structure, the experimental pIC_{50} values at both 4°C and 37°C, and the predictions from the respective models employing CoMFA or CoMSIA molecular descriptors at the two temperatures. Diamidines have an overall +2 charge and arylimidamides have an overall +1 charge.

Name	Structure	Experimental		CoMFA		CoMSIA	
		4°C	37°C	4°C	37°C	4°C	37°C
DB1627		-1.5	-1.5	-1.6	-1.7	-1.6	-1.7
10SAB031		-0.29	-0.61	-0.13	-0.36	0.48	-0.28
DB1362		-0.85	-0.82	-0.66	-1.3	-0.05	-1.1
14SMB013		-1.5	-0.97	-1.3	-1.5	-1.0	-1.3
DB613A		-1.5	-1.5	-1.1	-2.1	-0.33	-1.4
DB1868		0.55	1.2	1.2	-1.2	1.3	-1.4

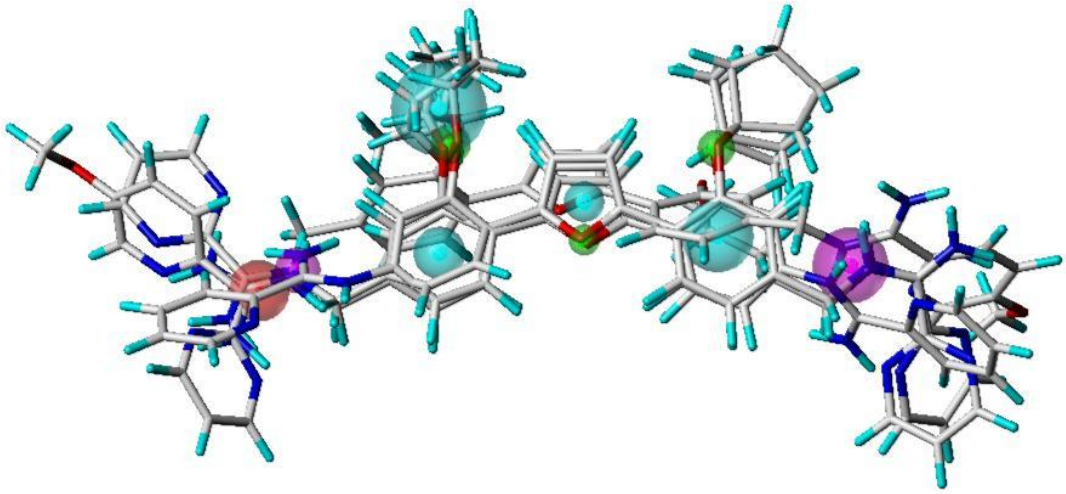


Figure 1. GALAHAD potentials as identified by simulations employing four arylimidamide compounds (DB1831, DB1853, DB1868 and DB766, Supplemental Table 4, Appendix C) with high inhibitory affinity. The identified features are color coded: cyan, hydrophobes; magenta, donor atoms; green, acceptor atoms; red, positive nitrogens.

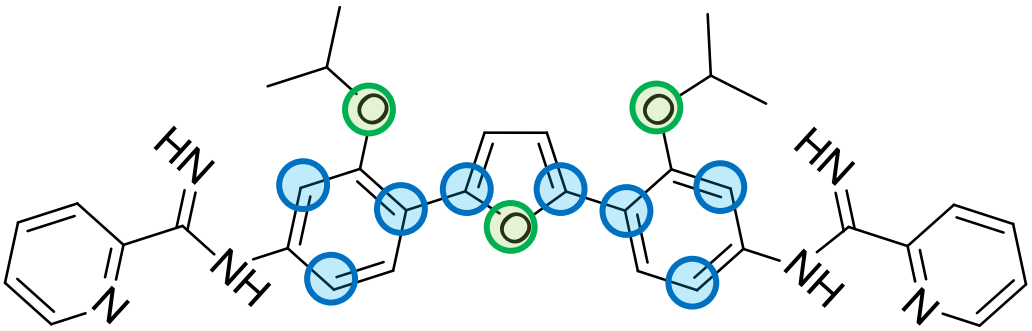


Figure 2. Alignment atoms identified on the arylimidamide DB766; these are color coded as in Figure 1. The atoms representative of the hydrophobes are in cyan and those of the acceptor atoms are in green.

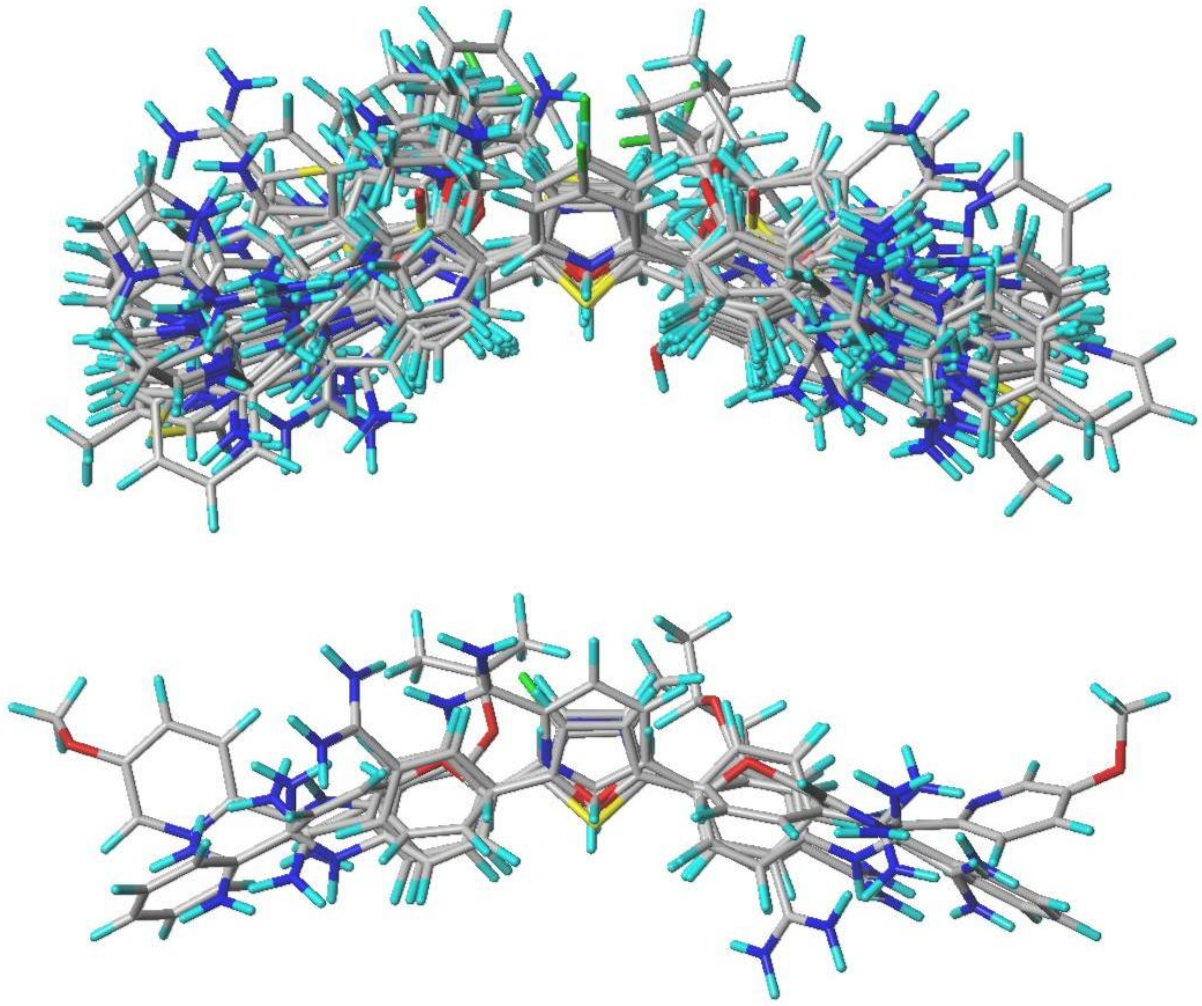


Figure 3. The training dataset of 41 compounds (top) was employed to construct partial least squares regression models, whereas the testing dataset of 6 compounds (bottom) was employed to assess the models constructed.

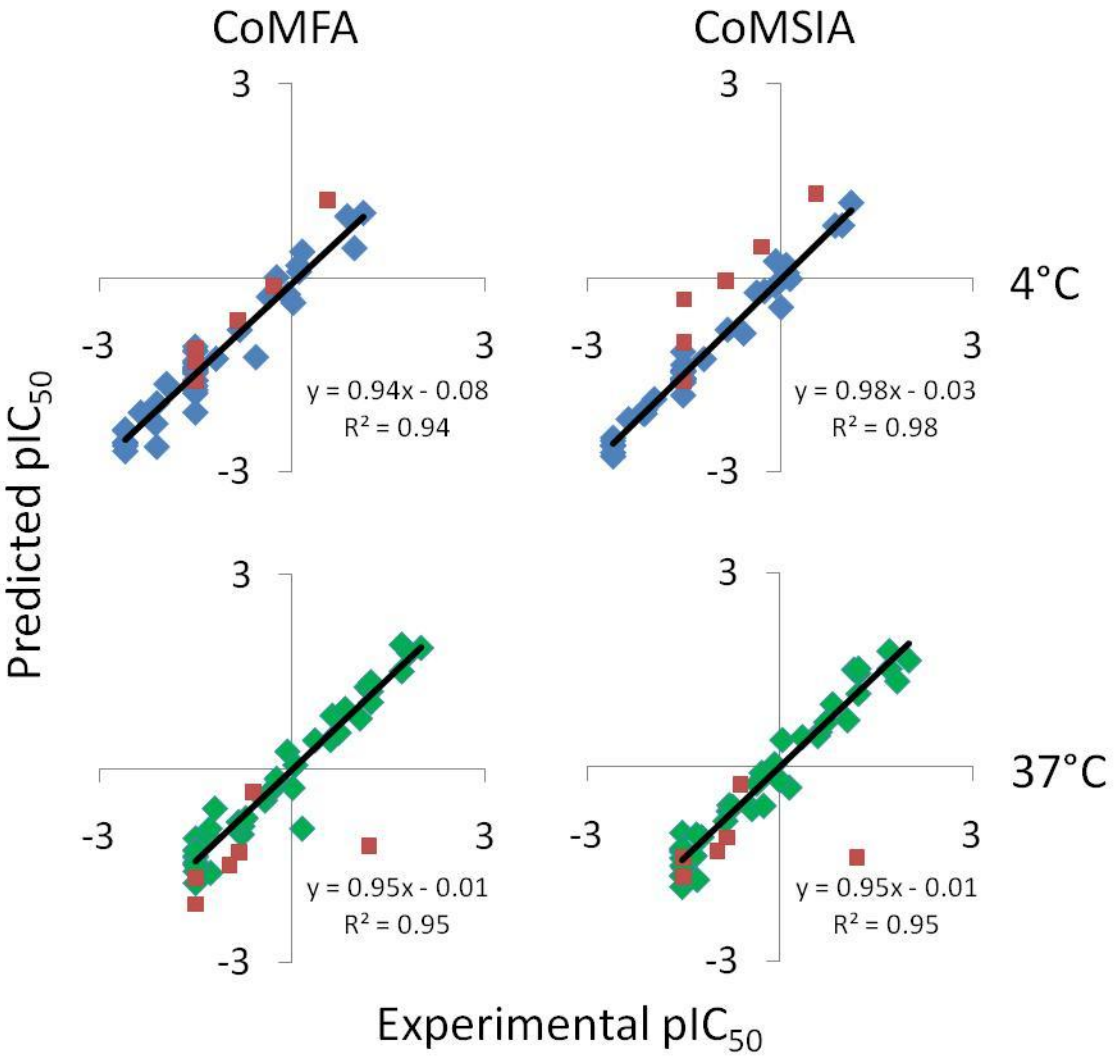


Figure 4. Predictions for the training (blue and green) and testing (red) datasets are displayed with respect to experimental data. The data from models employing CoMFA molecular descriptors are on the left, whereas those using CoMSIA molecular descriptors are on the right. Experimental data for assays and respective predictions at 4°C are displayed above those at 37°C. The trendlines for the training dataset predictions are displayed along with their respective equations and R^2 values.

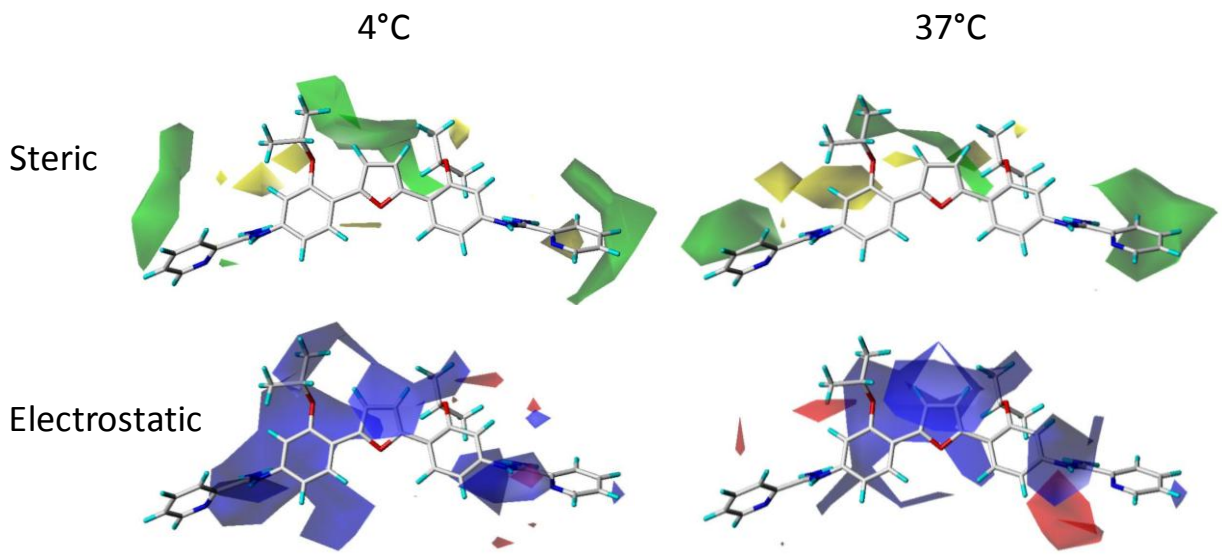


Figure 5. Potentials for models employing CoMFA molecular descriptors and biological pIC_{50} values. Favored steric and positive electrostatic potentials are shown in green and blue, whereas disfavored potentials are displayed in yellow and red, respectively. DB766 is used as a reference compound; the data displayed are for the overall models. The models for inhibition at 4°C are displayed to the left of the models for inhibition at 37°C and the steric molecular descriptor potentials are shown above respective electrostatic potentials.

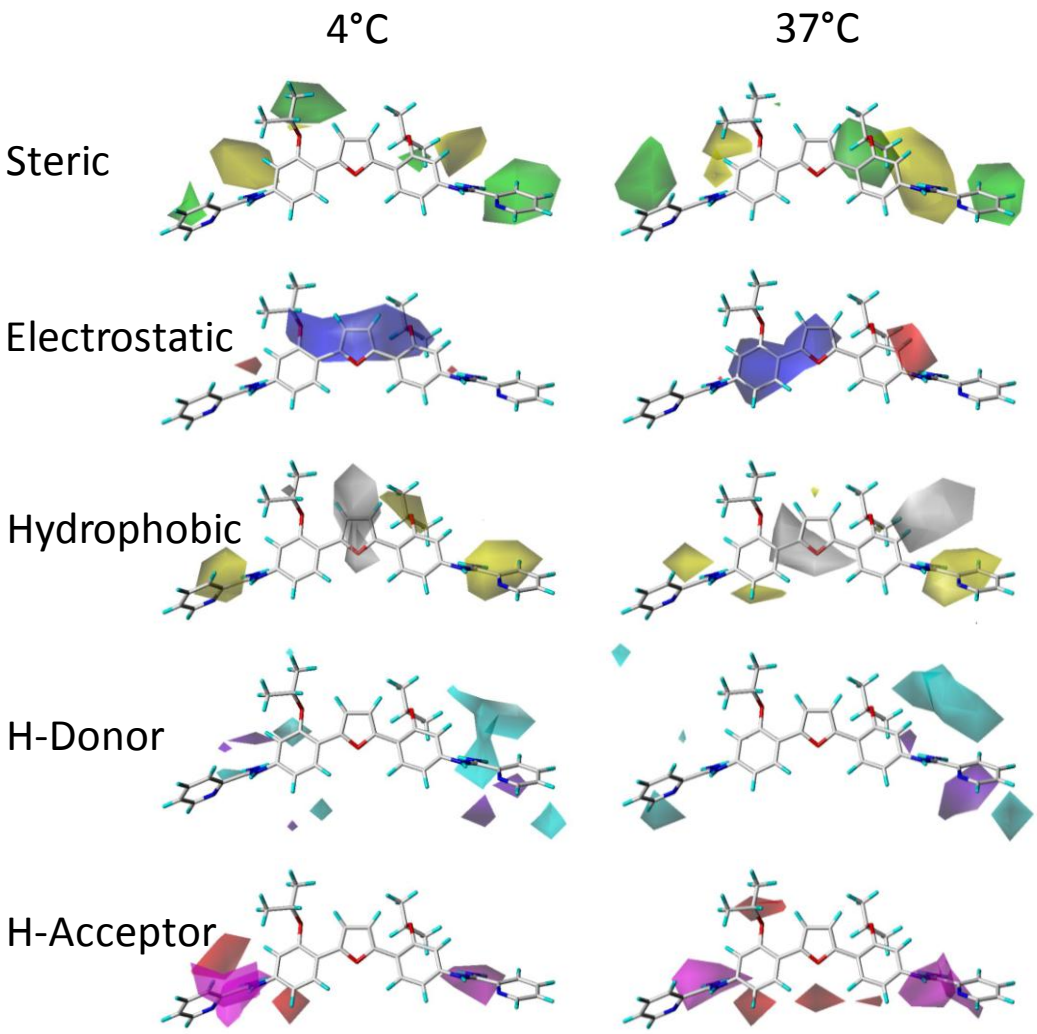


Figure 6. Potentials for models employing CoMSIA molecular descriptors and biological pIC_{50} values. Favored steric, positive electrostatic, hydrophobic, H-donor, and H-acceptor potentials are shown in green, blue, yellow, cyan, and magenta; negative potentials are displayed in yellow, red, white, purple, and red, respectively. As in Figure 5, DB766 is used as a reference compound and the data are displayed for the overall models.

CHAPTER 5: SETTING ANCHOR IN THE MINOR GROOVE: IN SILICO INVESTIGATION INTO FORMAMIDO N-METHYLPYRROLE AND N-METHYLIMIDAZOLE POLYAMIDES BOUND BY COGNATE DNA SEQUENCES

SETTING ANCHOR IN THE MINOR GROOVE: *IN SILICO* INVESTIGATION INTO FORMAMIDO *N*-METHYLPYRROLE AND *N*-METHYLIMIDAZOLE POLYAMIDES BOUND BY COGNATE DNA SEQUENCES

Catharine J. Collar[1], Moses Lee[2], and W. David Wilson[1]

From Department of Chemistry, Georgia State University, Atlanta, Georgia 30303[1] and Department of Chemistry and the Division of Natural and Applied Sciences, Hope College, Holland, Michigan 49423[2].

Running head: Base pair recognition by f Py Im polyamides

Address correspondence to W. David Wilson. Telephone: +1-404-413-5503. Fax: +1-404-413-5551. E-mail: wdw@gsu.edu.

The abbreviations used are: Py, *N*-methylpyrrole; Im, *N*-methylimidazole; f, formamido; DNA, Deoxyribonucleic Acid; A, Adenine; T, Thymine; G, Guanine; C, Cytosine; NMR, Nuclear Magnetic Resonance; ASA, Accessible Solvent Area; NSF, National Science Foundation.

Tricyclic *N*-Methylpyrrole (Py) and *N*-methylimidazole (Im) containing polyamide monocations are known to bind as stacked dimers within the minor groove of DNA and those with *N*-terminal formamido (f) substituents bind in a staggered configuration with high specificity over a range of affinities. Although binding constants have been reported, there is not a clear understanding of why such constants vary significantly for polyamide dimers and their respective cognate DNA sequences. By employing computational tools, the following homo-dimer complexes have been addressed in this study: f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3', f-ImPyPy in complex with 5'-d(GAATGCATTC)-3' and f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3'. These complexes were selected based on their 10 to 100-fold differences in binding constants. From this study, it was possible to determine how polyamides anchor themselves within the minor groove of specific DNA sequences. This is done through several interactions that provide stability for specific recognition: (1) Py groups secure themselves between DNA base pairs, (2) lone-pair-Π interactions are formed between DNA deoxyribose O4' and Im groups nearest f, (3) minor groove bases hydrogen bond to Im groups and amides of the polyamide backbone, (4) the f substituents rotate without leaving the minor groove of DNA and with this rotation form specific hydrogen bonds with electron rich sites on the floor of the minor groove, and (5) flexible charged *N*,*N*-dimethylaminoalkyl substituents reside favorably in the minor groove of DNA. Results displayed the greatest amount of interactions and stability for dimer f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3' and the least amount in dimer f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3'. Hence, for cognate

DNA sequences, the relative binding strength of compounds was determined as f-ImPyIm > f-ImPyPy > f-PyPyIm. This force-field-based computational study is in agreement with experimental results and provides a molecular rational for the binding constant values.

Introduction

The antibiotics netropsin and distamycin A, along with synthetic, tricyclic *N*-methylpyrrole (Py) and *N*-methylimidazole (Im) polyamides, bind within the minor groove of cognate DNA sequences with high specificity but with a surprisingly wide range of affinities (10^5 M^{-1} to 10^8 M^{-1}).[1, 2] The binding sequence specificity of these compounds follows a well-defined set of rules that have been established and confirmed *via* experimental techniques.[1, 3-5] Polyamide compounds align anti-parallel within the minor groove, tail-to-head and head-to-tail, to form stacked dimers able to recognize specific base pairs: Py overlapped with Py (Py-Py) recognizes adenine (A) thymine (T) base pairs or TA base pairs, Im overlapped with Im (Im-Im) recognizes guanine (G) cytosine (C) base pairs or CG base pairs, Im overlapped with Py (Im-Py) recognizes GC, and Py overlapped with Im (Py-Im) recognizes CG. These relationships are displayed visually through experimental findings, including X-ray diffraction[5-7] and NMR,[7, 8] for complexes of Im- and Py-containing polyamides with duplex oligodeoxyribonucleotides. As a result of their ability to recognize specific DNA sequences, polyamides are being developed as potential gene control agents with applications in cancer treatment as well as biotechnology.[9-14]

Experimental studies have uncovered two structural components of polyamide dimers that significantly affect DNA binding affinity: the *N*-terminal formamido group (f) and the combination and order of the pyrrole and imidazole moieties.[1, 2, 6, 8, 15-22] Compared to non-modified polyamides, those with an *N*-terminal f substituent displayed increased affinity for sequence specific binding within the minor groove.[1, 2] This general trend can be illustrated with

results for ImPyPy and f-ImPyPy. The binding affinity of f-ImPyPy (Figure 1) with cognate DNA increased by approximately 10^2 M^{-1} compared to ImPyPy with cognate DNA when a terminal f was present. The association and dissociation rates were slower for the f derivative and the polyamide stacking mode for the complex with f compounds were staggered while the others were overlapped.[2] Similar results have also been observed for other non-modified and f modified polyamides with their respective cognate DNA sequences.[1, 2]

In silico docking and molecular dynamics studies have also provided valuable insight into DNA-polyamide, and other compound, interactions.[23-27] For example, the flexible β-Dp tails of the ImHpPyPy-β-Dp polyamides contributed to binding through water mediated contact with phosphate oxygen.[28] Docking studies have also aided in the construction of DNA-polyamide complexes to examine the movements and interactions of individual bases, such as the roll of base pairs when computationally constructed polyamides were examined in complex with DNA.[24, 29] Experimental thermodynamic data have been examined through limited docking of f-ImPyIm (Figure 1) polyamides.[17]

The unanswered fundamental question in the experimental studies is why the combination and arrangement of Py and Im moieties have such a significantly different effect on binding affinity with cognate DNAs.[1, 18] Given these observed differences, analyzing experimental binding constants with regard to molecular structure interactions of DNA-polyamide complexes can provide valuable insight. The goal of this study is to use in-depth docking methods to compare and examine how polyamides interact with DNA structure and

groups in the minor groove. Through the use of computational tools, several complexes were examined: f-PyPyIm in complex with cognate sequence 5'-d(GAACTAGTTC)-3', f-ImPyPy in complex with cognate sequence 5'-d(GAATGCATTC)-3', and f-ImPyIm in complex with cognate sequence 5'-d(GAACGCGTTC)-3'. These complexes have experimentally determined binding constants of 1×10^6, 1×10^7, and 2×10^8 M^{-1}, respectively.[1, 18] This study represents the first in-depth docking approach to examine these polyamides and their cognate sequences to address the experimental affinity variations. Significant differences were found between the strongest and weakest binding polyamide dimers.

Experimental Procedures

Polyamides f-PyPyIm, f-ImPyPy, and f-ImPyIm were previously synthesized and examined in complex with 5'-d(GAACTAGTTC)-3', 5'-d(GAATGCATTC)-3', and 5'-d(GAACGCGTTC)-3', respectively. Surface plasmon resonance (SPR) was employed to determine binding constants.[1, 18]

Docking Preparation. The three polyamide-DNA complexes were evaluated by employing SYBYL 8.1[30] software on a Fedora Core 5 Linux Workstation. Solution nuclear magnetic resonance (NMR) structure 1B0S[31] was obtained from the protein data bank; this structure was used as a template and as a reference complex. 1B0S, an f-ImImIm dimer in complex with 5'-d(GAACCGGTTC)-3', was mutated using the Biopolymer and Building and Editing modules in SYBYL to form: the f-PyPyIm dimer in complex with 5'-

d(GAACTAGTTC)-3', the f-ImPyPy dimer in complex with 5'-d(GAATGCATTC)-3', and the f-ImPyIm dimer in complex with 5'-d(GAACGCGTTC)-3'. These modified complexes were then minimized for 100 iterations using the Tripos force field; thus, allowing the somewhat rigid DNA to accommodate the mutated bases and polyamides through slight changes to the width of the minor groove. Polyamide dimers were then moved to second memory locations, separate molecular areas within the SYBYL graphical user interface. The ability to move, or rather extract, the compounds into a separate molecular area allowed the compounds to explore torsional angles, translation, and rotational angles independent of the DNA when the FlexiDock genetic algorithm was employed. The two compounds of the dimer were given torsional, translational, and rotational freedom independent of each other; yet, they were docked simultaneously into the DNA. The structures of the polyamides are displayed in Figure 1.

The FlexiDock module of the SYBYL software suite was then implemented. Ten different random starting locations were assigned and employed by the genetic algorithm one at a time for a total of ten docking trials. Calculated and assigned as in previous studies, the large amount of generations ensured that lowest energy conformations were obtained.[32, 33] Each docking trial consisted of 516 000 generations. The dimers and the DNA were permitted torsional, rotational, and translational flexibility throughout the docking process. Atomic charges for the DNA were calculated using the Kollman All-Atom protocol, while the dimer was assigned Gasteiger-Huckel charges. All possible hydrogen bonding sites on the dimer and cognate DNA were targeted for function where possible. From each docking the 20 lowest

energy structures were selected. Hence, 200 structures were produced for each complex examined; 10 random starting locations × 20 low energy structures from each docking. The energy values (E_{MM}) for the overall lowest energy complexes are displayed in Table 1.

Docked Structure Analyses. Interactions were calculated and viewed using modules and tools of the SYBYL software package. The FlexiDock module optimizes torsional angles, translation, and rotational angles to minimize the energy function. Compounds were examined using the FlexiDock scoring function, which is based on the Tripos force field and estimates the energy for the dimer, the receptor, and the complex. The score is evaluated with van der Waals, electrostatic, torsional, and hydrogen bonding energies; lower energy in the complex state suggests better binding. Hydrogen bond distances were analyzed using the "Display H-Bonds" and "Measure" options. Values obtained were averaged for the 20 lowest energy conformations of each complex. All measurements were from heavy atom to heavy atom. The 40 lowest energy complex conformations displayed variations of the f group. The Advanced Computation and Dock modules, of the SYBYL software suite, were employed to gain further explanation into the Im and Py similarities and differences with respect to electrostatics and dipoles.

Grid Search, an application of the Advanced Computation module, was used to examine the f groups, of the lowest energy conformation of each complex, in Cartesian space through systematic rotation about bonds using defined increments of 20^{O} for a total of 360^{O}. At each increment the torsional bond angle was constrained and the conformation was minimized. The minimization of complexes employed default parameters.[30] This allowed for a systematic

exploration of torsional freedoms with regard to respective energies. These complexes were then arranged by f group rotation in increments of 20^{O} and then averaged so that the general trend of the energies could be viewed, analyzed, and compared effectively.

The Dock module of SYBYL 7.3[34] was then employed to re-examine structures and calculate energy values for low energy complexes obtained *via* FlexiDock and Grid Search. This software was ideal since structures could be observed and energy values could be calculated without changes to DNA-polyamide complex conformations.

Accessible solvent area (ASA) was examined for each lowest energy complex with Chimera software.[35] The module employed was the Area/Volume from Web (StrucTools server) with calculation options Gerstein ASA, surface probe 1.4 and all atoms except water. ASA was acquired for each complex, DNA, and individual polyamide. The ASA was summed for each atom of each residue and polyamide.

The Spartan '04[36] software package was employed to examine geometry optimized Py and Im groups using a single point *ab initio* calculation employing the Hartree-Fock 6-31G** level. This allowed for comparisons of *ab initio* calculated electrostatic potential maps and dipoles.

Results

The modification of NMR solution structure 1B0S resulted in the construction of the f-PyPyIm dimer in complex with 5'-d(GAACTAGTTC)-3', the f-ImPyPy dimer in complex with 5'-d(GAATGCATTC)-3', and the f-ImPyIm dimer in complex with 5'-d(GAACGCGTTC)-3'. Each of the polyamides underwent extensive docking, as did reference polyamides from 1B0S, within their respective DNAs to yield optimized structures.

Reference complex, 1B0S, displayed only slight deviations from the refined average structure obtained *via* solution NMR. The calculated root mean squared error between the NMR structure and the lowest energy docked structure was approximately 0.60. Figure 2 displays the alignment of the 10 lowest energy 1B0S-docked reference structures.

f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3'. The low energy structures of f-PyPyIm are hydrogen bonded as a staggered dimer to 5'-d(GAACTAGTTC)-3' (Figure 3). The base pairs involved in hydrogen bonding include those within the center of the DNA, 5'-d(ACTAGT)-3', the AT base pair followed by CG, TA, AT, GC, and TA, respectively. Each image of Figure 2 (Right) was taken as the DNA was rotated to the right, so that the bases would take on a view that was as linear as possible.

The hydrogen bond displayed in the top AT base pair is between the upper dimer compound amide NH of the charged polyamide tail and the C2 O of the T base. Both of the hydrogen bonds displayed in the following image of CG also exist between the upper compound

of the dimer and the G base subsequent to the T on the same DNA chain. The hydrogen bonds are between the Im N and the G C2 NH_2 and between the following amide NH and the G N3. Because of the staggered stacking, the TA base pair represents the first image in which heterocycles from both compounds of the dimer are present and both are hydrogen bonding. In this image the upper dimer compound amide NH following the Py is forming a hydrogen bond with the A N3, while the lower compound of the dimer is forming a hydrogen bond between the amide NH above the Py and the T C2 O. The image of AT and the dimer compounds also displays both compounds forming hydrogen bonds to respective DNA bases. The upper compound amide NH is hydrogen bonded to the T C2 O, while the lower compound amide NH is hydrogen bonded to the A N3. The following image displays base pair GC which hydrogen bonds with the lower dimer compound. The amide NH prior to the Im is hydrogen bonded to G N3 and the Im N is hydrogen bonded to the G C2 NH_2. The TA base pair, of the last image, is similar to the first base pair AT. In this region the NH following the Im of the lower dimer compound hydrogen bonds to the T C2 O.

f-ImPyPy in complex with 5'-d(GAA<u>TGCA</u>TTC)-3'. Docking results display f-ImPyPy to be hydrogen bonded more favorably to 5'-d(GAA<u>TGCA</u>TTC)-3' (Figure 4) than the f-PyPyIm dimer described above. Further stabilization is obtained from lone-pair-Π interactions between DNA deoxyribose O4' and the Im groups of the dimer polyamides. Similar to the compounds observed in Figure 3, f-ImPyPy is bound in a staggered dimer conformation and the base pairs involved in hydrogen bonding include those within the center of the DNA, in this case 5'-

d(ATGCAT)-3'. The images of Figure 4 (Right) show the AT base pair followed by TA, GC, CG, AT, and TA.

The hydrogen bonds displayed in the top AT base pair region are between the upper compound amide NH of the charged polyamide tail and the C2 O of the T base and between the upper compound amide NH and the lower compound f O. The hydrogen bond displayed in the following image of TA exists between the upper compound amide NH subsequent to the Py and the A N3. The GC base pair represents the first image in which both compounds of the dimer are visible and both are partaking in hydrogen bonding. In this image the upper dimer compound amide NH following the Py is forming a hydrogen bond with the C C2 O, while the lower compound of the dimer is forming hydrogen bonds between the f amide NH and the G N3 and the Im N and the G C2 NH_2. The image of CG and the dimer compounds also displays both compounds forming hydrogen bonds to respective DNA bases. The upper compound Im N is hydrogen bonded to the G C2 NH_2 and the following amide NH is hydrogen bonded to the G N3. The lower compound amide NH is hydrogen bonded to the C C2 O. The following image displays base pair AT. Hydrogen bonds displayed in this base pair region exist between the compounds of the dimer and between the lower compound amide NH, above the Py, and the A N3. The last image, of the TA base pair, displays a single hydrogen bond between an amide NH and the T C2 O.

f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3'. Docking results display the f-ImPyIm dimer to be hydrogen bonded to 5'-d(GAACGCGTTC)-3' (Figure 5) even more

favorably than either of the previous dimers to their cognate sequences. Similar to what was seen in Figure 4, Figure 5 shows stabilization in lone-pair-Π interactions between DNA deoxyribose O4' and the Im group nearest the f of the polyamides. As seen in Figures 3 and 4, these compounds take on a staggered dimer conformation and the base pairs involved in hydrogen bonding include those within the center of the DNA, in this case 5'-d(A<u>TGCA</u>T)-3'. As displayed in Figure 3 and 4, the images of Figure 5 (Right) show the AT base pair followed by TA, CG, GC, CG, GC, and TA.

The hydrogen bond displayed in the top AT base pair region is between the upper compound polyamide charged tail amide NH and the C2 O of the T base. Four hydrogen bonds exist in the following image of CG. These hydrogen bonds are between the upper compound amide NH and lower compound f O, between the upper compound Im and the G C2 NH_2, between the amide NH subsequent to the Im of the upper compound to the G N3 and between the f O of the lower compound and the G C2 NH_2. The GC base pair represents the first image in which both compounds of the dimer are present and both are partaking in hydrogen bonding to both strands of the DNA. In this image the upper dimer compound amide NH following the Py is forming a hydrogen bond with the C C2 O, while the lower compound of the dimer is forming hydrogen bonds between the amide NH and the G N3 and the Im N and the G C2 NH_2. The image of CG and the dimer compounds also displays both compounds forming hydrogen bonds to respective DNA bases. The upper compound Im N is hydrogen bonded to the G C2 NH_2 and the following amide NH is hydrogen bonded to the G N3. The lower compound amide NH is

hydrogen bonded to the C C2 O. The following image displays base pair GC. Similar to the first CG region, four hydrogen bonds are displayed in this base pair region. These hydrogen bonds are between the upper compound f O and lower compound amide NH, between the lower compound Im N and the G C2 NH_2, between the amide NH prior to the Im of the upper compound to the G N3 and between the f O of the upper compound and the G C2 NH_2. The last image, of the TA base pair, displays a single hydrogen bond between an amide NH and the T C2 O.

Terminal Interactions. Overall docking results for all three complexes display significant flexibility in the polyamide charged tails, as expected, and significant rotation of the f substituent, which provides unexpected extra insights, while the heterocycles and amide groups exhibit less flexibility. As noted above, the charged tail amide NH forms hydrogen bonds to T C2 O base pairs (Figures 3, 4, and 5). The remainder of each polyamide charged tail resides favorably within respective cognate DNA minor grooves. Rotation of the f group allows for different hydrogen bonds to form and stabilize the complex. Figure 6 displays such changes within overlaid lowest energy structures of f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3'. The lowest energy structure, previously addressed in Figure 5, is shown as caped sticks, while a second low energy conformation is displayed as ball and stick structures. The f NH of the most common lowest energy structures obtained from docking displays hydrogen bonds to G N3 of the first GC base pair of the recognition sequence (Figures 5 and 6, Upper Right); however, upon rotation of f this interaction is lost and new hydrogen bonds are formed between the f O and the G C2 NH_2 of the first CG base pair and the G C2 NH_2 of the first GC base pair (Figure 6, Lower

Right). When this rotation occurs the hydrogen bond between the dimer polyamides, f O and charged tail NH, can no longer form. The rotation of f can occur at either end of the dimer formation. These formations, obtained *via* FlexiDock, were analyzed further through Grid Search. Figure 7 displays the energy fluctuations as rotation occurs in the f group. In the plot, the lowest energy complex conformations, with respect to f torsional angles, are for the two hydrogen bonded conformations in Figure 6.

Complex Energies. Relative total energies for lowest energy complexes subsequent to FlexiDock and Grid Search, calculated *via* Dock, are reported in Table 2. Total energies are the sum of steric and electrostatic energies. For complexes from FlexiDock, the steric and electrostatic energies are respectively, -23.5 and -56.6 kcal/mol for the complex with dimer f-PyPyIm; -73.9 and -44.9 kcal/mol for the complex with f-ImPyPy; and -77.9 and -52.4 kcal/mol for the complex with dimer f-ImPyIm. For each complex acquired *via* Grid Search the steric and electrostatic energies are respectively, -75.3 and -28.0 kcal/mol for the complex with dimer f-PyPyIm, -75.3 and -42.6 kcal/mol for the complex with f-ImPyPy, and -72.7 and -53.3 kcal/mol for the complex with dimer f-ImPyIm. Since the 180° rotation of f occurred in the top twenty percent of lowest energy complexes, as viewed in Figure 6, energies for these complexes were also calculated subsequent to Grid Search (Table 2). For each complex acquired steric and electrostatic energies are respectively, -72.6 and -26.8 kcal/mol for the complex with dimer f-PyPyIm, -72.4 and -44.8 kcal/mol for the complex with f-ImPyPy, and -73.6 and -52.6 kcal/mol for the complex with dimer f-ImPyIm.

Accessible Surface Area (ASA). Buried surface on complex formation was addressed through ASA calculations. Figure 8 displays the surfaces for the complexes, DNA and polyamides; blue surfaces encompass positively charged regions, while red cover those that are negatively charged. Positive and negative regions of the polyamides can align with those of the DNA minor groove to maximize electrostatic interactions. This was further supported with the ASA values for the complexes, DNA and polyamide (Figure 9). The red area represents the DNA. Notice that the DNA ASA is fairly similar for all three complexes, as are the three polyamide areas displayed in green. The blue areas show differences related to the respective DNAs binding their specific polyamides for complex formation: (1) the complex with f-PyPyIm displays more ASA at the T bases of the recognition sequence than the other two complexes, (2) the complex with f-ImPyPy displays a decreased ASA near the A bases of the recognition sequence, and (3) the complex with f-ImPyIm is the most uniform and consists of the most buried ASA.

Ab Initio Electrostatic Potential Maps. To understand the energy contributions from polyamide structures, it is informative to compare the *ab initio* calculated electrostatic potential maps for the Py, Im, and amide units of the polyamide dimers (Figure 10) with each other as well as with the low energy stacked complexes shown in Figure 11. Although the dipole moments of the Py and Im heterocycles point in the same direction, the magnitude of the dipole is larger for the Im and the electrostatic potential maps clearly show a significantly different distribution of molecular electrostatic potential. With both the Py and Im the positive potential is distributed on

the N-Me group and close vicinity. With the Py the highest negative potential is on the Py Π-system while in the Im, it is on the unprotonated Im-N (Figure 10). As expected, the negative potential on the amide is highest on the carbonyl O while the positive potential is on the -NH. With this distribution, the dipole moment of the amide points in the opposite direction to the heterocycles (Figure 10) in the orientation of DNA binding (Figure 11).

Each of the stacked polyamides has six heterocycles that can be evaluated in terms of the maps in Figure 10. Starting with the weakest binder, f-PyPyIm (Figure 11), the heterocycles interact as follows: At top of the Figure, (1) the first Im is relatively unstacked; (2) the next heterocycle (lower molecule of the dimer) is stacked favorably with a positive amide -NH over the negative area of the pyrrole; (3) the next two pyrroles are stacked such that their positive regions are near negative carbonyl O atoms, a fairly favorable orientation; (4) The next Py has its most positive region closely stacked with a positive -NH, an unfavorable interaction; and (5) the last Im is not well stacked. In the strongest binding f-ImPyIm complex (Figure 11), (1) the first Im is not strongly stacked; (2) the next Im (lower molecule) is favorably stacked with an amide with the negative carbonyl O of the amide near the most positive region of the Im (Figure 10) and the amide positive H of the -NH stacked near the negative Im-N; (3) the next two Py groups are also favorable stacked with amide negative O atoms near their positive N-Me groups; (4) the next Im is in a similar favorable orientation with an amide O near the positive N-Me of the imidazole while the positive H of the amide -NH is near the Im negative N; and (5) the last Im is not as strongly stacked as the internal heterocycles. It should be noted, however, that the terminal

two Im groups are stacked with their most negative regions near the most positive regions of the adjacent Ims and, given the larger dipole moment of the Im *versus* Py groups (Figure 10), this should be a favorable contribution. In summary, the electrostatic interactions between the stacked heterocycles appear to make favorable contributions to dimer binding of both f-PyPyIm and f-ImPyIm but there are more and stronger favorable interactions in the f-ImPyIm dimer.

Discussion

Experimental results for simple tricyclic polyamides, such as those in Figure 1, have a puzzling, large variation in energies when bound by their cognate DNA sequences.[1, 2] We have conducted a docking study for the polyamides of Figure 1 and respective cognate DNA to provide some initial molecular level information on the different complexes. Three components that contribute to polyamide dimer-DNA interactions were investigated in the docked structures: (1) hydrogen bonding, (2) buried surface on complex formation, and (3) electrostatic interactions of the polyamide units in the stacked dimer. All of these interactions have been evaluated and these results provide insight into the large variations in binding constants.

When analyzing the dimer of f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3' and its resulting position due to interactions with the minor groove, it is important that the dimer overlap is a staggered conformation of central Py-Py/Py-Py (Figures 3 and 11). The stacked Py groups form a stable motif with the ability to anchor these polyamides into stable dimer conformations in the minor groove (Figures 2). Figure 12 illustrates the models from Figure 3 in two-

dimensions and shows the interactions of the Py-Py stacked motif with the central base pairs of the cognate binding sequence. The Py groups fit between the bases and aid in position indexing so that amide NH groups and the Im N can form favorable hydrogen bonds with base functional groups. The electrostatic potentials also play a significant role in both the specific interactions and complex stabilization (Figures 8 and 9).

The f-ImPyPy dimer polyamides also overlap in a staggered conformation of Py-Im/Im-Py (Figures 4 and 11). As in Figure 3, steric interactions of the stacked Py groups appear to play an important role in anchoring these polyamides within their recognition sequence. Py groups index themselves with steric complementary between the base pairs. The Im groups form favorable hydrogen bonds, due in part to the added stability provided by the steric positioning interactions of the Py groups. The optimum positioning of the Py and Im groups allows the dimer to form hydrogen bonds between the ends of the upper and lower stacked compounds in Figures 4 and 13. Polyamide f-ImPyPy, in complex with 5'-d(GAATGCATTC)-3', displays stacking differences that vary from those of f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3' and these appear to be due to electrostatic interactions between the polyamides and the DNA (Figures 8, 10, and 11). DNA and polyamides were mobile throughout the docking process; the f-ImPyPy polyamides moved into minor groove regions that are more optimal than those of the f-PyPyIm complex (Figures 4 and 13). This allows for more favorable interactions and a larger negative calculated energy, E_{MM} (Table 2).

Similar to f-PyPyIm and f-ImPyPy dimers, f-ImPyIm polyamide dimers overlap in their minor groove location in a staggered conformation of Py-Im/Im-Py (Figures 5 and 14). As seen in Figures 3 and 4, the Py groups index themselves between base pairs with steric complementary and play an important role in anchoring the compound into stable low energy docked conformations with the best possible positioning. The Im groups contribute to the GC base pair recognition and general affinity. The optimum positioning of the Py and Im groups also allows the dimer to form hydrogen bonds between the charged *N*,*N*-dimethylaminoalkyl tail NH of the upper compound and the f O of the lower compounds. The Im groups on the ends also contribute significantly to the amount of hydrogen bonding. This is by far the most stable of the three structures evaluated, as shown by the wealth of hydrogen bonding and the positioning of the compounds. The terminal Im, not involved in the Py-Im/Im-Py stacking, forms tight hydrogen bonds and the compounds are pulled in close to the DNA. These interactions are enhanced by favorable electrostatic interactions that reduce the ASA (Figures 8 and 9). The DNA and compounds form tight favorable interactions and the E_{MM} value for the complex with f-ImPyIm is more negative than the values obtained for complexes with compounds f-PyPyIm or f-ImPyPy.

Given that the compounds of the dimer are binding to the same sequences on opposite DNA strands one may expect the hydrogen bonds to be quite similar. The small differences in observations of individual structures are as expected for flexible docking. The complex with f-PyPyIm does not exhibit hydrogen bonding between the two compounds of the dimer and these

compounds exhibit more mobility within the DNA minor groove (Figures 3 and 12). The hydrogen bond length similarities in compound binding to respective DNA strand are only found in two locations, at the terminal T and the center A of recognition sequence 5'-d(A<u>CT**A**G</u>**T**)-3'. The complex with f-ImPyPy exhibits hydrogen bonding within the dimer, at both ends of the compounds, and this results in a greater amount of consistent hydrogen bond length similarities between compounds and their respective DNA strands (Figures 4 and 12). The hydrogen bond length similarities in compound binding to respective DNA strand are found in three locations central the recognition site, at the G, C, and A of recognition sequence 5'-d(AT<u>**GCA**</u>T)-3'. The complex with f-ImPyIm exhibits hydrogen bonding within the dimer and to the bases of the parallel DNA strand (Figures 5 and 14). The hydrogen bond length similarities in compound binding to respective DNA strands are found throughout the recognition site, at the G, C, G, and T of recognition sequence 5'-d(AC<u>**GCG**</u>**T**)-3'.

The terminal groups of the polyamides are flanked by a flexible charged tail and a small f substituent. The movements of the charged *N*,*N*-dimethylaminoalkyl tail were minor in comparison to a previous molecular dynamics study examining polyamides with longer tails.[28] The charged N resided toward the center of the minor groove between the phosphates of the DNA backbone. Perhaps, possible interactions with phosphates are limited by the size of the tail and the stable hydrogen bonding of the tail NH with T C2 O.

The rotation of the f group in the stacked complexes is a significant observation in these experiments. The terminal f substituents are small enough to rotate in the dimer widened minor

groove without the polyamide leaving the minor groove and energies obtained suggest that the two orientations shown in Figure 6 contribute to binding (Figures 6, 7, and Table 2). The steric contributions to energy are similar, most likely due to the similar size of all three polyamides and the areas occupied; electrostatics differ much more. It is also important to note that stability of structures resulted in consistency of E_{MM} values, even when f rotated 180^{O}. These results suggest that when a single polyamide begins to deviate from its recognition site, a rotation of f occurs and new bonds are formed; thus, keeping the complex longer than if the f substituent was absent. This discovery explains our recent observations that modifications of the f group with other acyl groups results in diminished binding affinity.[22] Specifically, the order of binding constants was f-ImPyIm >> Acetyl-ImPyIm > *N*-methylureidoacetyl-ImPyIm > trifluoroacetyl-ImPyIm. This is consistent with the suggestion that small and planar *N*-terminus subsitituents promote favorable binding with DNA. Furthermore, consistent with the role of the f or acyl group, ImPyIm analogs bearing an NH_2 at the *N*-terminus and non-formamido-ImPyIm gave the weakest binding indicating the importance of having an f group to form favorable hydrogen bonds with sites on the floor in the minor groove.

Previously, experimental studies employing surface plasmon resonance (SPR) acquired binding constants for f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3', f-ImPyPy in complex with 5'-d(GAATGCATTC)-3', and f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3'; these constants are approximately 1×10^6, 1×10^7, and 2×10^8 M^{-1}, respectively.[1, 18] In our studies, E_{MM} values were calculated for the lowest energy complexes obtained *via* FlexiDock and Grid

Search (Table 2); the more negative the value, the stronger the binding. Both the experimental and the *in silico* data are in agreement. The ranked binding from strongest to weakest is: f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3' > f-ImPyPy in complex with 5'-d(GAATGCATTC)-3' > f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3'.

This in-depth docking approach provides useful new molecular information about polyamide complexes and how they are anchored within the minor groove. Hydrogen bonding, steric and electrostatic interactions all play a role, along with compound conformation, to determine how a compound will recognize specific DNA sequences. Specifically, f-ImPyIm binds better than the other dimers as a result of the greater amount of intra-dimer and intra-complex hydrogen bonds, lone-pair-Π interactions, optimum dipole interactions, as well as excellent steric fit and electrostatic interactions. We are currently employing these findings to improve compound design. Findings suggest that dimer spacing provided by Py groups and hydrogen bonding interactions of Im groups can be employed to recognize even longer DNA sequences. This of course is given that recognition compounds: (1) keep a curvature that parallels that of DNA, (2) stack efficiently maintaining electrostatic interactions, and (3) have the ability to form hydrogen bonds on both ends. Insights from this study suggest that compounds such as f-ImPyImPyIm should bind and recognize 5'-d(-ACGCGCGT-)-3 with similar affinity and greater specificity than f-ImPyIm binds and recognizes 5'-d(GAACGCGTTC)-3'. These studies are in progress and the results will be reported in due course.

Acknowledgments

We thank the National Science Foundation (NSF) Grants CHE 0550992 and 0809162 for support (to M. L. and W. D. W.), along with the Georgia State University Molecular Basis of Disease Fellowship (to C. J. C.).

References

1. Buchmueller, K. L.; Staples, A. M.; Uthe, P. B.; Howard, C. M.; Pacheco, K. A.; Cox, K. K.; Henry, J. A.; Bailey, S. L.; Horick, S. M.; Nguyen, B.; Wilson, W. D.; Lee, M., Molecular recognition of DNA base pairs by the formamido/pyrrole and formamido/imidazole pairings in stacked polyamides. *Nucleic. Acids. Res.* **2005,** *33*, (3), 912-921.
2. Lacy, E. R.; Le, N. M.; Price, C. A.; Lee, M.; Wilson, W. D., Influence of a terminal formamido group on the sequence recognition of DNA by polyamides. *J. Am. Chem. Soc.* **2002,** *124*, (10), 2153-2163.
3. White, S.; Baird, E. E.; Dervan, P. B., On the pairing rules for recognition in the minor groove of DNA by pyrrole-imidazole polyamides. *Chem. Biol.* **1997,** *4*, (8), 569-578.
4. White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E.; Dervan, P. B., Recognition of the four Watson-Crick base pairs in the DNA minor groove by synthetic ligands. *Nature* **1998,** *391*, (6666), 468-471.

5. Kielkopf, C. L.; Bremer, R. E.; White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E.; Dervan, P. B.; Rees, D. C., Structural effects of DNA sequence on T.A recognition by hydroxypyrrole/pyrrole pairs in the minor groove. *J. Mol. Biol.* **2000,** *295*, (3), 557-567.
6. Chenoweth, D. M.; Dervan, P. B., Allosteric modulation of DNA by small molecules. *Proc. Natl. Acad. Sci. U S A* **2009,** *106*, (32), 13175-13179.
7. Yang, X. L.; Hubbard, R. B.; Lee, M.; Tao, Z. F.; Sugiyama, H.; Wang, A. H., Imidazole-imidazole pair as a minor groove recognition motif for T:G mismatched base pairs. *Nucleic. Acids. Res.* **1999,** *27*, (21), 4183-4190.
8. Zhang, Q.; Dwyer, T. J.; Tsui, V.; Case, D. A.; Cho, J.; Dervan, P. B.; Wemmer, D. E., NMR structure of a cyclic polyamide-DNA complex. *J. Am. Chem. Soc.* **2004,** *126*, (25), 7958-7966.
9. Hochhauser, D.; Kotecha, M.; O'Hare, C.; Morris, P. J.; Hartley, J. M.; Taherbhai, Z.; Harris, D.; Forni, C.; Mantovani, R.; Lee, M.; Hartley, J. A., Modulation of topoisomerase IIalpha expression by a DNA sequence-specific polyamide. *Mol. Cancer Ther.* **2007,** *6*, (1), 346-354.
10. Chou, C. J.; Farkas, M. E.; Tsai, S. M.; Alvarez, D.; Dervan, P. B.; Gottesfeld, J. M., Small molecules targeting histone H4 as potential therapeutics for chronic myelogenous leukemia. *Mol. Cancer Ther.* **2008,** *7*, (4), 769-778.
11. Lai, Y. M.; Fukuda, N.; Ueno, T.; Matsuda, H.; Saito, S.; Matsumoto, K.; Ayame, H.; Bando, T.; Sugiyama, H.; Mugishima, H.; Serie, K., Synthetic pyrrole-imidazole

polyamide inhibits expression of the human transforming growth factor-beta1 gene. *J Pharmacol. Exp. Ther.* **2005,** *315*, (2), 571-575.

12. Matsuda, H.; Fukuda, N.; Ueno, T.; Tahira, Y.; Ayame, H.; Zhang, W.; Bando, T.; Sugiyama, H.; Saito, S.; Matsumoto, K.; Mugishima, H.; Serie, K., Development of gene silencing pyrrole-imidazole polyamide targeting the TGF-beta1 promoter for treatment of progressive renal diseases. *J. Am. Soc. Nephrol.* **2006,** *17*, (2), 422-432.
13. Olenyuk, B. Z.; Zhang, G. J.; Klco, J. M.; Nickols, N. G.; Kaelin, W. G., Jr.; Dervan, P. B., Inhibition of vascular endothelial growth factor with a sequence-specific hypoxia response element antagonist. *Proc. Natl. Acad. Sci. U S A* **2004,** *101*, (48), 16768-16773.
14. Nickols, N. G.; Jacobs, C. S.; Farkas, M. E.; Dervan, P. B., Improved nuclear localization of DNA-binding polyamides. *Nucleic Acids Res.* **2007,** *35*, (2), 363-370.
15. Bremer, R. E.; Szewczyk, J. W.; Baird, E. E.; Dervan, P. B., Recognition of the DNA minor groove by pyrrole-imidazole polyamides: comparison of desmethyl- and N-methylpyrrole. *Bioorg. Med. Chem.* **2000,** *8*, (8), 1947-1955.
16. Brown, T.; Mackay, H.; Turlington, M.; Sutterfield, A.; Smith, T.; Sielaff, A.; Westrate, L.; Bruce, C.; Kluza, J.; O'Hare, C.; Nguyen, B.; Wilson, W. D.; Hartley, J. A.; Lee, M., Modifying the N-terminus of polyamides: PyImPyIm has improved sequence specificity over f-ImPyIm. *Bioorg. Med. Chem.* **2008,** *16*, (9), 5266-5276.
17. Buchmueller, K. L.; Bailey, S. L.; Matthews, D. A.; Taherbhai, Z. T.; Register, J. K.; Davis, Z. S.; Bruce, C. D.; O'Hare, C.; Hartley, J. A.; Lee, M., Physical and structural

basis for the strong interactions of the -ImPy- central pairing motif in the polyamide f-ImPyIm. *Biochemistry* **2006,** *45*, (45), 13551-13565.

18. Buchmueller, K. L.; Staples, A. M.; Howard, C. M.; Horick, S. M.; Uthe, P. B.; Le, N. M.; Cox, K. K.; Nguyen, B.; Pacheco, K. A.; Wilson, W. D.; Lee, M., Extending the language of DNA molecular recognition by polyamides: unexpected influence of imidazole and pyrrole arrangement on binding affinity and specificity. *J. Am. Chem. Soc.* **2005,** *127*, (2), 742-750.
19. Mackay, H.; Brown, T.; Uthe, P. B.; Westrate, L.; Sielaff, A.; Jones, J.; Lajiness, J. P.; Kluza, J.; O'Hare, C.; Nguyen, B.; Davis, Z.; Bruce, C.; Wilson, W. D.; Hartley, J. A.; Lee, M., Sequence specific and high affinity recognition of 5'-ACGCGT-3' by rationally designed pyrrole-imidazole H-pin polyamides: thermodynamic and structural studies. *Bioorg. Med. Chem.* **2008,** *16*, (20), 9145-9153.
20. Melander, C.; Herman, D. M.; Dervan, P. B., Discrimination of A/T sequences in the minor groove of DNA within a cyclic polyamide motif. *Chemistry* **2000,** *6*, (24), 4487-4497.
21. O'Hare, C. C.; Uthe, P.; Mackay, H.; Blackmon, K.; Jones, J.; Brown, T.; Nguyen, B.; Wilson, W. D.; Lee, M.; Hartley, J. A., Sequence recognition in the minor groove of DNA by covalently linked formamido imidazole-pyrrole-imidazole polyamides: effect of H-pin linkage and linker length on selectivity and affinity. *Biochemistry* **2007,** *46*, (42), 11661-11670.

22. Westrate, L.; Mackay, H.; Brown, T.; Nguyen, B.; Kluza, J.; Wilson, W. D.; Lee, M.; Hartley, J. A., Effects of the N-terminal acylamido group of imidazole- and pyrrole-containing polyamides on DNA sequence specificity and binding affinity. *Biochemistry* **2009,** *48*, (24), 5679-5688.
23. Holt, P. A.; Chaires, J. B.; Trent, J. O., Molecular docking of intercalators and groove-binders to nucleic acids using Autodock and Surflex. *J. Chem. Inf. Model.* **2008,** *48*, (8), 1602-1615.
24. Ge, W.; Schneider, B.; Olson, W. K., Knowledge-based elastic potentials for docking drugs or proteins with nucleic acids. *Biophys. J.* **2005,** *88*, (2), 1166-1190.
25. Kielkopf, C. L.; White, S.; Szewczyk, J. W.; Turner, J. M.; Baird, E. E.; Dervan, P. B.; Rees, D. C., A structural basis for recognition of A.T and T.A base pairs in the minor groove of B-DNA. *Science* **1998,** *282*, (5386), 111-115.
26. Wellenzohn, B.; Loferer, M. J.; Trieb, M.; Rauch, C.; Winger, R. H.; Mayer, E.; Liedl, K. R., Hydration of hydroxypyrrole influences binding of ImHpPyPy-beta-Dp polyamide to DNA. *J. Am. Chem. Soc.* **2003,** *125*, (4), 1088-1095.
27. Cashman, D. J.; Buscaglia, R.; Freyer, M. W.; Dettler, J.; Hurley, L. H.; Lewis, E. A., Molecular modeling and biophysical analysis of the c-MYC NHE-III1 silencer element. *J. Mol. Model.* **2008,** *14*, (2), 93-101.
28. Wellenzohn, B.; Flader, W.; Winger, R. H.; Hallbrucker, A.; Mayer, E.; Liedl, K. R., Complex of B-DNA with polyamides freezes DNA backbone flexibility. *J. Am. Chem. Soc.* **2001,** *123*, (21), 5044-5049.

29. Correa, B. J.; Canzio, D.; Kahane, A. L.; Reddy, P. M.; Bruice, T. C., DNA sequence recognition by Hoechst 33258 conjugates of hairpin pyrrole/imidazole polyamides. *Bioorg. Med. Chem. Lett.* **2006,** *16*, (14), 3745-3750.
30. *SYBYL Molecular Modeling Software, 8.1 ed.*, Tripos Inc.: St. Louis, MO, 2008.
31. Yang, X. L.; Kaenzig, C.; Lee, M.; Wang, A.H., Binding of AR-1-114, a tri-imidazole DNA minor groove binder to CCGG sequence analyzed by NMR spectroscopy. *Eur. J. Biochem.* **1999,** *263*, (3), 646-655.
32. Munde, M.; Ismail, M. A.; Arafa, R.; Peixoto, P.; Collar, C. J.; Liu, Y.; Hu, L.; David-Cordonnier, M. H.; Lansiaux, A.; Bailly, C.; Boykin, D. W.; Wilson, W. D., Design of DNA minor groove binding diamidines that recognize GC base pair sequences: a dimeric-hinge interaction motif. *J. Am. Chem. Soc.* **2007,** *129*, (44), 13732-43.
33. Liu, Y.; Collar, C. J.; Kumar, A.; Stephens, C. E.; Boykin, D. W.; Wilson, W. D., Heterocyclic diamidine interactions at AT base pairs in the DNA minor groove: effects of heterocycle differences, DNA AT sequence and length. *J. Phys. Chem. B.* **2008,** *112*, (37), 11809-18.
34. *SYBYL Molecular Modeling Software, 7.3 ed.*, Tripos Inc.: St. Louis, MO, 2006.
35. *UCSF Chimera*, Regents of the University of California: San Francisco, CA, 2007.
36. *Spartan '04*, Wavefunction Inc.: Irvine, CA, 2004.

Tables and Figures

Table 1. Energies (E_{MM}) gained from FlexiDock docking studies. These values are in kcal/mol.

	E_{MM}
f-PyPyIm with 5'-d(ACTAGT)-3'	-594
f-ImPyPy with 5'-d(ATGCAT)-3'	-670
f-ImPyIm with 5'-d(ACGCGT)-3'	-765

Table 2. Total energies, reported as E_{MM} values, gained from Dock for the lowest energy complexes obtained *via* FlexiDock and Grid Search. The E_{MM} values for both low-energy structures acquired *via* Grid Search are labeled 1 and 2, respectively. Grid Search 1 relates to the structures with the f positioned as in the top right image of Figure 6, while Grid Search 2 relates to the structures with the f positioned as in the bottom right image of Figure 6. All E_{MM} values are in kcal/mol.

	FlexiDock	Grid Search 1	Grid Search 2
f-PyPyIm with 5'-d(ACTAGT)-3'	-80.1	-103	-99.4
f-ImPyPy with 5'-d(ATGCAT)-3'	-119	-118	-117
f-ImPyIm with 5'-d(ACGCGT)-3'	-130	-126	-126

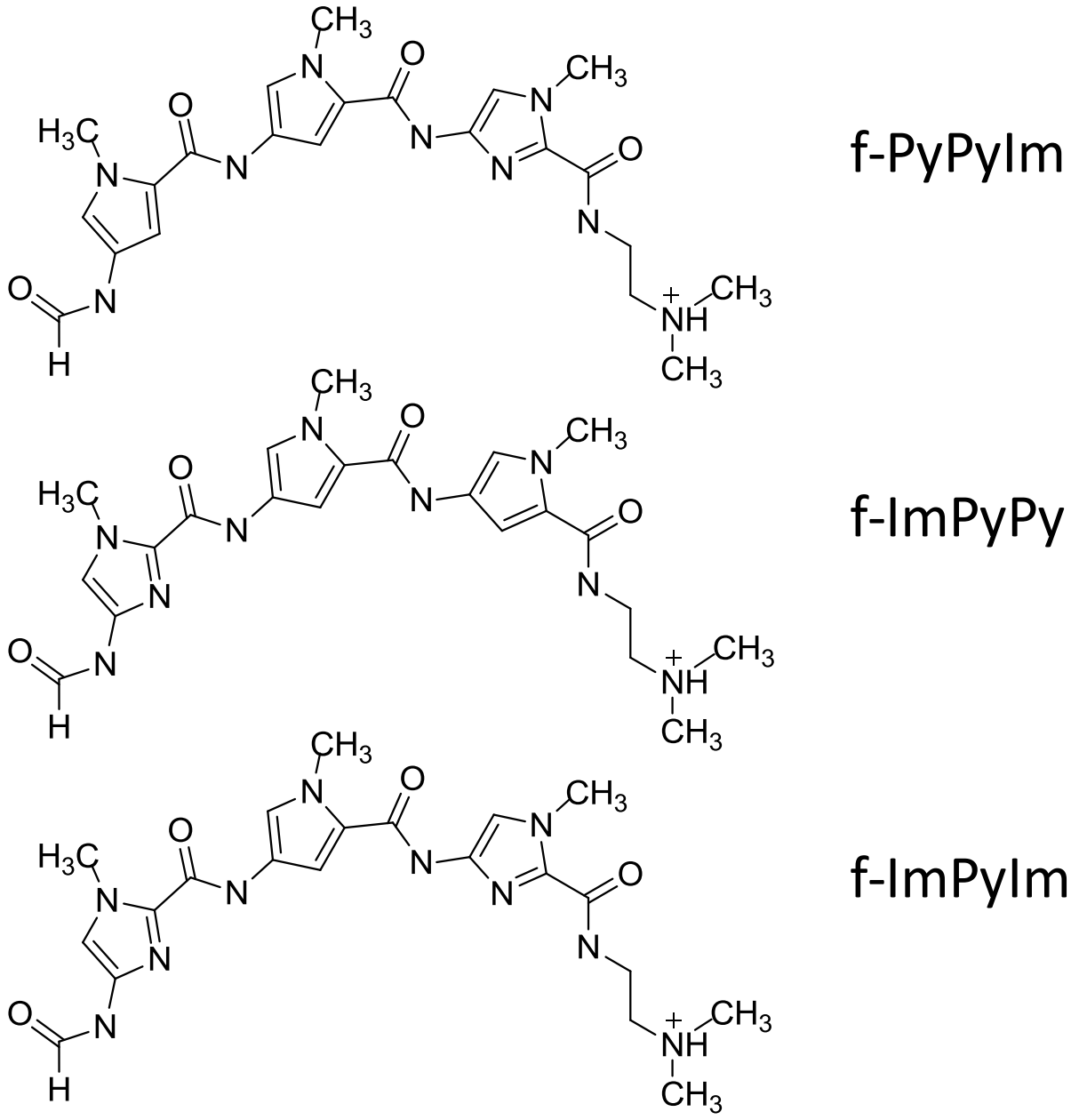


Figure 1. Two-dimensional illustration of polyamide structures (Left) with abbreviations (Right): formamido (f), *N*-methylpyrrole (Py) and *N*-methylimidazole (Im). Dimer complexes of these compounds are shown docked into cognate DNA sequences in Figures 3-5.

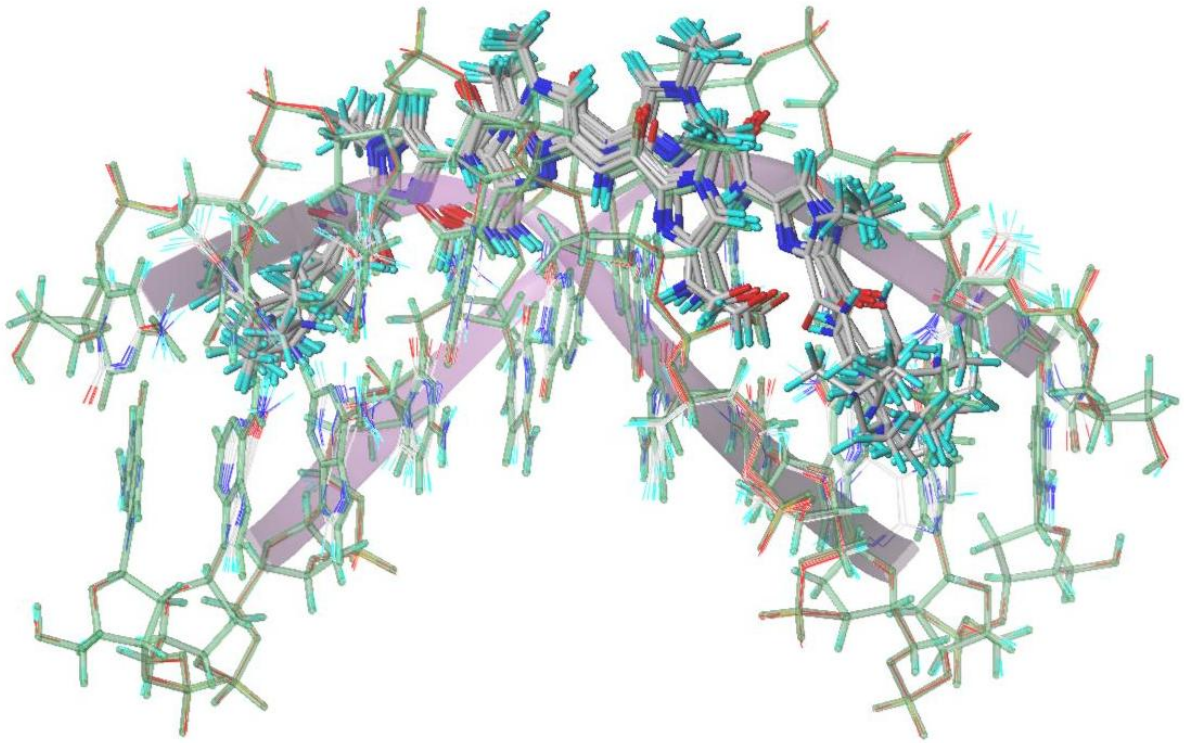


Figure 2. Overlay of the 10 lowest energy structures for the docking of reference structure, 1B0S, polyamides into cognate DNA. The refined average structure obtained from the protein data bank is displayed in green, while all other low energy complexes are displayed by atom type.

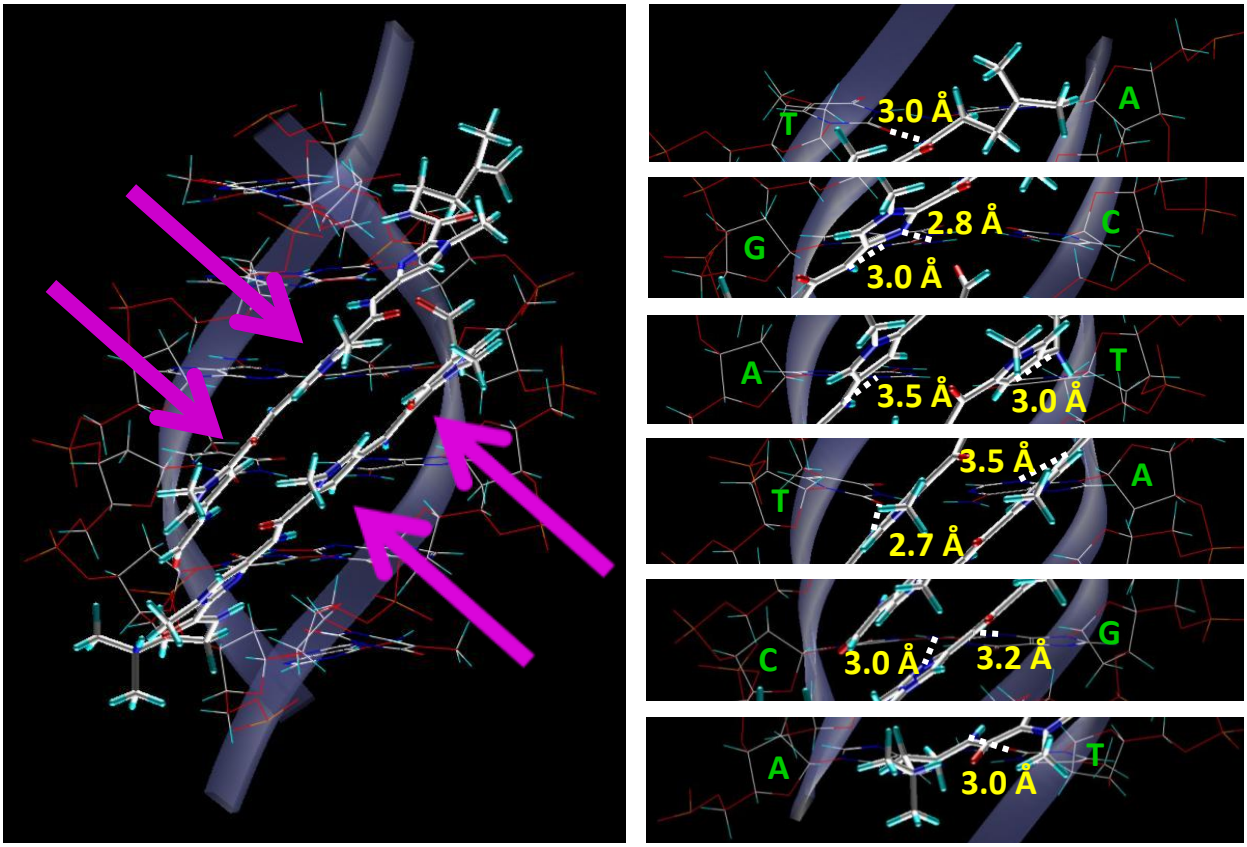


Figure 3. f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3'. For clarity, the terminal bases are not displayed in the images and the images are of only the lowest energy conformation. The image on the left displays the complex as a whole, while the segmented images on the right show the individual bases as the polyamide-DNA complex is rotated to the right. Magenta arrows are displayed on the left image; these point to the Py groups. On the right, the bases are labeled in green, the hydrogen bonds are in white and the average respective hydrogen bond lengths are in yellow. Notice that the Py groups index themselves with steric complementary between base pairs, this is pointed out on the left and more clearly viewed in the central TA and AT images on the right.

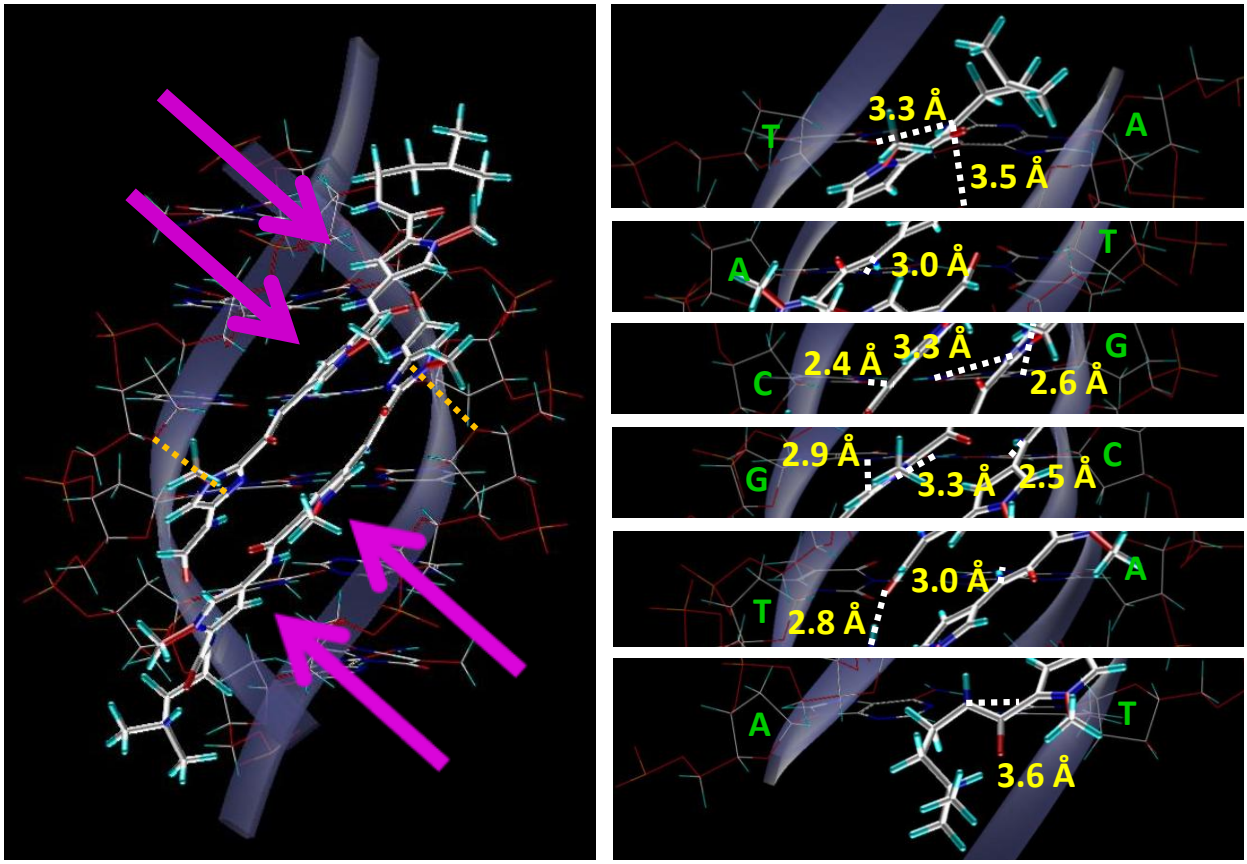


Figure 4. f-ImPyPy in complex with 5'-d(GAATGCATTC)-3'. For clarity, the terminal bases are not displayed in the images and the images are of only the lowest energy conformation. The image on the left displays the complex as a whole and identified lone-pair-Π interactions (orange), while the segmented images on the right show the individual bases as the polyamide-DNA complex is rotated to the right (labeled as in Figure 3).

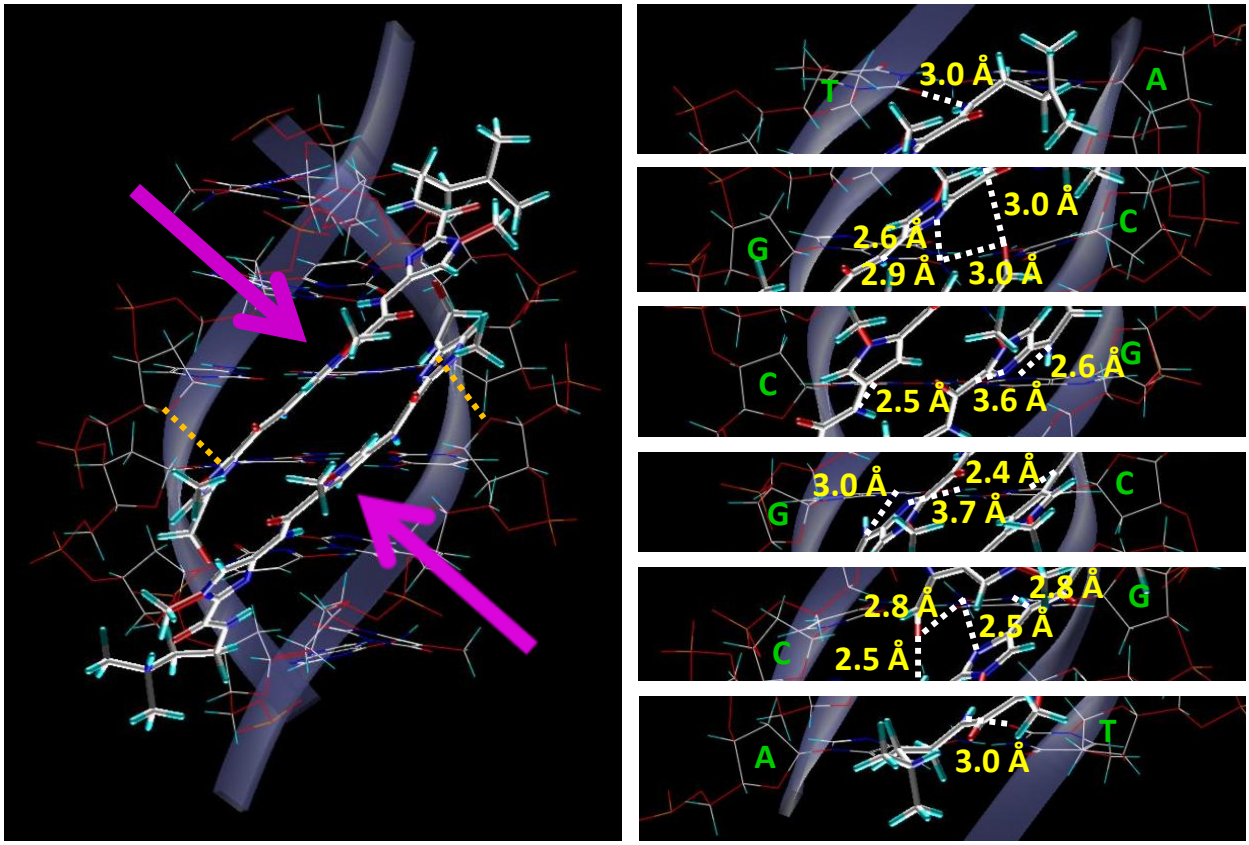


Figure 5. f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3'. For clarity, the terminal bases are not displayed in the images and the images are of only the lowest energy conformation. The image on the left displays the complex as a whole, while the segmented images on the right show the individual bases as the polyamide-DNA complex is rotated to the right (labeled as in Figures 3 and 4).

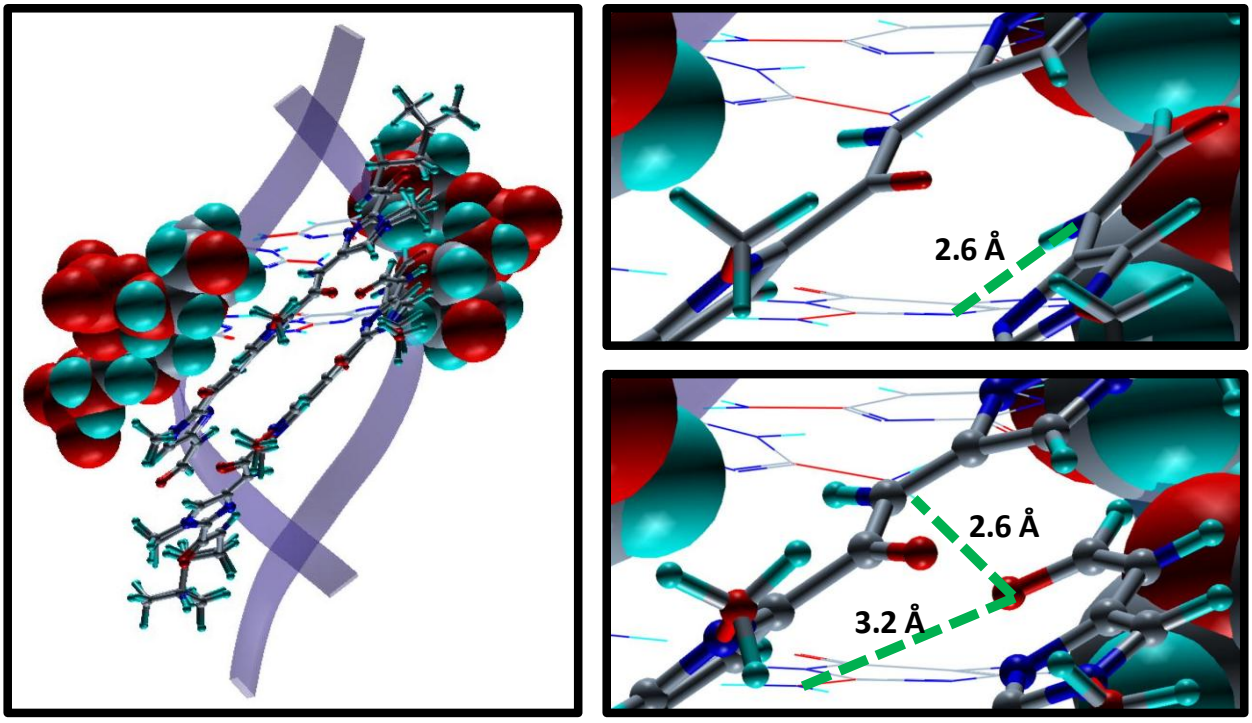


Figure 6. f-ImPyIm in complex with cognate sequence 5'-d(GAACGCGTTC)-3'. The image on the left displays the overlap of the two low energy dimer conformations in the minor groove of the lowest energy DNA base pairs affected by f rotation. In an enlarged view for clarity, the images on the right display the two low energy conformations individually. These conformations consist of different hydrogen bonding interactions, which are shown in green. The upper right image displays the f N hydrogen bonding to the G N3 of the first GC base pair of the recognition sequence; whereas the lower right image displays the hydrogen bonding of f O to the G C2 NH_2 of the first CG base pair, as well as the G C2 NH_2 of the first GC base pair.

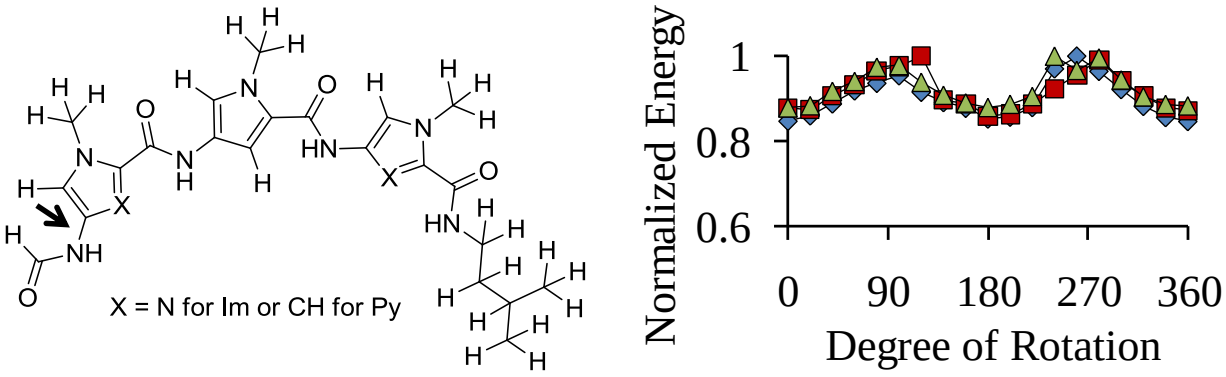


Figure 7. Polyamide structure (Left) with an arrow pointing to the bond rotated *via* Grid Search. X represents *N*-methylpyrrole (Py) and/or *N*-methylimidazole (Im) depending on the complex employed for formamido (f) bond rotation. The graph (Right) displays the averaged, normalized energy values for each structure obtained after a 20^{O} rotation of one or both f bonds within the dimer. Data for complexes with f-PyPyIm, f-ImPyPy, and f-ImPyIm, are displayed in blue, red and green, respectively. The lowest energy conformations are at 0^{O} and 180^{O}. At 0^{O} the f is positioned as in Figure 6, Upper Right, and position 180^{O} is shown in Figure 6, Lower Right.

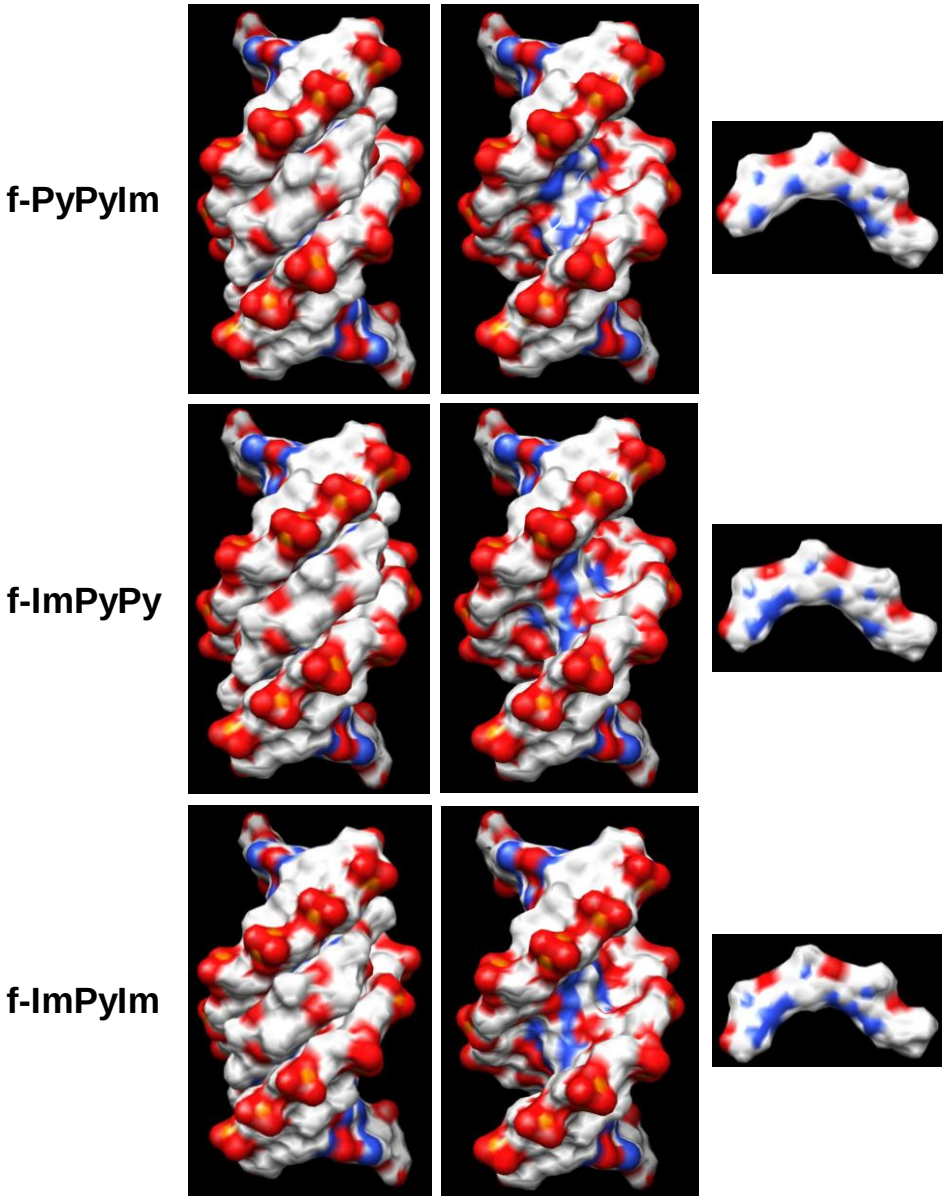


Figure 8. Surfaces displaying electrostatic potentials with respect to coulombic coloring for the complexes (Left), DNA (Center) and polyamides (Right); blue surfaces encompass positively charged regions, while red cover those that are negatively charged.

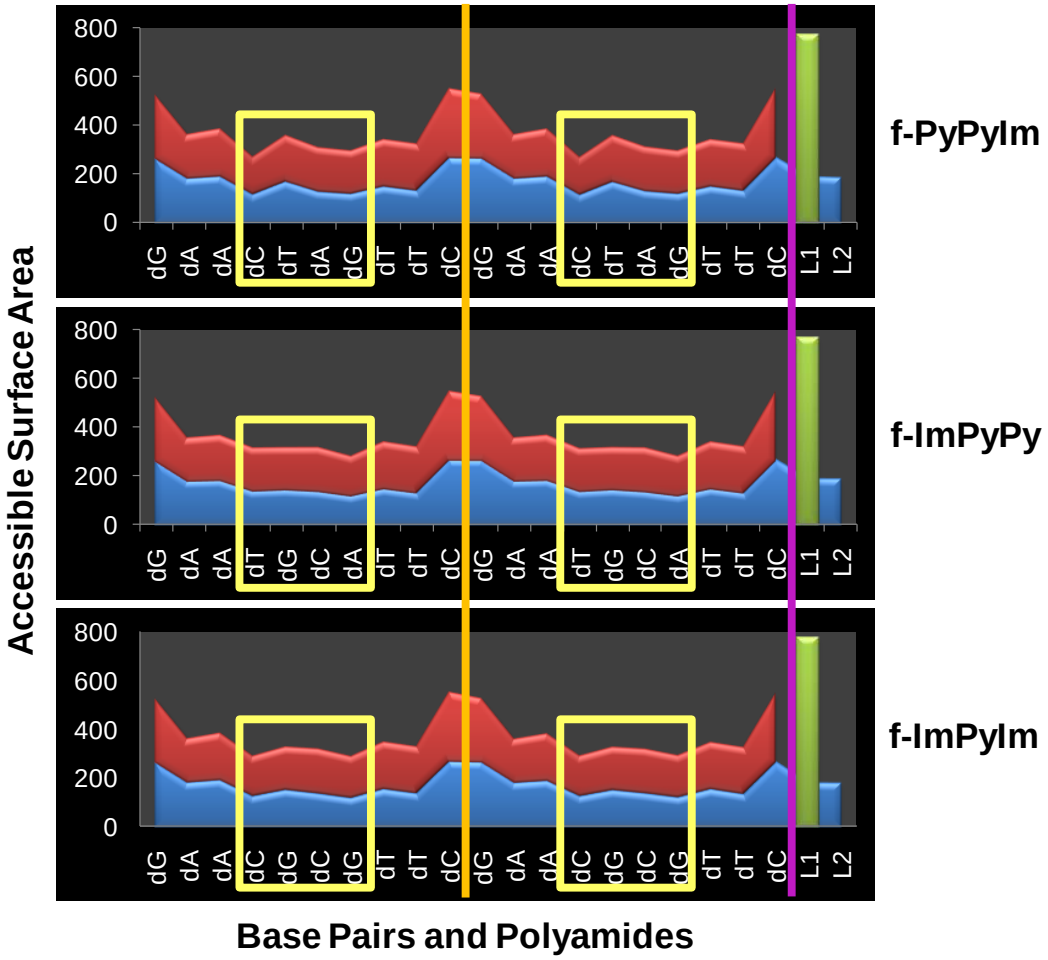


Figure 9. Accessible Surface Area (ASA) calculated for each base pair and polyamide in complex (blue) and alone (red for DNA and green for single polyamide). The ASA is reported in $Å^2$ and the DNA bases are denoted as dA, dT, dG and dC for adenine, thymine, guanine and cytosine, respectively. Orange and purple lines spanning the three ASA graphs separate the two DNA strands and the polyamides. Both DNA strands are shown from 5' to 3', displaying the differences in each strand with (blue) and without (red) compound interaction. Yellow boxes highlight the specific recognition sites for each complex. The polyamides are displayed as L1 and L2. When analyzing only a single polyamide (green), this compound is L1.

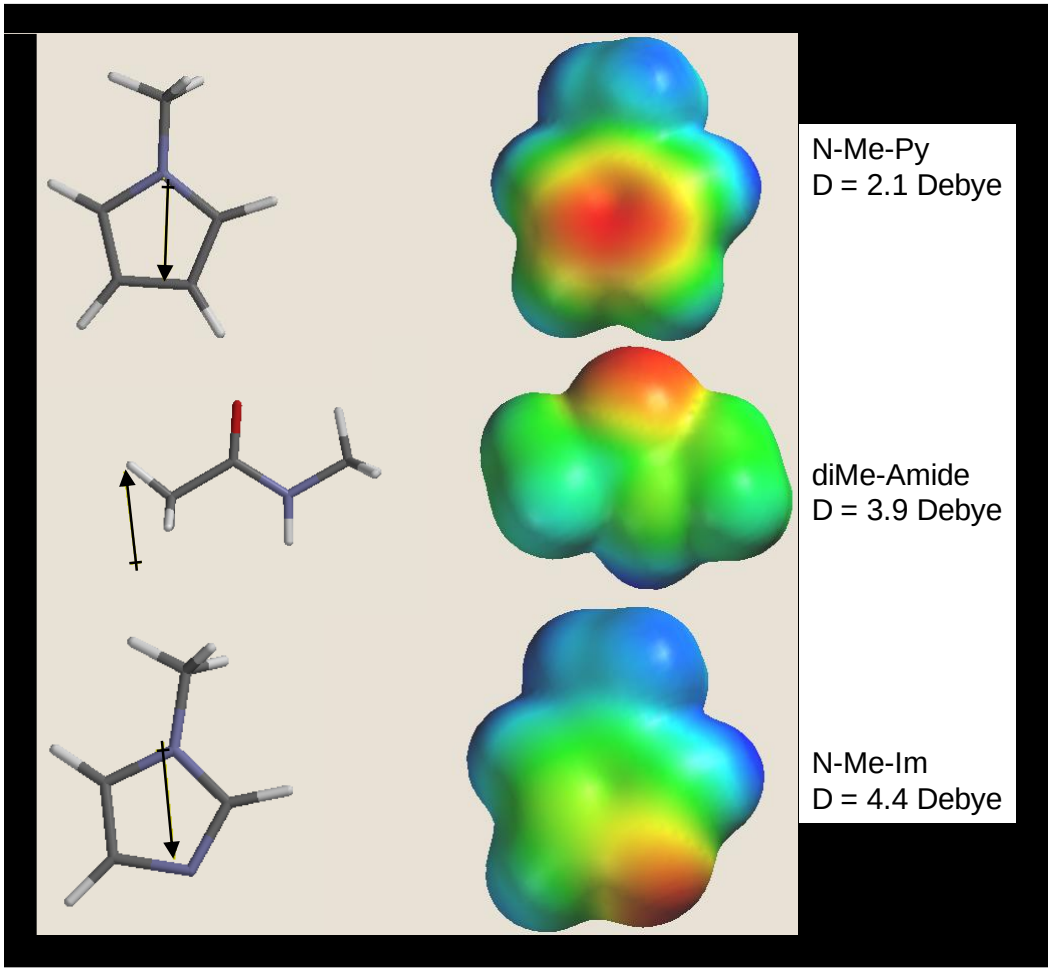


Figure 10. *Ab initio* calculated electrostatic potential maps for the Py, Im and amide units of the polyamide dimers, respectively these units are shown on the left with their dipole moments. The electrostatic potentials are shown in the center, blue is positive and red is negative, and the magnitudes of the dipoles are displayed on the right.

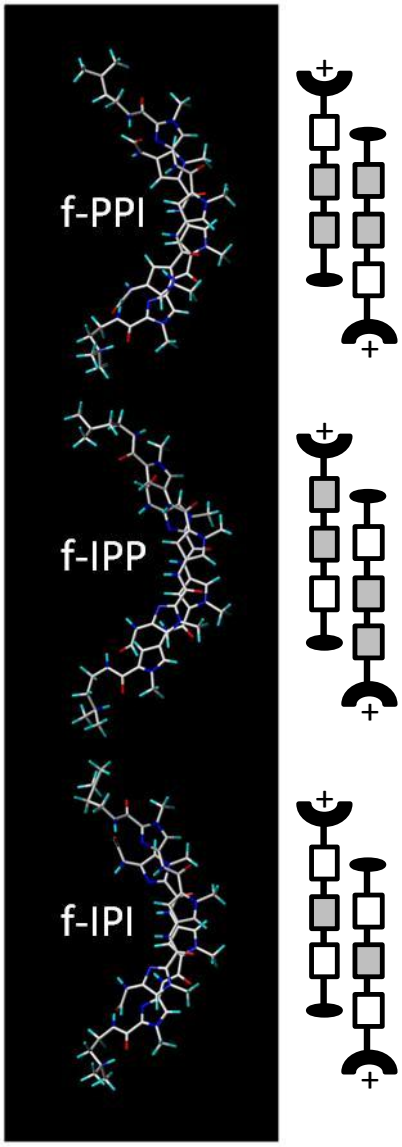


Figure 11. Top view of dimers formed during docking (Left) and schematic representation (Right) with Py in gray and Im in white. The DNA has been removed so that preferred staggered conformations can be viewed. From top to bottom, the dimers come from f-PyPyIm in complex with 5'-d(GAACTAGTTC)-3', f-ImPyPy in complex with 5'-d(GAATGCATTC)-3' and f-ImPyIm in complex with 5'-d(GAACGCGTTC)-3'. Notice the spacing of the dimers.

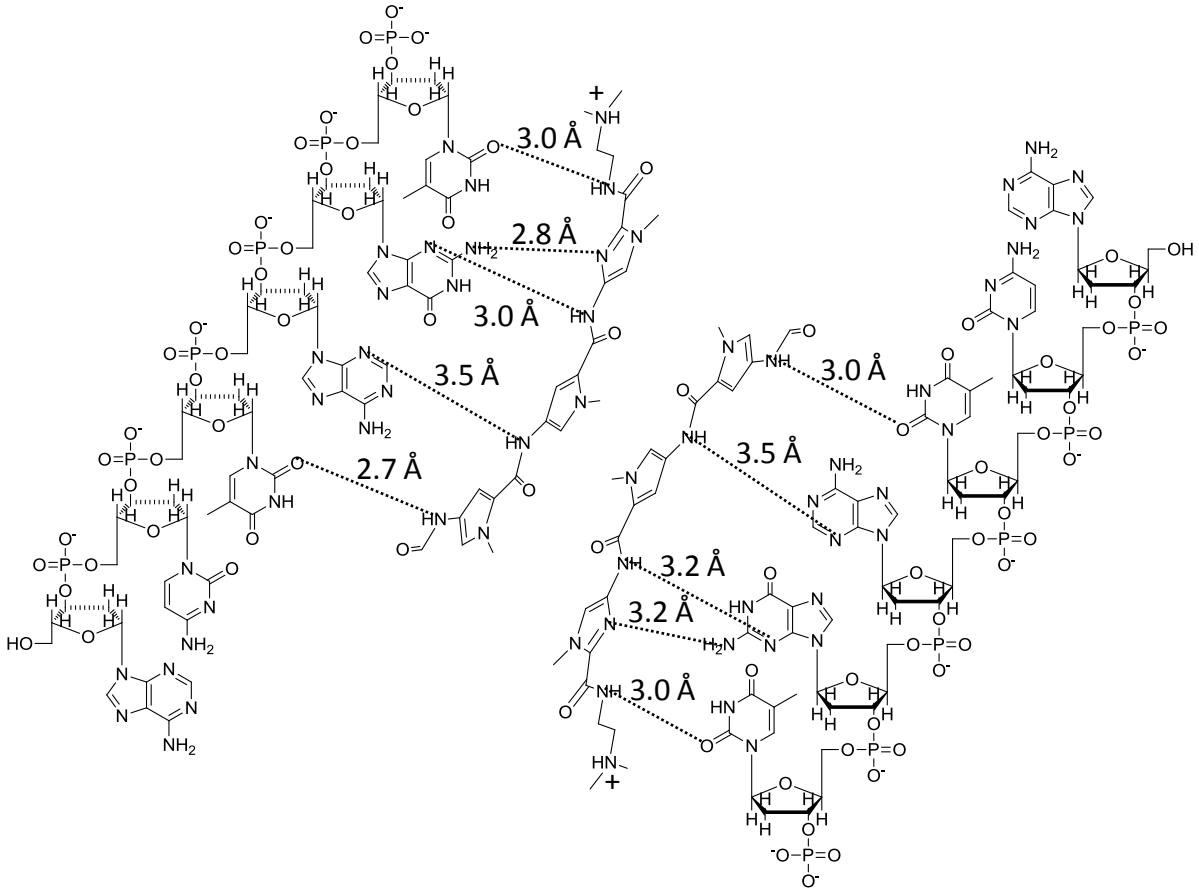


Figure 12. Two-dimensional illustration of f-PyPyIm in complex with cognate sequence 5'-d(GAACTAGTTC)-3'. Hydrogen bonds are displayed by dashed lined with respective distances.

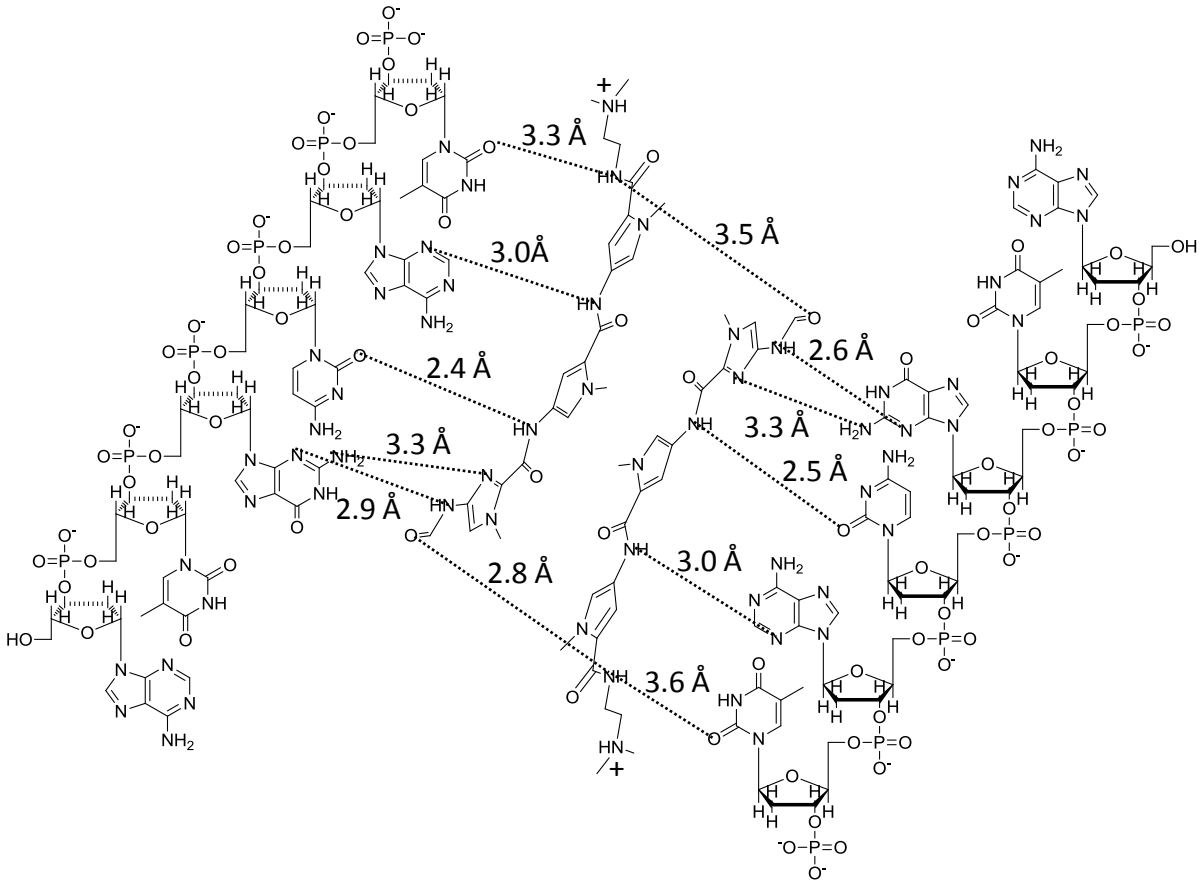


Figure 13. Two-dimensional illustration of f-ImPyPy in complex with cognate sequence 5'-d(GAATGCATTC)-3'. Hydrogen bonds are displayed by dashed lined with respective distances.

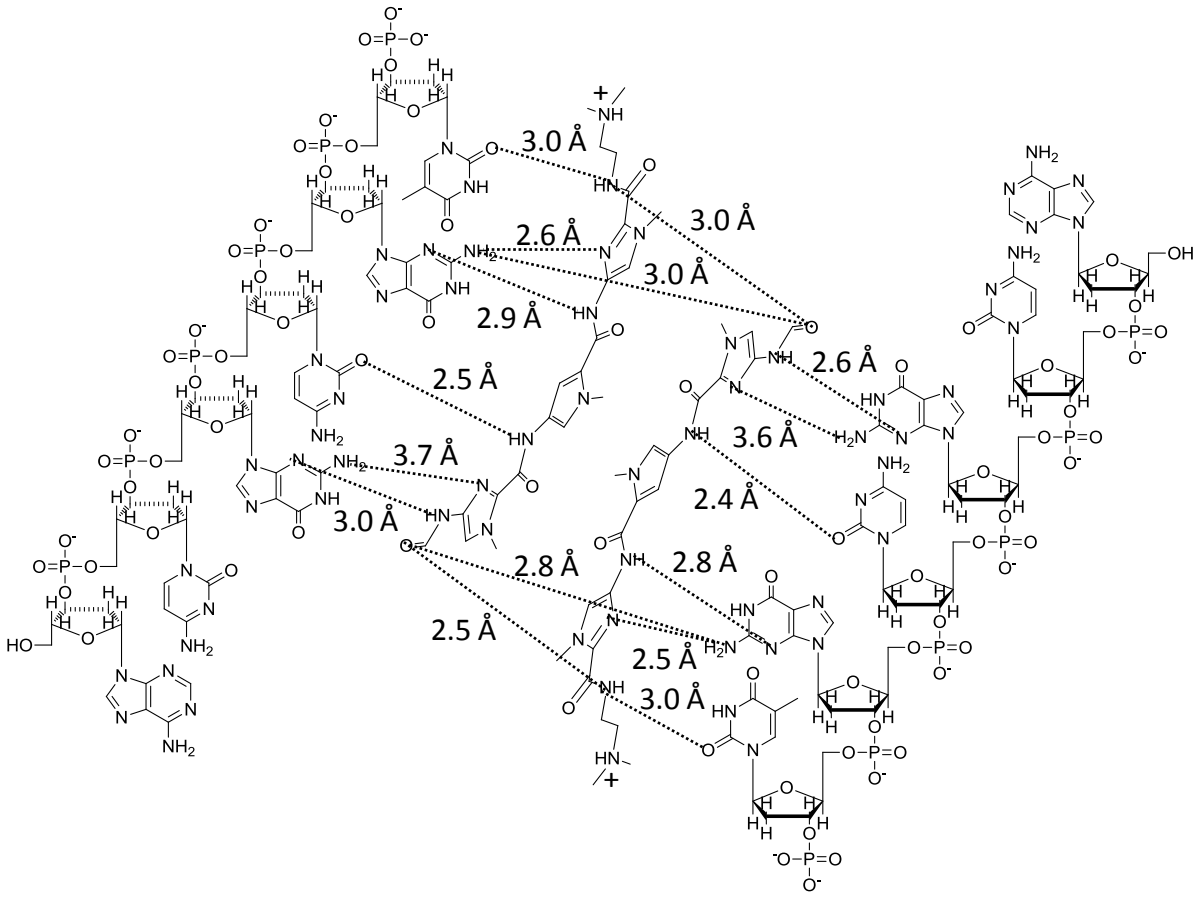
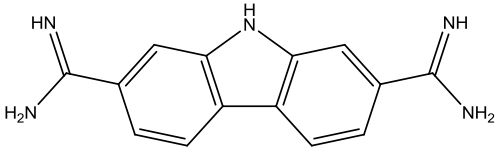


Figure 14. Two-dimensional illustration of f-ImPyIm in complex with cognate sequence 5'-d(GAACGCGTTC)-3'. Hydrogen bonds are displayed by dashed lined with respective distances.

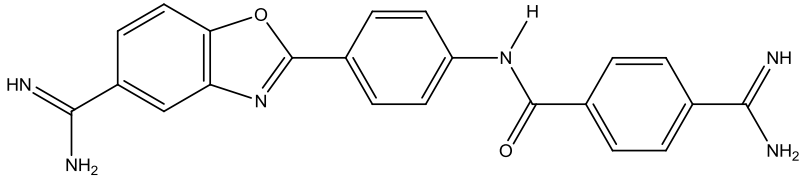
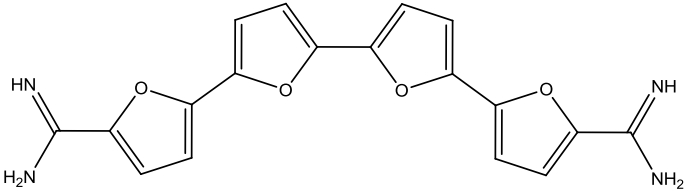
APPENDICES

Appendix A

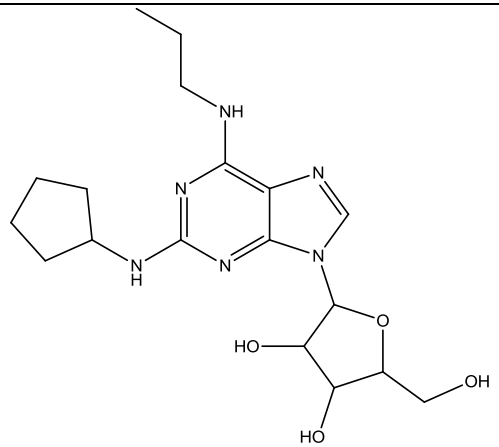
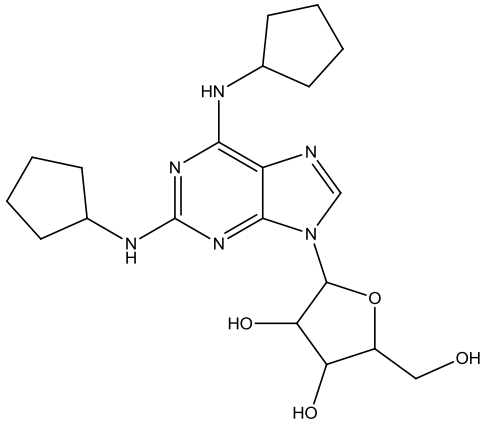
Supplemental Table 1. Compounds employed for training and testing.

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
A	1-deazaadenosine	45.4	-24.8	(1)
A	1-deazapurine	131	-22.2	Aldrich
A	2,6-diaminopurine-2'-d-riboside	4.44	-30.6	MP Biomedicals
A	2-chloro-adenosine	9.65	-28.6	TriLink Biotech
A		7.5	-29.3	(10)
A	2'-deoxyadenosine	0.23	-37.9	Sigma
A	2'-deoxyinosine	165	-21.6	Sigma
A	2-hydroxy-6-aminopurine	9.7	-28.6	Acros
A	2-nitoradenosine	81	-23.4	(1)
A	3-deaza-adenosine	0.29	-37.3	Sigma
A	6-chloropurine riboside	15.4	-27.5	TriLink Biotech

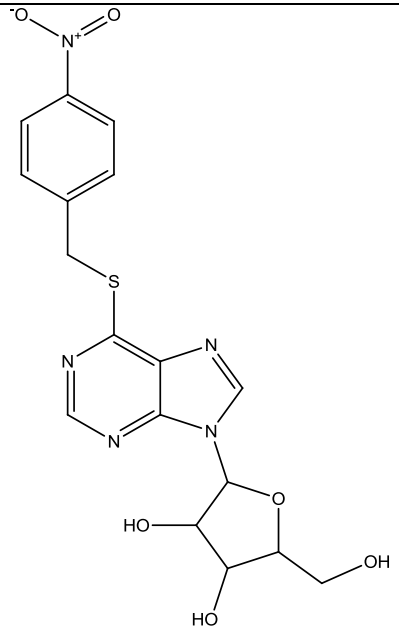
Supplemental Table 1 (continued)

Scaffold	Compound		K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
A	8-azidoadenosine		331	-19.9	(8)
A	8-bromoadenosine		37.8	-25.2	Acros
A	9-deazaadenosine		12.2	-28	(7)
A	adenine		0.3	-37.2	Sigma
A	adenosine		0.92	-34.5	Sigma
A	allopurinol		255	-20.5	Sigma
A	DAPI		0.47	-36.1	Fluka
A	DB1208		0.37	-36.7	(5)
A	DB1464		0.15	-39	(5)
A	Dilazep		150	-21.8	Sigma
A	Dipyridamole		51.6	-24.5	Sigma
A	formycin A		36.5	-25.3	(8)
A	Hypoxanthine		500	-18.8	Sigma

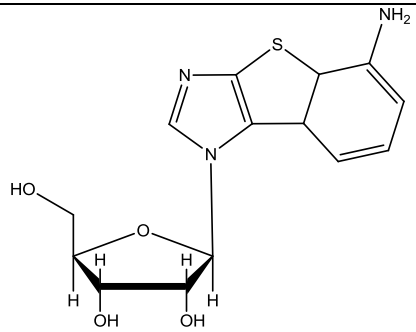
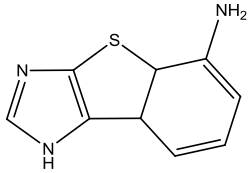
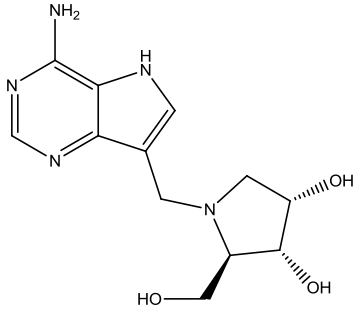
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
A		9	-28.8	(1)
A		19.9	-26.8	(1)

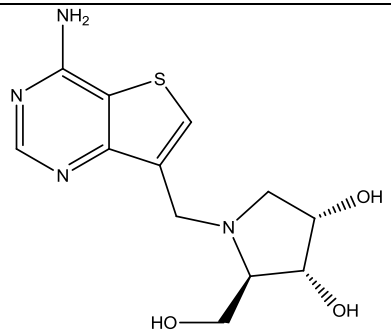
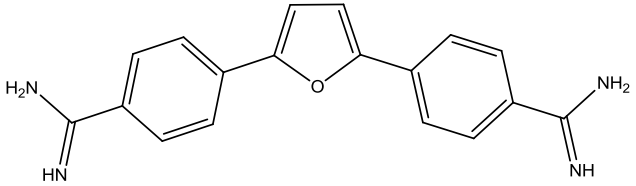
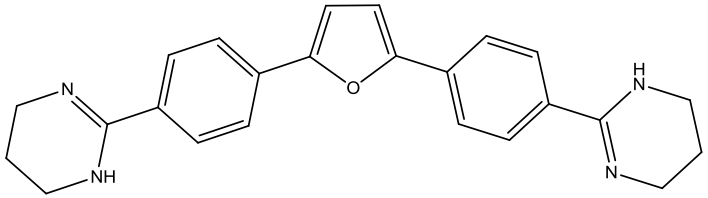
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
A		74.5	-23.6	Sigma
A	Nebularine (purine riboside)	17.1	-27.2	Sigma
A	Oxypurinol	303	-20.1	Sigma
A	Purine	18.1	-27.1	Sigma

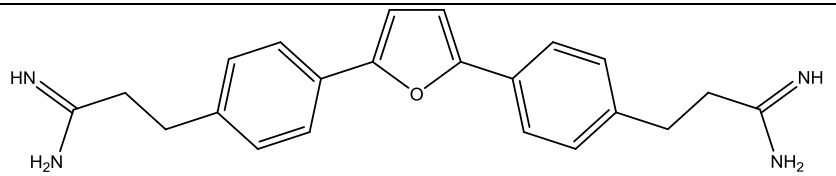
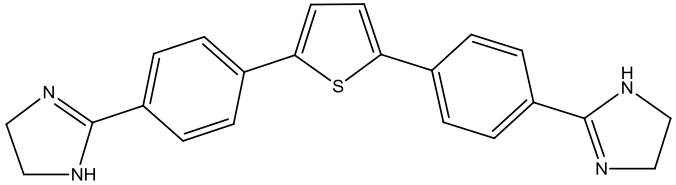
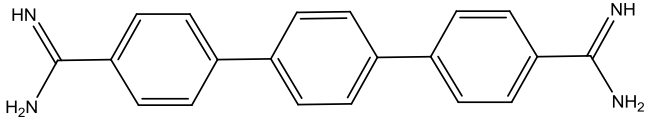
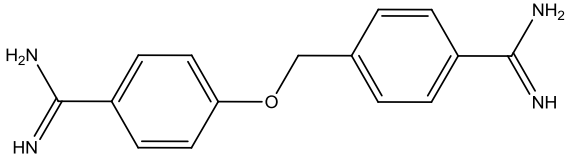
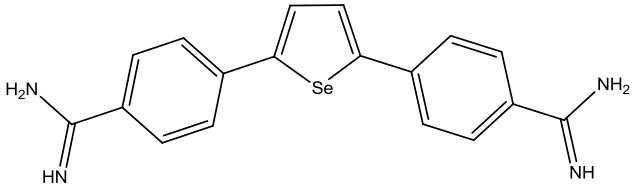
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
A		16.2	-27.3	(3)
A		9.7	-28.6	(3)
A	Tubercidin (7-deazaadenosine)	3.81	-30.9	Fluka
A		8.2	-29	(6)

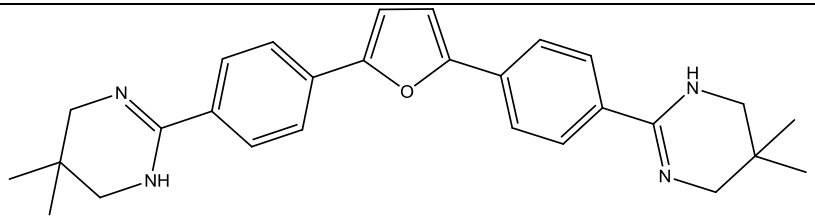
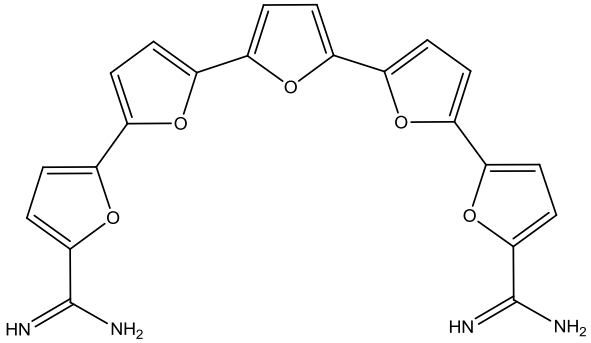
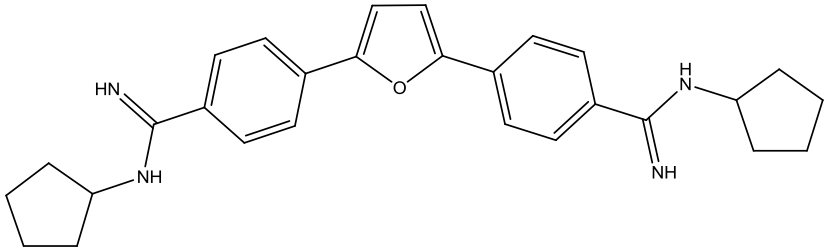
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
A		125	-22.3	(6)
A	Xanthine	106	-22.7	Sigma
B	2-hydroxybenzamidine	2030	-15.4	Acros
B	3-aminobenzamidine	722	-17.9	Acros
B	4-aminobenzamidine	22.9	-26.5	Acros
B	Benzamidine	111	-22.6	Sigma
B	furamidine 	1.19	-33.8	(5)
B	DB103 	31.4	-25.7	(5)

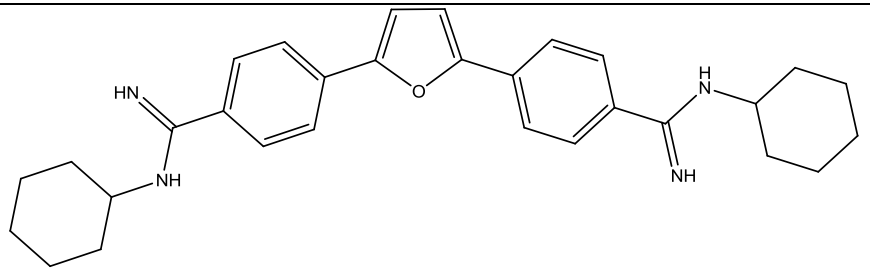
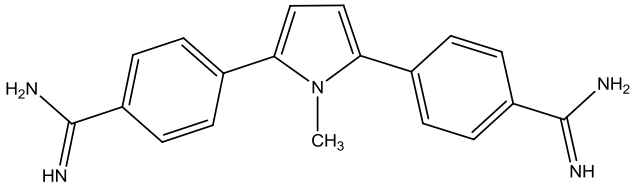
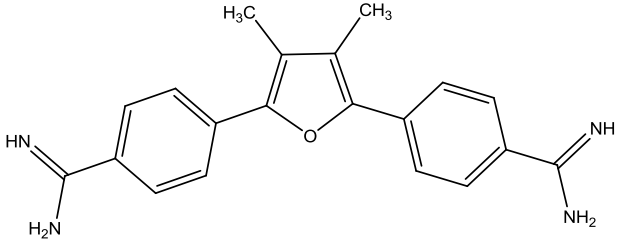
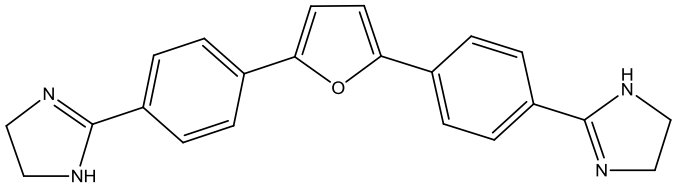
Supplemental Table 1 (continued)

Scaffold		Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
B	DB1061		7.07	-29.4	(5)
B	DB1064		33.2	-25.6	(5)
B	DB1111		5.5	-30	(5)
B	DB1138		8.1	-29.1	(5)
B	DB1213		1.09	-34	(5)

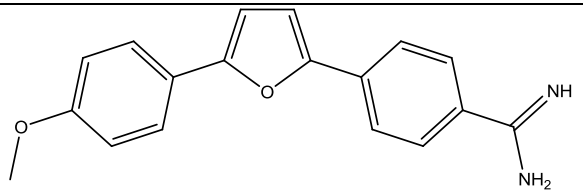
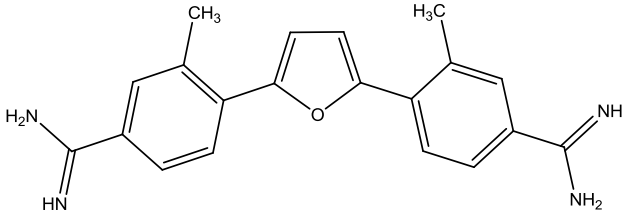
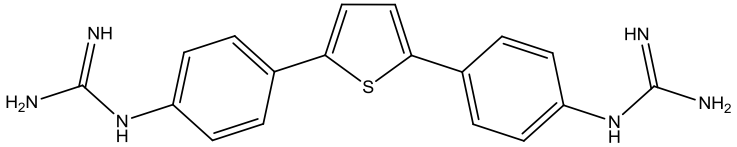
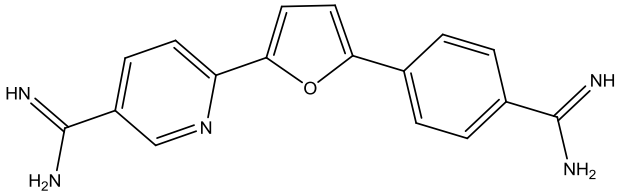
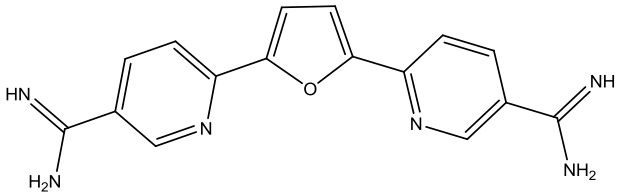
Supplemental Table 1 (continued)

Scaffold		Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
B	DB1339		25	-26.3	(5)
B	DB1680		0.95	-34.4	(5)
B	DB244		3.23	-31.3	(5)

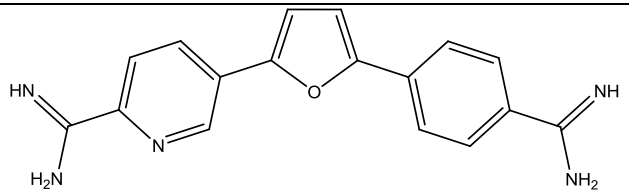
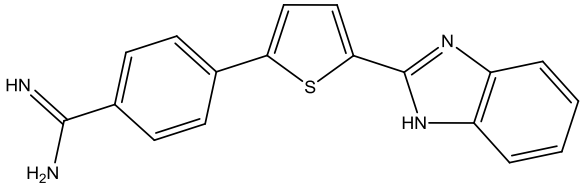
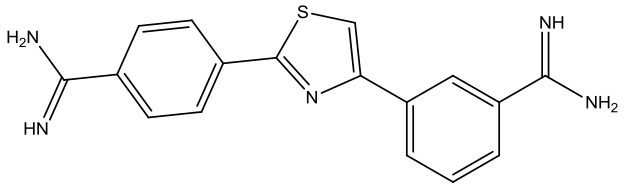
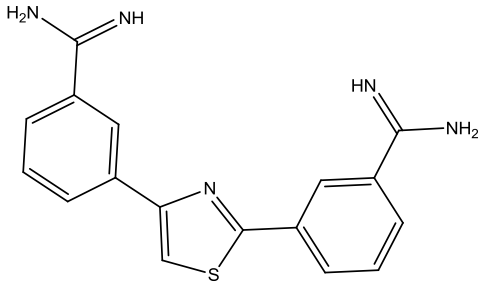
Supplemental Table 1 (continued)

Scaffold		Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
B	DB249		4.49	-30.5	(5)
B	DB320		0.39	-36.6	(5)
B	DB544		3.02	-31.5	(5)
B	DB60		13.8	-27.7	(5)

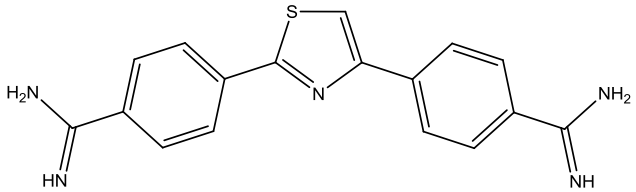
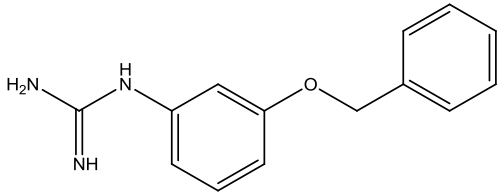
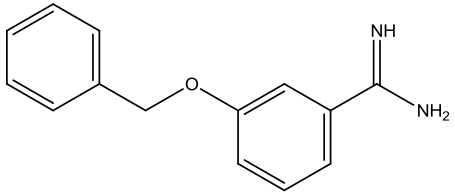
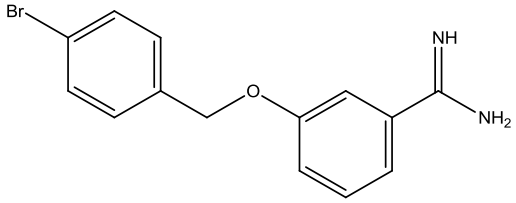
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source	Scaffold
B	DB607		4.1	-30.8	(5)
B	DB629		0.88	-34.6	(5)
B	DB686		5.4	-30.1	(5)
B	DB820		1.95	-32.6	(5)
B	DB829		1.4	-33.4	(5)

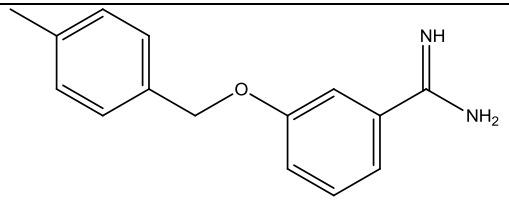
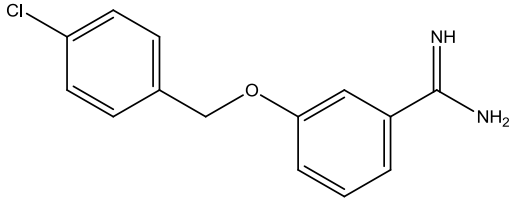
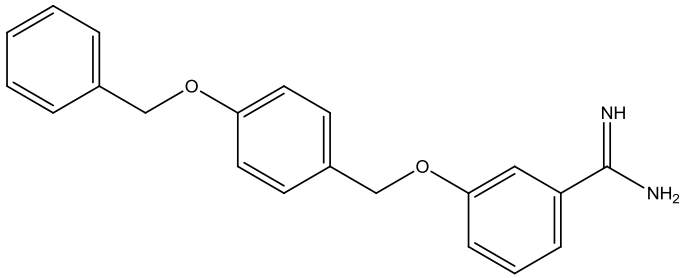
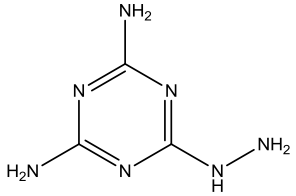
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source	Scaffold
B	DB867		4	-30.8	(5)
B	DB931		1.23	-33.7	(5)
B	Distamycin A		10.6	-28.4	Sigma
B			1.36	-33.5	(9)
B			0.92	-34.5	(9)

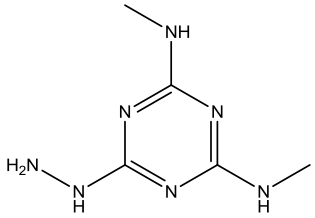
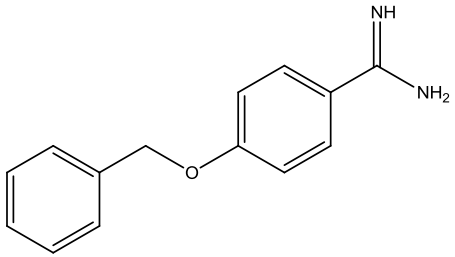
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
B		3.92	-30.9	(9)
C		8.05	-29.1	(12)
C		0.38	-36.6	(12)
C		0.81	-34.8	(12)

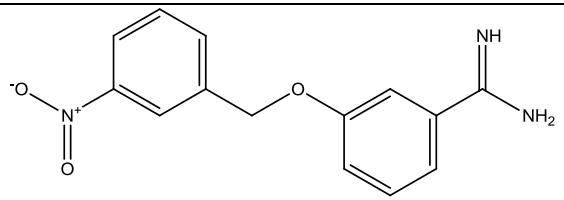
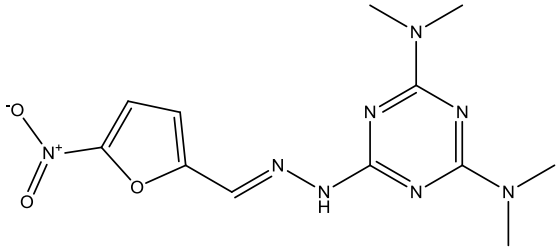
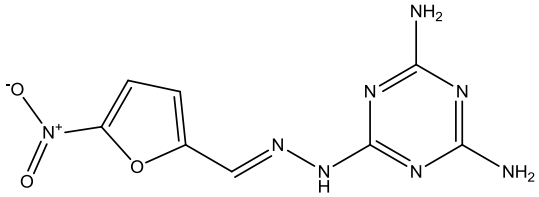
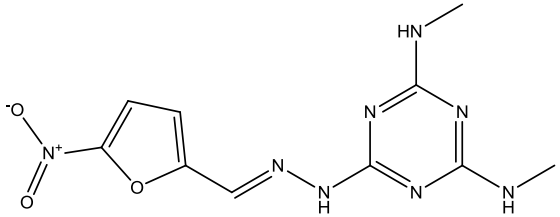
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
C		0.21	-38.1	(12)
C		1.01	-34.2	(12)
C		1.57	-33.1	(12)
C		11.9	-28.1	(13)

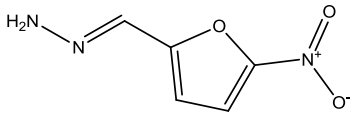
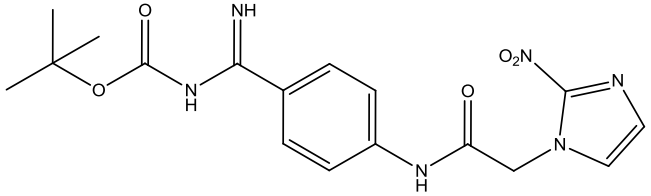
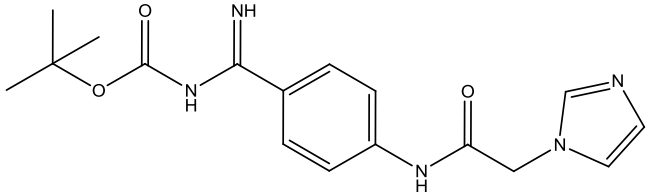
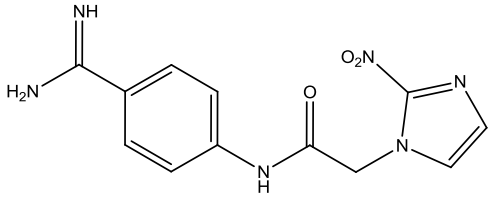
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
C		59.3	-24.1	(13)
C		5.8	-29.9	(10)
C	2-aminopyridine	14.3	-27.7	Aldrich
C	4,6-diaminopyrimidine	3.22	-31.3	Aldrich
C	4-aminopyridine	145	-21.9	Aldrich
C	4-aminopyrimidine	137	-22.1	Acros
C	4-hydroxybenzamidine	235	-20.7	Aldrich
C	butamidine	1.04	-34.2	(2)
C	stilbamidine	2.42	-32.1	Sanofi-Aventis

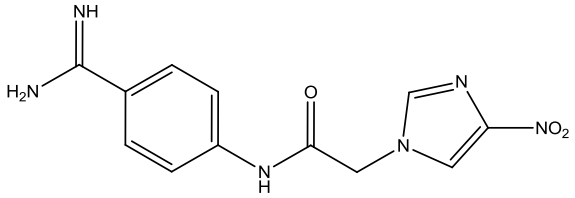
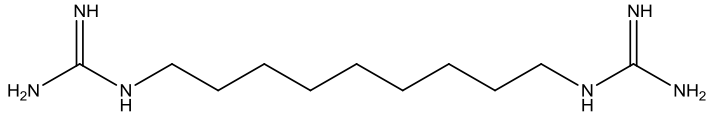
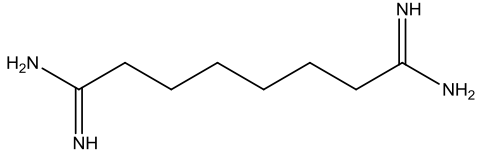
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
D		0.33	-37	(14)
D		129	-22.2	(14)
D		53.4	-24.4	(14)
D		4.6	-30.5	(14)

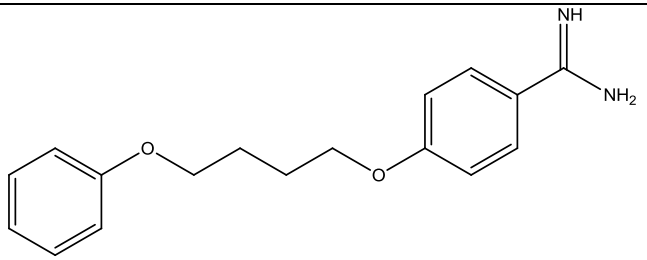
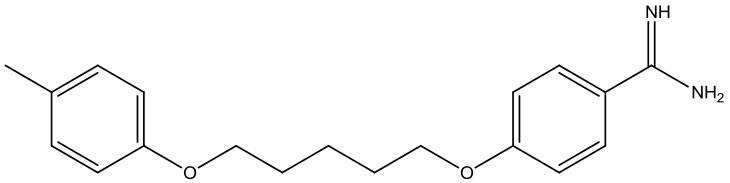
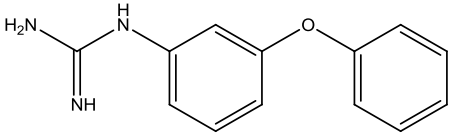
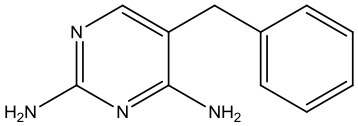
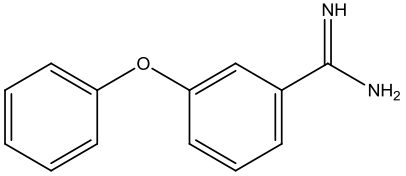
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
D		404	-19.4	(14)
D		13.1	-27.9	(14)
D		3.65	-31	(14)
D		1.58	-33.1	(14)

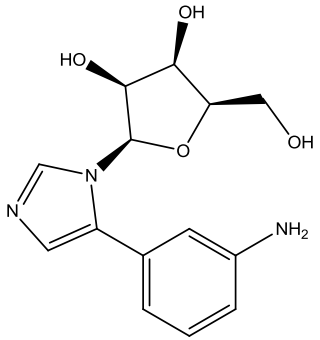
Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
D		2.88	-31.6	(14)
D	Heptamidine	0.28	-37.4	(2)
D	Hexamidine	0.43	-36.3	(2)
D	iodo-pentamidine	0.27	-37.5	(4)
D	Megazol	192	-21.2	(15)
D	Octamidine	0.48	-36.1	(2)
D	Pentamidine	0.37	-36.7	Sigma
D	Propamidine	1.92	-32.6	(11)
E	1,1'-(nonane-1,9-diyl)diguanidine 	45.4	-24.8	Biomol
E		200	-21.1	(9)

Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
E		3.25	-31.3	(9)
E		8.75	-28.9	(9)
F		9.3	-28.7	(12)
F		0.38	-36.6	(13)
F		0.38	-36.6	(12)

Supplemental Table 1 (continued)

Scaffold	Compound	K_i of P2 (μM)	$\Delta(G^0)$ (kJ/mol)	Source
F		37.7	-25.3	(3)
F	melarsen oxide	9.7	-28.6	Sanofi-Aventis
F	Melarsoprol	0.54	-35.8	Sanofi-Aventis
F	Thiamine	364	-19.6	Sigma
G	Aminopterin	78.4	-23.4	Sigma
G	diminazene aceturate (berenil)	2.36	-32.1	Sigma
G	Ethidium	5.96	-29.8	Sigma
G	Isometamidium	0.21	-38.1	May & Baker

(1) Gift of Professor Gerrit-Jan Koomen, University of Amsterdam; Amsterdam, The Netherlands.
(2) Gift of Professor Alan Fairlamb, University of Dundee; Dundee, UK.
(3) Gift of Professor Katherine Radtke-Seley, University of Maryland, Baltimore Co; Baltimore, MA, USA.
(4) Gift of Dr Philip Blower, University of Kent at Canterbury; Canterbury, UK.
(5) Gift of Professor David Boykin, Georgia State University; Atlanta, GE, USA.
(6) Gift of Professor Achiel Haemers, University of Antwerp; Antwerp, Belgium.
(7) Gift of Professor Mahmoud H. el Kouni, University of Alabama at Birmingham; Birmingham, AL, USA.
(8) Gift of Professor Simon Jarvis, University of Westminster; London, UK.
(9) Gift of Dr Paul O'Neil, University of Liverpool; Liverpool, UK.
(10) Gift of Professor Richard Tidwell, University of North Carolina, Chapel Hill; NC, USA.
(11) Gift from Dr Christophe Dardonville, Instituto de Química Médica; Madrid, Spain.
(12) Gift from Professor Ian Gilbert, University of Dundee, UK; see Stewart et al. (2005) Antimicrob. Ag. Chemother. 49, 5169-5171.
(13) Gift from Professor Ian Gilbert, University of Dundee, UK; see Tye et al (1998) Bioorg Med Chem Lett 8, 811-816 and Klenke et al. (2001) J. Med. Chem. 44, 3440-3352.
(14) Gift from Professor Ian Gilbert, University of Dundee, UK; see Stewart et al (2004) Antimicrob. Ag. Chemother. 48, 1733-1738 and Baliani et al (2005) J. Med. Chem. 48, 5570-5579.
(15) Gift from Professor Bernard Bouteille, Institut d'Epidémiologie Neurologique et de Neurologie Tropicale, Limoges, France.

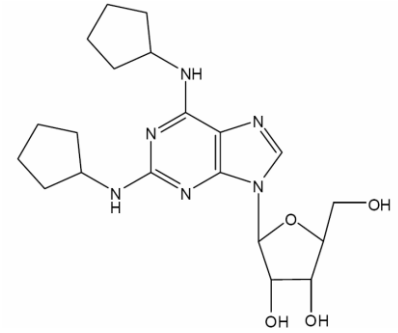
Supplemental Table 2. Listing of K_i values, Gibbs free energy ΔG^0 and energy gain/loss relative to a control compound for some of the compounds utilised in this study and listed in Supplemental Table 1. Conclusions drawn from the data with respect to substrate binding of the P2 transporter are listed in the final column.

Compound	Ki value (μM)	Δ(G0) (KJ/mol)	δ(Δ(G0)) (KJ/mol)	Relative to	Conclusion
Adenosine	0.92 ± 0.06	34.5		N/A	
Position 1					Average contribution of 7.7 kJ/mol to binding the purine ring.
1-Deazaadenosine	45.4 ± 8.7	24.8	9.7	Adenosine	N1 contributes of 9.7 kJ/mol to binding of adenosine.
1-Deazapurine	181 ± 33	21.4	5.7	Purine	N1 contributes of 5.7 kJ/mol to binding of adenine.

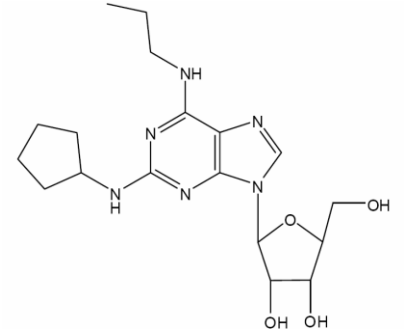
Supplemental Table 2 (continued)

Compound	Ki value (μM)	Δ(G0) (KJ/mol)	δ(Δ(G0)) (KJ/mol)	Relative to	Conclusion
Position 2					Depending on the group, substitutions on position reduce binding energy of adenosine analogs with 5 – 11 kJ/mol.
2-Nitroadenosine	81 ± 22	23.4	11.1	Adenosine	
2-Hydroxy-6-aminopurine	9.7 ± 2.3	28.6	8.7	Adenine	
2,6-Diamino, 2'deoxypurine riboside	4.4 ± 1.3	30.5	7.4	2'-Deoxyadenosine	
2-Chloroadenosine	9.7 ± 3.4	28.6	5.9	Adenosine	
Position 3					N3 does not contribute to binding. Its removal re-distributes charge around the molecule, resulting in a slightly higher affinity.
3-deazaadenosine	0.29 ± 0.06	37.3	-2.8	Adenosine	
Position 6					The 6-NH2 group contributes an average of 8.2 kJ/mol to binding of aminopurines.
6-chloropurine riboside	15.4 ± 0.8	27.5	7.0	Adenosine	
Purine	18.1 ± 3.2	27.1	10.2	Adenine	
Purine riboside	17.1 ± 2.1	27.2	7.3	Adenosine	

Supplemental Table 2 (continued)

Compound	Ki value (μM)	Δ(G0) (KJ/mol)	δ(Δ(G0)) (KJ/mol)	Relative to	Conclusion
Positions 6 and 1					The binding energies of positions 1 and 6 are additive, resulting in very low affinity for inosine and guanosine, and is estimated at 15.7 kJ/mol.
Guanosine	>500				
Inosine	>500				
2'-deoxyinosine	165 ± 23	21.6	16.3	2'-Deoxyadenosine	
1-deazapurine	131 ± 34	22.2	15.1	Adenine	
Positions 6 and 2					Loss of affinity can be attributed solely to the substitution on position 2 (see above). The single substitution at the 6-amine position is therefore not (greatly) detrimental to binding, especially when the substitution is small or flexible.
	21 ± 9.2		8.6	Adenosine	

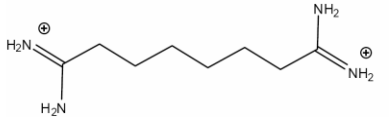
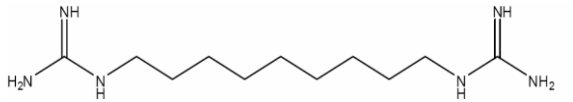
Supplemental Table 2 (continued)

Compound	Ki value (μM)	Δ(G0) (KJ/mol)	δ(Δ(G0)) (KJ/mol)	Relative to	Conclusion
	9.0 ± 1.7		5.7	Adenosine	
Position 7					Small apparent contribution to binding from N7, though too small to represent a full hydrogen bond. The absence of N7, however, decreases the positive charge on N9.
7-deazaadenosine (tubercidin)	3.8 ± 0.7	30.9	3.6	Adenosine	
Formycin A	36.5 ± 6.6	25.3	9.2	Adenosine	

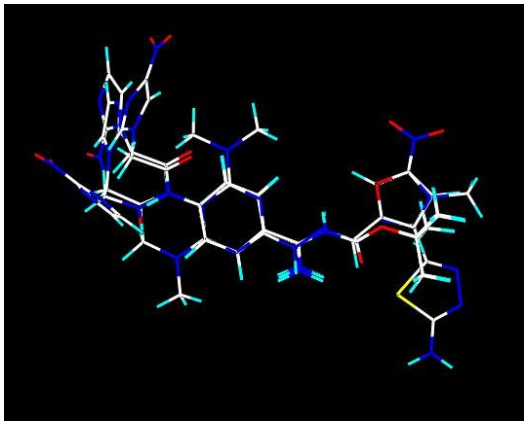
Supplemental Table 2 (continued)

Compound	Ki value (μM)	Δ(G0) (KJ/mol)	δ(Δ(G0)) (KJ/mol)	Relative to	Conclusion
Position 8					Substitutions at position 8 are detrimental to binding.
8-azidoadenosine	331 ± 142	19.9	14.6	Adenosine	
8-bromoadenosine	37.8 ± 8.2	25.2	9.2	Adenosine	
Position 9					Significant contribution of N9 to adenosine binding.
9-deazaadenosine	12.1 ± 3.2	28.1	6.4	Adenosine	
Ribose					Absence of the ribose or of just the 2'-hydroxyl group increases affinity by approximately 3 kJ/mol.
Adenine	0.30 ± 0.02	37.3	-2.8	Adenosine	
2'-deoxyadenosine	0.23 ± 0.04	37.9	-3.4	Adenosine	

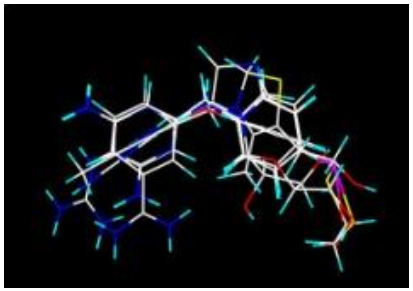
Supplemental Table 2 (continued)

Compound	Ki value (μM)	Δ(G0) (KJ/mol)	δ(Δ(G0)) (KJ/mol)	Relative to	Conclusion
Aromaticity					The aromaticity of the purines and diamidines contributes importantly to their high affinity binding of P2, as non-aromatic diamidines or diguanidines of similar length and flexibility as the aromatic diamidines display ~100-fold less affinity, corresponding to 10-11 kJ/mol in binding energy. Presumably π−π-bonds with aromatic amino acid residues are involved in substrate-transporter interactions.
	>200				
	45 ± 15	24.7	11.9	Pentamidine	
Pentamidine	0.37 ± 0.04	36.7			
Propamidine	1.9 ± 0.8	32.6			
Stilbamidine	2.4 ± 0.3	32.0			

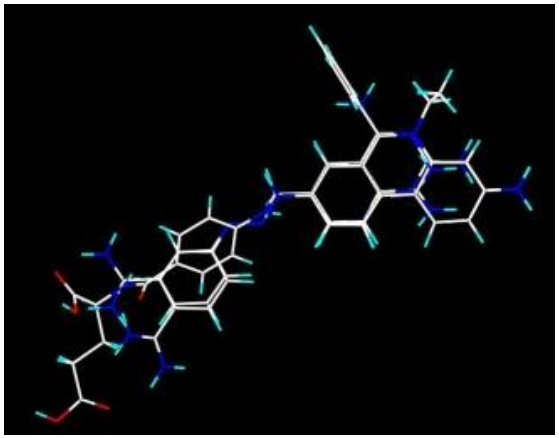
Dataset E



Dataset F



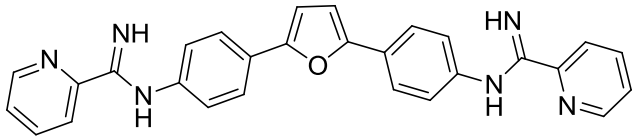
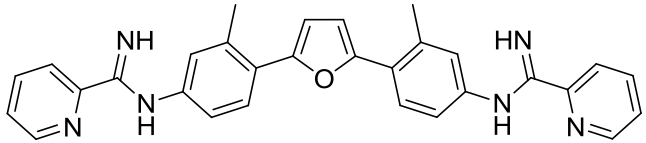
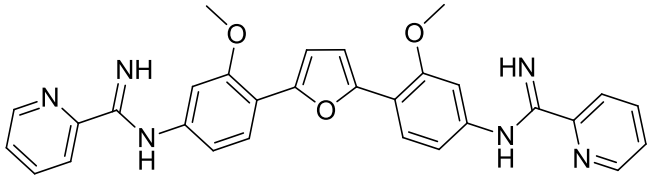
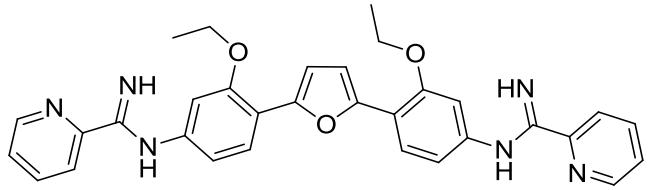
Dataset G



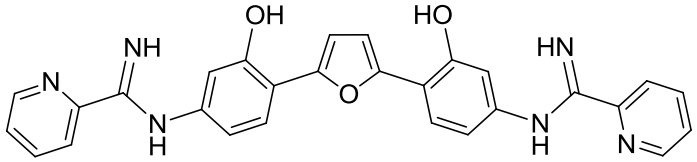
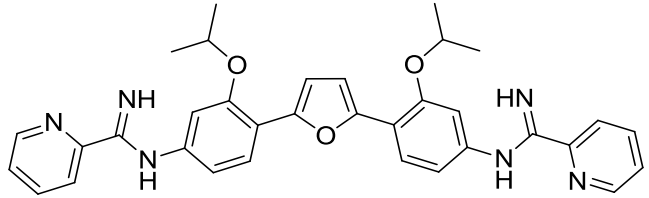
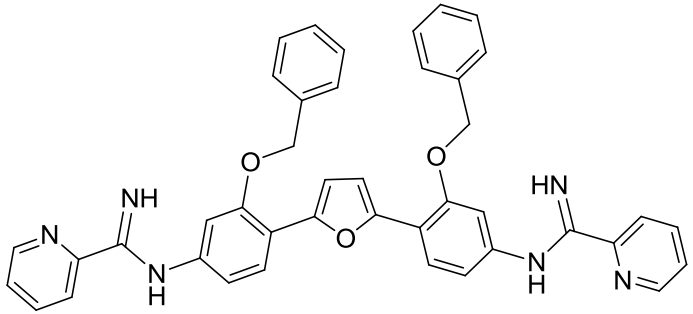
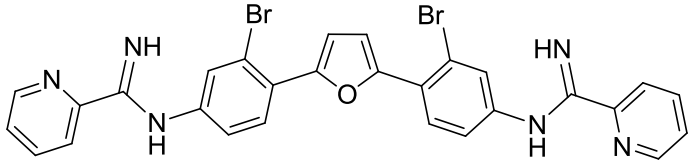
Supplemental Figure 1. Initial alignments for datasets E, F and G.

Appendix B

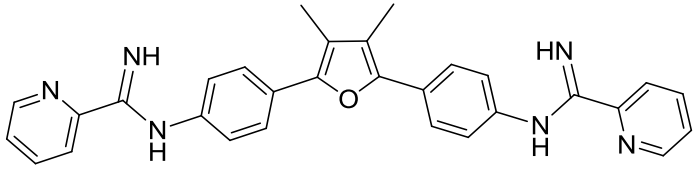
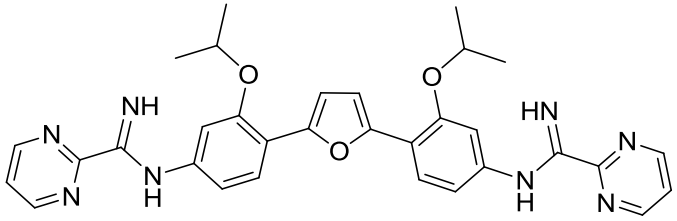
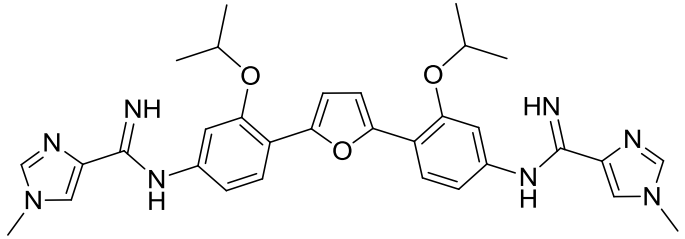
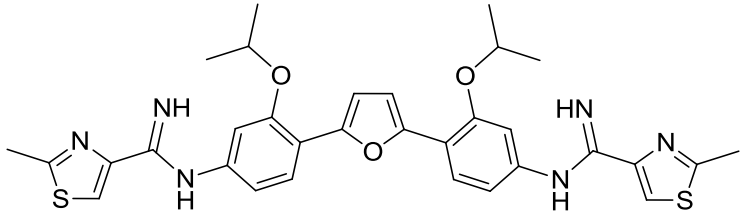
Supplemental Table 3. Training dataset of compounds with experimentally determined inhibition (IC_{50}) values against *L. donovani* (LD) and *L. amazonensis* (LA).

Name	Structure	LD	LA
DB667		1.6 ± 0.4	0.53 ± 0.19
DB702		0.67 ± 0.14	0.29 ± 0.13
DB709		0.53 ± 0.19	0.37 ± 0.15
DB745		0.50 ± 0.15	0.12 ± 0.02

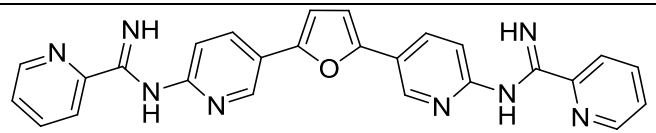
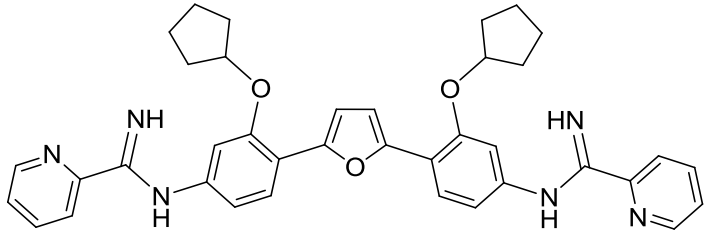
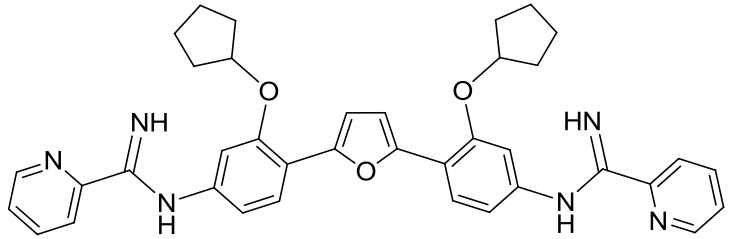
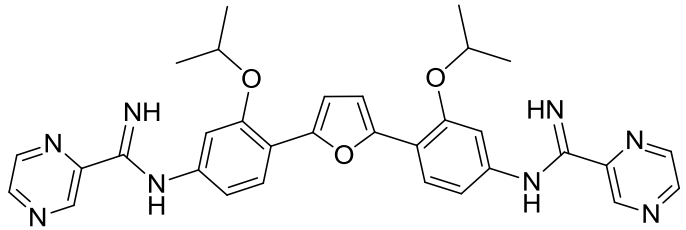
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB750		1.50 ± 0.13	0.96 ± 0.33
DB766		0.50 ± 0.10	0.087 ± .015
DB780		4.5 ± 0.9	0.51 ± 0.09
DB894		1.2 ± 0.1	0.62 ± 0.20

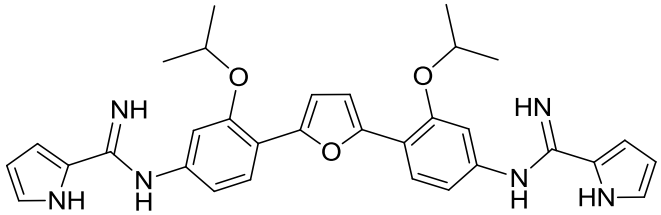
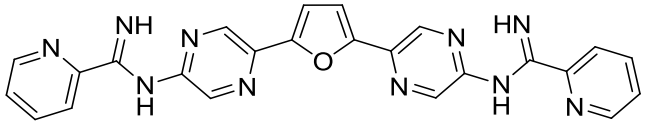
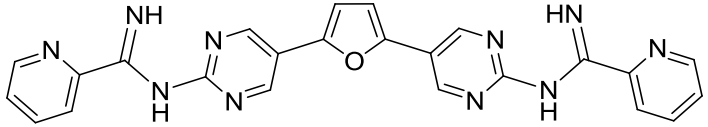
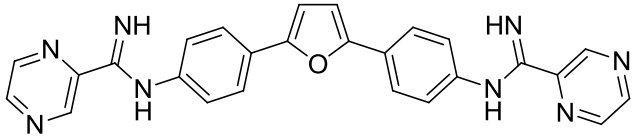
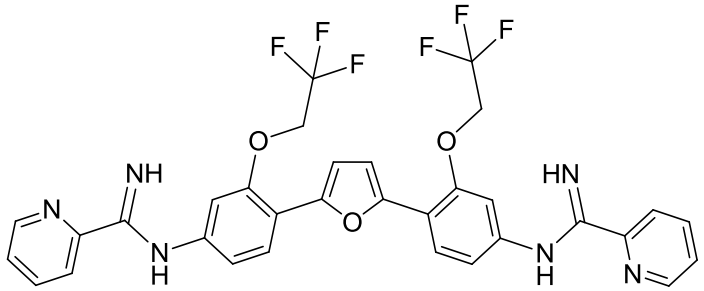
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB946		0.37 ± 0.04	0.11 ± 0.03
DB1831		0.55 ± 0.17	0.70 ± 0.16
DB1848		31 ± 7	5.0 ± 1.5
DB1850		1.4 ± 0.4	0.17 ± 0.06

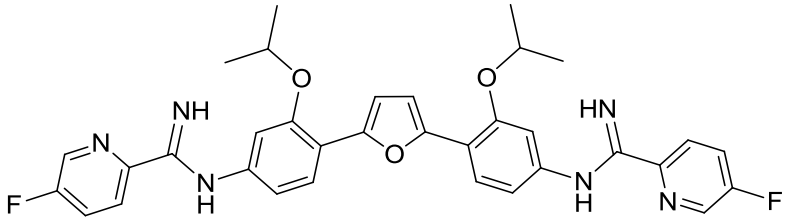
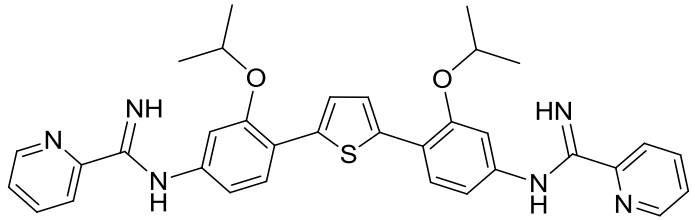
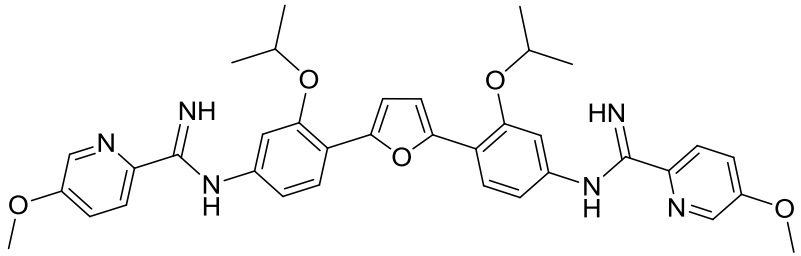
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1851		>50	>10
DB1852		1.4 ± 0.2	0.17 ± 0.06
DB1853		1.5 ± 0.3	0.21 ± 0.07
DB1855		20 ± 6	3.7 ± 3.2

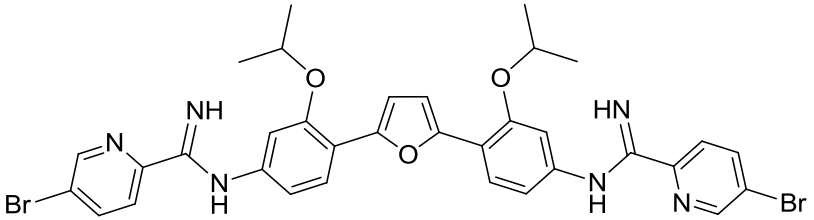
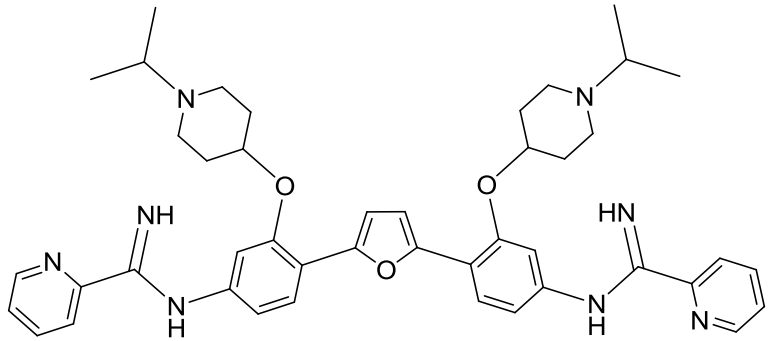
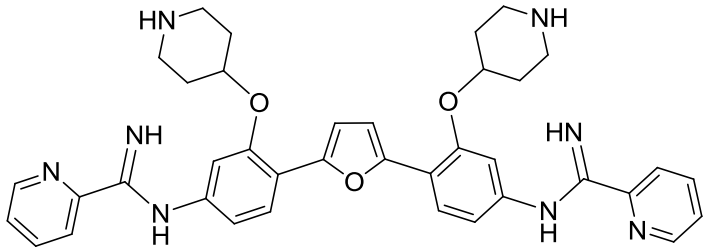
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1858		16 ± 4	0.90 ± 0.07
DB1859		>100	>10
DB1860		>100	>10
DB1861		>100	>10
DB1862		1.1 ± 0.4	0.25 ± 0.04

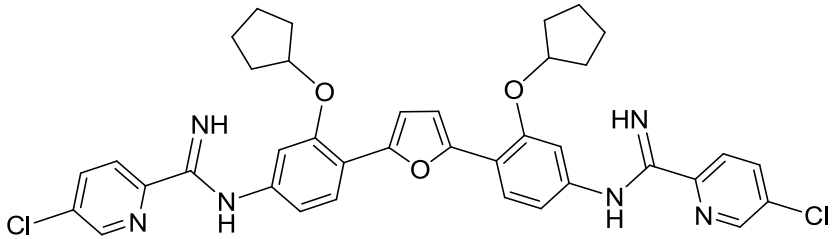
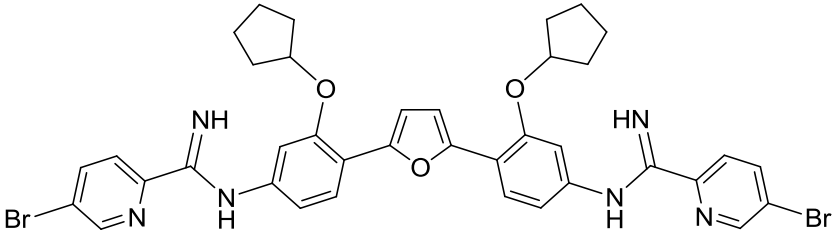
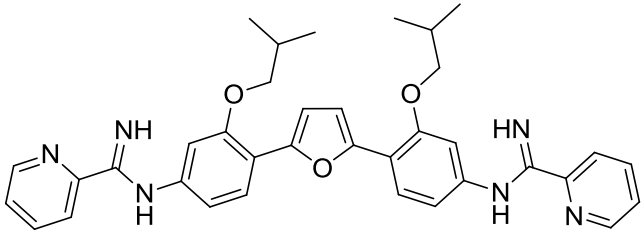
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1863		2.5 ± 0.2	0.59 ± 0.08
DB1867		0.68 ± 1.8	0.045 ± 0.011
DB1868		1.0 ± 0.1	0.13 ± 0.07

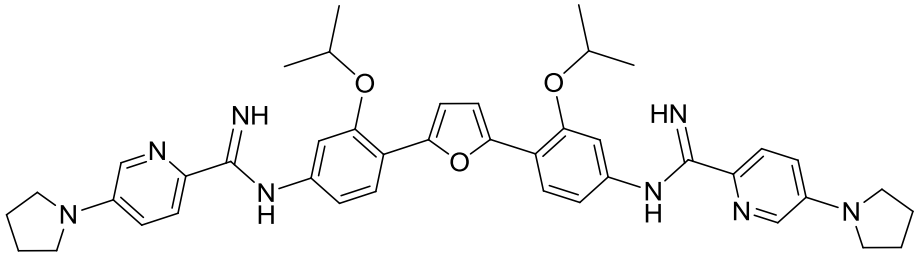
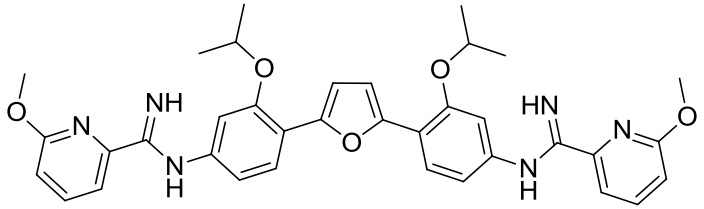
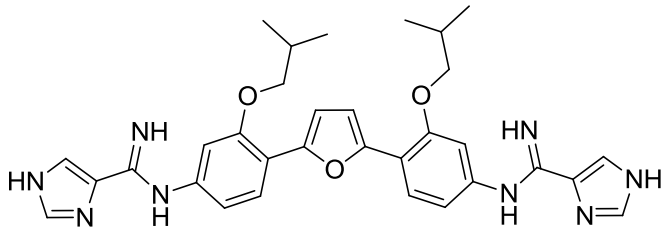
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1875		28 ± 4	>10
DB1876		28 ± 4	2.1 ± 0.9
DB1880		59 ± 9	> 10

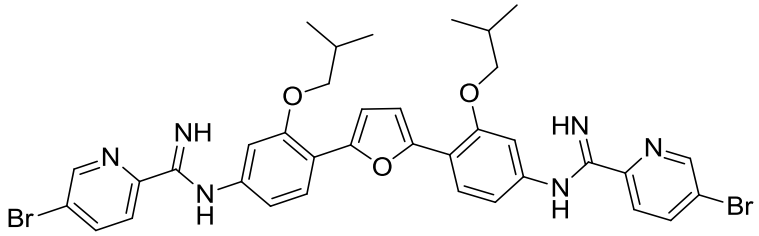
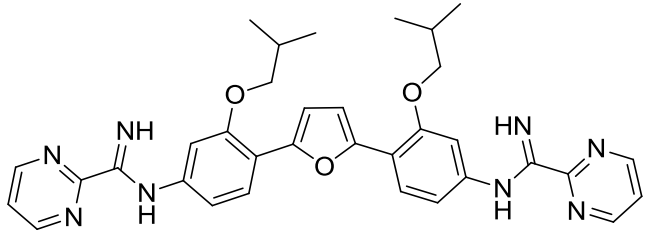
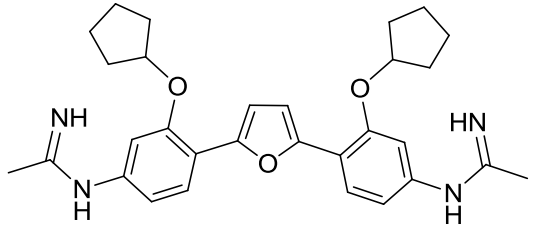
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1888		>100	>10
DB1889		>100	>10
DB1890		1.1 ± 0.2	0.095 ± 0.018

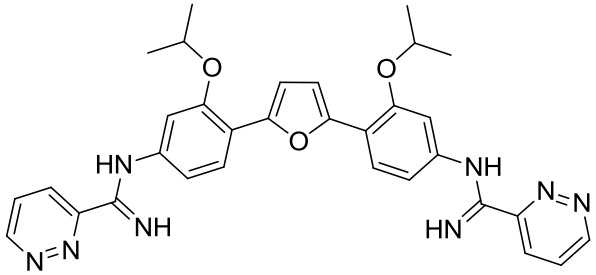
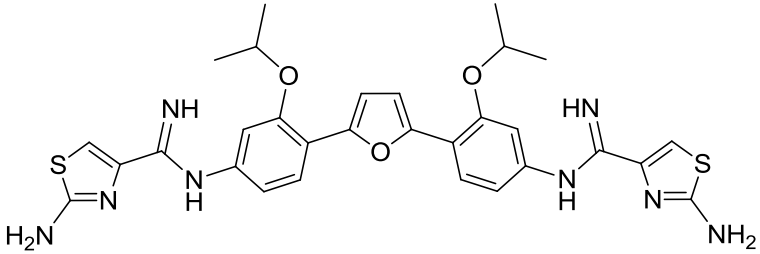
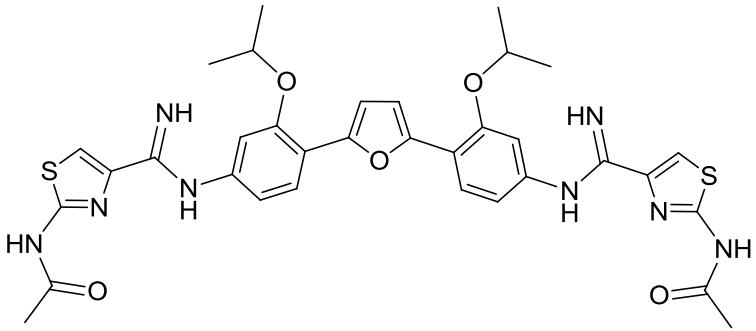
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1906		1.9 ± 0.5	0.27 ± 0.04
DB1907		4.5 ± 0.2	2.7 ± 0.7
DB1909		23 ± 4	>10

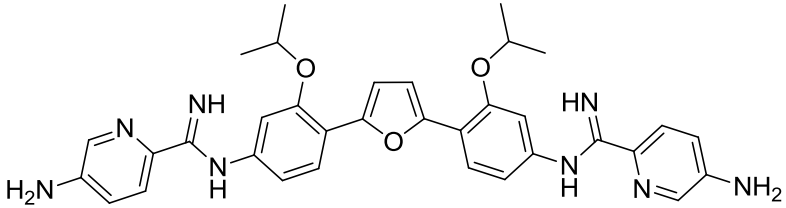
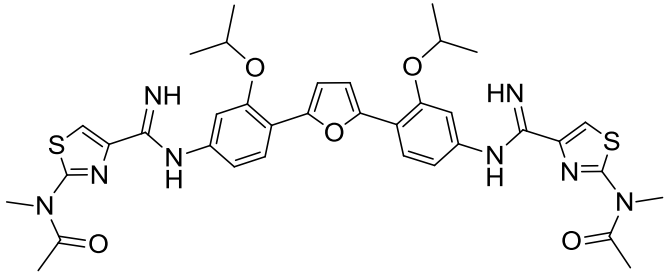
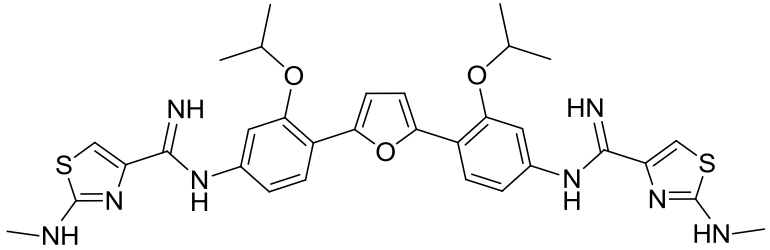
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1913		>50	>10
DB1920		2.4 ± 0.6	0.49 ± 0.08
DB1921		41 ± 3	4.7 ± 0.7

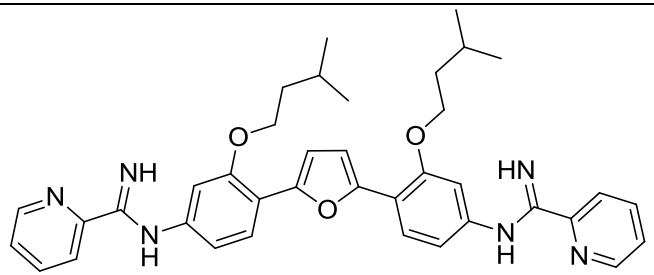
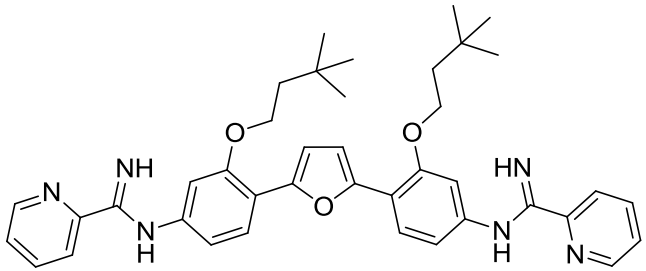
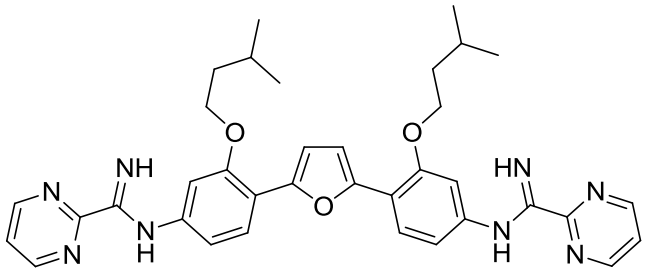
Supplemental Table 3 (continued)

Name	Structure	LD	LA
DB1937		1.1 ± 0.2	1.1 ± 0.1
DB1938		21 ± 2	1.7 ± 0.4
DB1939		>100	0.22 ± 0.02

Supplemental Table 3 (continued)

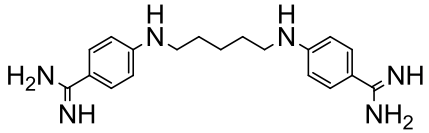
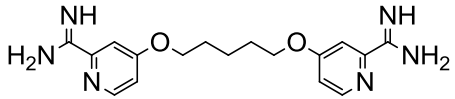
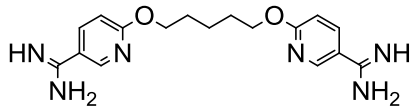
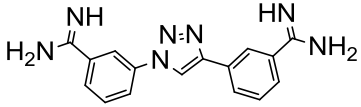
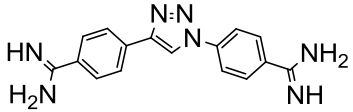
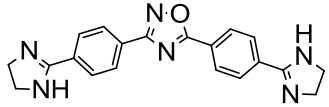
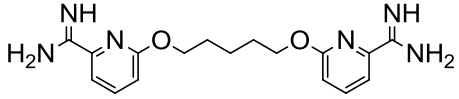
Name	Structure	LD	LA
DB1940		0.82 ± 0.10	0.64 ± 0.19
DB1942		3.6 ± 1.1	0.81 ± 0.20
DB1943		1.3 ± 0.3	0.30 ± 0.10

Supplemental Table 3 (continued)

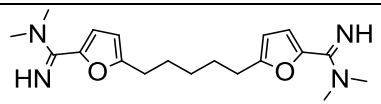
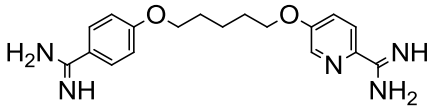
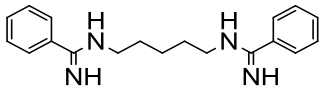
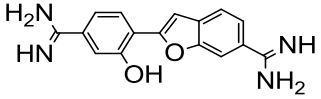
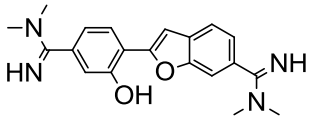
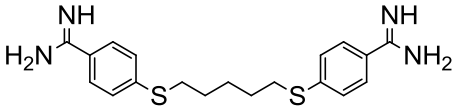
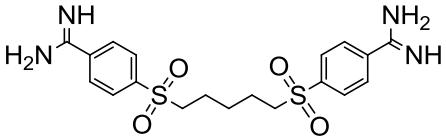
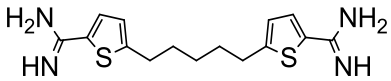
Name	Structure	LD	LA
DB1950		5.1 ± 1.2	>10
DB1951		28 ± 2	>10
DB1952		7.9 ± 1.5	0.32 ± 0.07

Appendix C

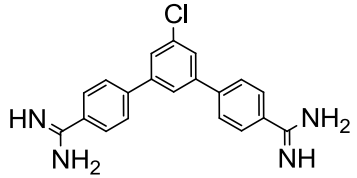
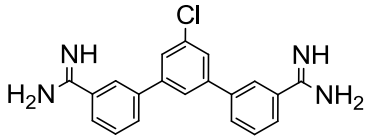
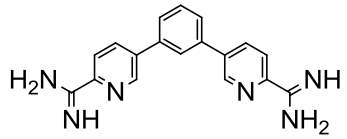
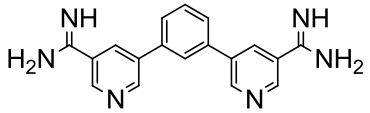
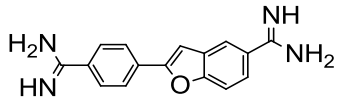
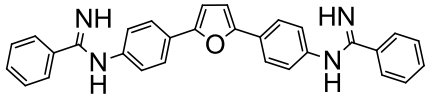
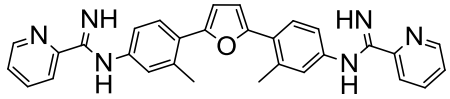
Supplemental Table 4. Compounds with experimentally determined inhibition (IC_{50}) values against *Trypanosoma cruzi*.

No.	ID	Compound	4°C	37°C
1	1MAA119		32	2.3
2	6SMB038		400	19
3	9SMB070		229.3	20.9
4	10SAB031		32	32
5	10SAB055		91.6	32
6	10SAB092		32	32
7	11SAB003		400	2.7

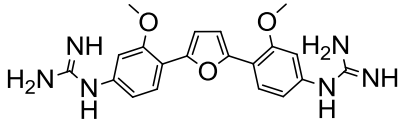
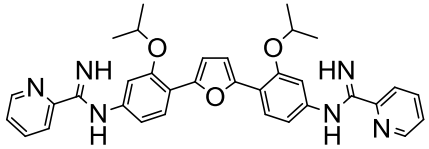
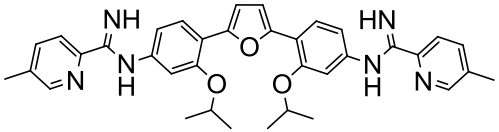
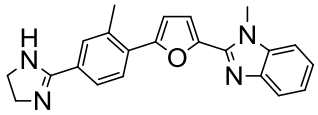
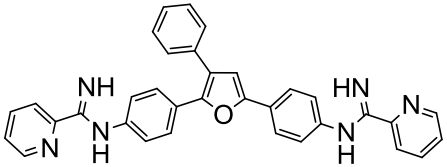
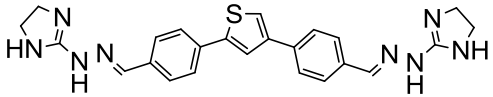
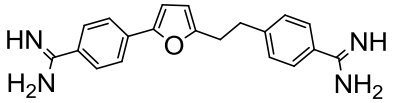
Supplemental Table 4 (continued)

No.	ID	Compound	4°C	37°C
8	12SMB032		32	32
9	14SMB013		32	9.3
10	16SAB065		400	32
11	18SMB092		32	32
12	18SMB096		32	32
13	21DAP023		128.6	0.7
14	21DAP027		400	32
15	24SMB001		128.6	1

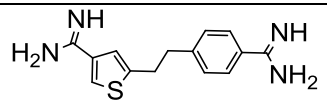
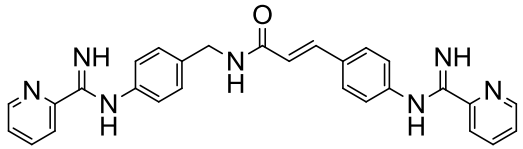
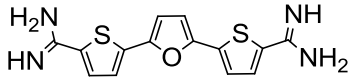
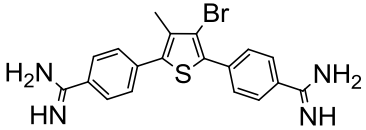
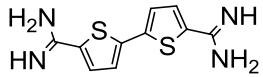
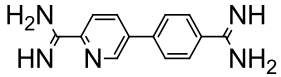
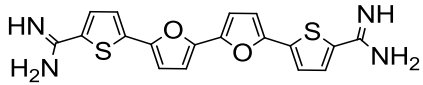
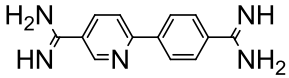
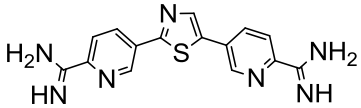
Supplemental Table 4 (continued)

No.	ID	Compound	4°C	37°C
16	25DAP009		400	1.9
17	25DAP013		32	6.1
18	27DAP060		135.8	16.3
19	27DAP080		400	32
20	150OXD049		32	32
21	DB613A		1.96	4.05
22	DB702		1.18	0.45

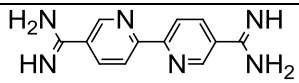
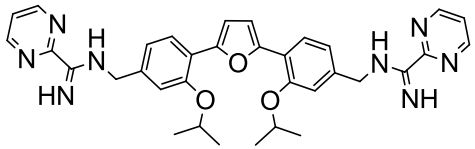
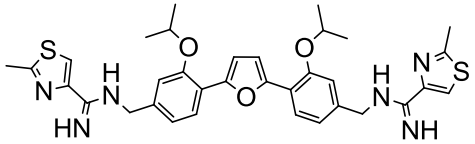
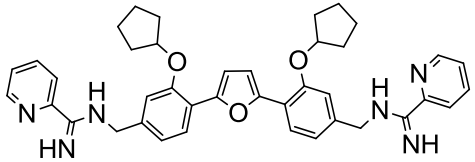
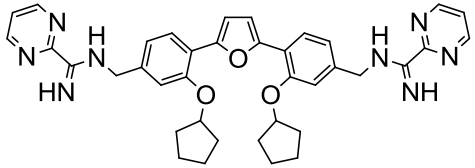
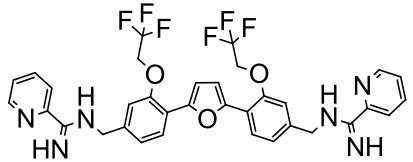
Supplemental Table 4 (continued)

No.	ID	Compound	4°C	37°C
23	DB711		32	19.4
24	DB766		0.11	0.06
25	DB786		32	0.015
26	DB824		15.54	4.43
27	DB889		0.97	0.09
28	DB1080		1.77	0.24
29	DB1195		1.21	0.26

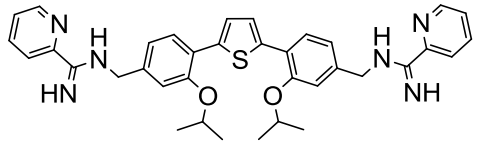
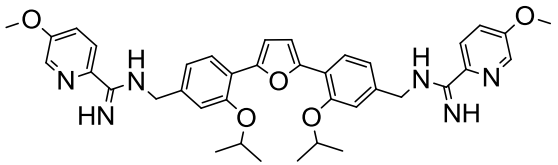
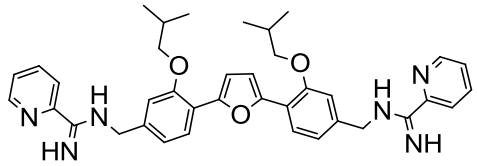
Supplemental Table 4 (continued)

No.	ID	Compound	4°C	37°C
30	DB1196		0.81	1.19
31	DB1201		6.6	1.77
32	DB1345		3.72	0.91
33	DB1362		7	6.6
34	DB1582		32	6
35	DB1627		32	32
36	DB1645		32	0.15
37	DB1646		32	31
38	DB1651		32	6.9

Supplemental Table 4 (continued)

No.	ID	Compound	4°C	37°C
39	DB1670		32	32
40	DB1831		0.08	0.02
41	DB1850		2.35	0.19
42	DB1852		32	0.06
43	DB1853		0.14	0.07
44	DB1862		0.79	0.06

Supplemental Table 4 (continued)

No.	ID	Compound	4°C	37°C
45	DB1867		0.7	0.02
46	DB1868		0.28	0.06
47	DB1890		32	0.01