

Distribution Agreement

In presenting this dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of medial, now or hereafter known, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this dissertation. I retain all ownership rights to the copyright of the dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Signature

Qi Shi

Date

Computer-Assisted Drug Discovery
Part I: Design, Development, Validation and Application of FRESH, a
Novel *In-Silico* High-throughput Screening Program
Part II: Monocarbonyl Curcumin Analogues: Heterocyclic Pleiotropic
Kinase Inhibitors that Mediate Anticancer Properties
Part III: Development of 2nd Generation NAMFIS Software Program
by Java

By

Qi Shi

Doctor of Philosophy

Chemistry

James P. Snyder

Advisor

Dennis C. Liotta

Advisor

Vince Conticello

Committee Member

Huw Davies

Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.

Dean of the James T. Laney School of Graduate Studies

Date

Computer-Assisted Drug Discovery

Part I: Design, Development, Validation and Application of FRESH, a

Novel *In-Silico* High-throughput Screening Program

Part II: Monocarbonyl Curcumin Analogues: Heterocyclic Pleiotropic

Kinase Inhibitors that Mediate Anticancer Properties

Part III: Development of 2nd Generation NAMFIS Software Program

by Java

By

Qi Shi

M.S., Emory University, 2012

B.S., Peking University, 2008

Advisors:

Dr. James P. Snyder

Dr. Dennis C. Liotta

An abstract of

A dissertation submitted to

the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Chemistry

2014

Abstract

Computer-Assisted Drug Discovery

Part I: Design, Development, Validation and Application of FRESH, a Novel *In-Silico* High-throughput Screening Program

Part II: Monocarbonyl Curcumin Analogues: Heterocyclic Pleiotropic Kinase Inhibitors that Mediate Anticancer Properties

Part III: Development of 2nd Generation NAMFIS Software Program

by Java

By Qi Shi

There is an ever growing effort to apply computational power as a routine component of medicinal chemistry and drug discovery. In Part I of this dissertation, a novel *in-silico* high-throughput screening program was developed and applied to several drug discovery projects. The program, termed FRESH (FRagment-based Exploitation of modular Synthesis by virtual High Throughput Screening), combines virtual library enumeration, rapid vHTS (virtual High Throughput Screening), pharmacological property prioritizing and 2D/3D QSAR (Quantitative Structure-activity Relationship) construction. It is designed to address the issue of balancing multiple factors during drug lead-optimization of the drug discovery process. The workflow programming platform Pipeline Pilot and the corresponding programming language PilotScript were used to construct the program. The second part of the dissertation explores the mechanism behind the pleiotropic properties of mono-carbonyl curcumin analogues by molecular modeling calculations and protein sequence alignment. The last part of the dissertation reveals the mathematical principles and the Java programming approach behind the new generation of the NAMFIS (Nuclear magnetic resonance Analysis of Molecular Flexibility in Solution) software program, together with improvements on the old version.

Computer-Assisted Drug Discovery

Part I: Design, Development, Validation and Application of FRESH, a

Novel *In-Silico* High-throughput Screening Program

Part II: Monocarbonyl Curcumin Analogues: Heterocyclic Pleiotropic

Kinase Inhibitors that Mediate Anticancer Properties

Part III: Development of 2nd Generation NAMFIS Software Program

by Java

By

Qi Shi

B.S., Peking University, 2008

Advisors:

Dr. James P. Snyder

Dr. Dennis C. Liotta

A dissertation submitted to the faculty of the
James T. Laney School of Graduate Studies of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Chemistry

2014

Acknowledgments

During my six years of graduate research at Emory, many people have provided help and encouragement. I deeply appreciate their support.

To Dr. Snyder, one of my research advisors: Not only did you offer crucial advice to guide my scientific research, but more importantly, you taught me through your own example of how to be a mature man. I still remember the first time we met in the corridor at Emerson Hall. At that time, I was still a young student just out of college, maybe “too young, too simple, sometimes naïve”. Your personal example and advice helped me to become more considerate, patient and collaborative with colleagues, students and team members. I also appreciate your invitations to your house for Christmas Eve celebration and Chinese Spring Festivals, and I’ll always remember the tasty smoked salmon. I especially appreciate your effort and encouragement during some of the difficult times I encountered during the past six years. I would not have made it through all the difficult moments without you.

To Dr. Liotta, my research advisor and the chair of the review committee: Although our communication moments have been relatively brief, your words act like the “finishing and most crucial touch to a painting” (“Hua Long Dian Jing” in Chinese, sorry I can’t find an exact English translation). Each time following a discussion with you, I’ve gained new ideas for my future work. I still remember how your single sentence “It’s not about IC_{50} , it’s about selectivity” has solve a key puzzle for me and integrated my discrete knowledge on various aspects into an entire knowledge network. I also appreciate your effort and support to help me through some of my difficult times during the past six years. Your advice and encouragement has helped me through eventually.

To Dr. Davies and Dr. Conticello: Your constructive advice and criticism throughout my graduate research time is valuable resource for me. I'll always remember your help and support during the time I encountered a failure or a problem. From the broad perspective of scientific research to tiny details like “excellent ability” or “nanocytotoxicity”, your advice has helped me significantly for the past six years.

To all the current and past Liotta group members: It is very fortunate for me to have you as colleagues. Our laboratory has been the friendliest and most helpful place I have ever experienced. Thanks a lot for sharing your precious knowledge and experience with me and for always offering help and support on all aspects. It was a wonderful experience working with all of you.

To Cindy, Tegest and other Liotta lab associates: Thanks for your effort to make sure everything in the lab runs smoothly and gets done in time.

To my collaborators: Thanks for all your support and critical assessment of my research work. Participation in collaborative teamwork has equipped me with precious experience on how to communicate with team members and how to become an excellent team player.

To BIOVIA (former Accelrys), Schrodinger, Openeye, Oracle and all the other software development groups and support teams: Thanks for your excellent modeling/programming products and customer service. My research work could not have proceeded without them. I especially thank Dr. Adrian Stevens (BIOVIA) for the invitation to speak at the Accelerate2014 meeting and for advice on job hunting and presentations. Thanks too to Dr. Fred Coughlin (BIOVIA) for the crucial training on Pipeline Pilot.

To all my friends in Atlanta: I always enjoy spending off-work time with you. Thanks for all your advice on my research, job hunting, personal emotional experiences and time spent during the holidays. Your presence has made my life in Atlanta much happier.

To my parents, families and relatives: Thanks for all your love, support and encouragement for so long. Even though we are far apart from one another, I can still feel your presence. During my irrational and uncertain times, I know you will always be there to provide reassurance.

Table of Contents

Part I: Design, Development, Validation and Application of FRESH, a Novel In-Silico High-throughput Screening Program

Program	1
Chapter 1: Introduction	2
1.1. Drug molecules	2
1.1.1 Available chemical space.....	2
1.1.2. The concept of drug likeness	3
1.1.3. CNS drug likeness.....	5
1.2. Computer assisted drug discovery	7
1.2.1 Estimation of ligand-protein interactions <i>in-silico</i>	7
1.2.2. <i>In-silico</i> estimation of physical/ADMET properties.....	9
1.3. Lead optimization challenges	11
1.3.1. Multi-target/site therapies	11
1.3.2. Balancing multiple factors	13
Chapter 2: Design and Development of the FRESH Program.....	15
2.1. Overall design strategy	15
2.2. Screening software program selection	16
2.3. Construction of the FRESH program.....	19
2.4. Algorithm design and optimization	21
2.5. Advantages of FRESH.....	24
Chapter 3: FRESH Validation Case Studies	25
3.1. Introduction.....	25
3.2. Case I: Phosphoinositide 3-Kinase, α isoform. (PI3K α) – Homology receptor model case.....	27
3.2.1. Case Background	27
3.2.2. Design of FRESH program, 1st round.....	28
3.2.3. Design of FRESH program, iterations	32
3.3. Case II: Carbonic Anhydrase 2 (CA II) – Crystal structure model case.....	34
3.3.1. Case Background	34
3.3.2. Design of FRESH program, 1st round.....	35
3.3.3. Design of FRESH program, iteration steps.....	38
3.3.4. Further experiments	41
3.4. Case III: Histone Deacetylase 1. (HDAC 1) – Ligand-only example.....	41
3.4.1. Case Background	41
3.4.3. Design of FRESH program, 1st round.....	42
3.4.4. Design of FRESH program, iterations	44
3.5. Conclusion and Future Work	45
Chapter 4: Application I: Designing novel SNRIs.....	47
4.1. Project background	47
4.2. Challenges in the lead optimization step	50
4.3. Receptor-based QSAR models of NET and SERT	51
4.4. Application of the FRESH program	55

4.4.1. The FRESH Program design.....	55
4.4.2. Resulting structures.....	57
4.4.3. Test result and comparison.....	58
4.5. Conclusion and future direction.....	59
Chapter 5: Application II: Identification of novel KCN1 analogs to block the p300/KCN1 interaction	61
5.1. Project background	61
5.2. Challenge in the lead optimization step.....	62
5.3. Receptor-based QSAR approach	63
5.3.1. Binding receptor selection	63
5.3.2. Binding site selection.....	66
5.3.3. Validate the scoring functions	69
5.4. Application of the FRESH program	73
5.4.1. The FRESH program design.....	73
5.4.2. Resulting structures.....	75
5.5. Application of FRESH to the new scaffold	79
5.5.1. New compound scaffold	79
5.5.2. Construction of QSAR.....	80
5.5.3. The FRESH program and the resulting structures	81
5.5.4. Future directions	83
Part I: Conclusions and Future Directions	85
Part II: Monocarbonyl Curcumin Analogues: Heterocyclic Pleiotropic Kinase Inhibitors That Mediate Anticancer Properties	86
Chapter 6: Curcumin analogs as Pleiotropic Kinase Blockers	87
6.1. Project background	87
6.2. Modeling of curcumin analogs	89
6.2.1. Comparison of AKT-1 and AKT-2.....	89
6.2.2. Docking pose analysis.....	90
6.2.3. Sequence Comparison of Kinase Binding Sites.....	96
6.3. Conclusion	97
Part III: Development of 2nd Generation NAMFIS Software Program by Java.....	99
Chapter 7: Design and Improvement of the NAMFIS Program	100
7.1. NAMFIS background.....	100
7.1.1. The Problem of Force Field Methods.....	100
7.1.2. Mathematical Background for NAMFIS	101
7.1.3. Problem with the Current Version of NAMFIS.....	102
7.2. New Generation of NAMFIS.....	104
7.2.1. Java VS Python.....	104
7.2.2. GUI and Backend Design	105
7.2.3. Minimization Step.....	107
7.2.4. Exception Handling	108
7.3. Conclusion and Future Work.....	110

Chapter 8: Experimental Details	112
8.1. Pipeline Pilot Tasks.....	112
8.1.1. Filter for desirable fragments/compounds by substructure matching.....	112
8.1.2. Select fragments/compounds according to physical/ADMET property values	114
8.1.3. Process the commercial library to synthetic fragments	116
8.1.4. Covalently attach fragments to the core structures.....	116
8.1.5. Remove duplicate structures.....	117
8.1.6. Merge data	118
8.1.7. Program debugging examples.....	118
8.2. Glide Task.....	120
8.2.1. Prepare protein receptor for docking and sequence comparisons.....	120
8.2.2. Generate receptor docking grids around ligand binding sites.....	121
8.2.3. Set ligand docking parameters to control output	121
8.3. MM-GBSA Task.....	121
8.3.1. Parameter settings for energy refinement	121
8.4. Miscellaneous Schrodinger Tasks.....	122
8.4.1. LigPrep – 2D to 3D structure conversion	122
8.4.2. QikProp – physical ADMET property estimates.....	122
8.4.3. Protein sequence alignment	122
8.4.4. Homology modeling	122
8.4.5. Induced fit docking – Flexible ligand and protein interaction.....	123
8.5. Additional Specific Details.....	123
8.5.1. PI3K α case study	123
8.5.2. CA II case study.....	127
8.6. Java Programming Language Implementation for NAMFIS.....	129
8.6.1. Perform user input validation.....	138
8.5.2. Perform constrained geometry and population minimization.....	140
Appendix I: A series of KCN1 analog with experimental IC ₅₀ values and predicted MM-GBSA values.....	141
Appendix II: The Java implementation of the “NAMFISOptimizer” class	159

List of Figures

Figure 1. Ladostigil is derived from the framework combination of Rivastigmine and Rasagiline.....	13
Figure 2. An example of a working Pipeline Pilot program.....	18
Figure 3. The general outline of a FRESH program.....	21
Figure 4. PI3K inhibitor scaffold.....	28
Figure 5. ROC curve for the ECFP method. AUC = 0.93.....	29
Figure 6. ROC curve for the Glide Score. AUC = 0.64.....	30
Figure 7. ROC curve for the MM-GBSA score. AUC = 0.72.....	30
Figure 8. Compound 19d.....	31
Figure 9. Docking pose of 19d on PI3K α	32
Figure 10. Compound 19f.....	33
Figure 11. Compounds 19k (left) and 19j (right).....	34
Figure 12. CA II Inhibitor Scaffold.....	35
Figure 13. ROC curve for the ECFP method. AUC = 0.88.....	36
Figure 14. ROC curve for the Glide score. AUC = 0.63.....	37
Figure 15. Compound 30.....	38
Figure 16. Docking pose of 30 at the catalytic pocket of CA II.....	39
Figure 17. Compound 3.....	39
Figure 18. Crystal structure of 3 at the catalytic pocket of CA II.....	40
Figure 19. Compound 24.....	40
Figure 20. FDA approved HDAC inhibitor SAHA Inhibitor Scaffold.....	42
Figure 21. HDAC1 Inhibitor Scaffold.....	42
Figure 22. ROC curve for the ECFP method, AUC = 0.87.....	43
Figure 23. Compound 5n.....	44
Figure 24. Compounds 5f (left) and 5g (right).....	44
Figure 25. Compounds 5i.....	44
Figure 26. Mechanism of monoamine transporter/reuptake inhibitors.....	49
Figure 27. Milnacipran.....	49
Figure 28. Correlations of estimated binding NET affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of 5.....	53
Figure 29. Correlations of estimated binding NET affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of 6.....	54
Figure 30. Correlations of estimated binding SERT affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of 5.....	54
Figure 31. Correlations of estimated binding SERT affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of 6.....	55
Figure 32. Synthetic route for the arylcyclopropylamine analogs.....	56
Figure 33. The FRESH protocol (main interface) for prioritizing arylcyclopropylamine analogs. Some sub-protocol components are not shown.....	57
Figure 34. Structure of KCN1 with three highlighted substituent groups.....	62
Figure 35. Affinity pull-down analysis of p300 and HIF-1 α proteins using KCN1-coupled agarose beads.....	64
Figure 36. ^{14}C -KCN1 binding experiment result.....	64
Figure 37. KCN1 attached to the gold surface.....	65

Figure 38. SPR sensorgrams for KCN1 binding to p300.....	66
Figure 39. p300-CH1 extracted from the complex.	67
Figure 40. Crucial residues Leu344, Leu345, Cys388 and Cys393 on p300 CH1. Leu344 is hidden under the helix behind Leu345.	68
Figure 41. Four clefts chosen for the docking sites(left) The top two sites with docked KCN1 (right).....	68
Figure 42. Two helices on HIF-1 α (purple) superimpose with docked KCN1 at Site 1 and Site 2	69
Figure 43. Crucial residues on HIF-1 α , namely Leu795, Cys800, Leu818 and Leu822..	69
Figure 44. ROC at Site 1. AUC = 0.68	72
Figure 45. ROC at Site 2. AUC = 0.70	72
Figure 46. Synthetic route for KCN1 and its analogs.....	74
Figure 47. The FRESH program (illustration interface only) for prioritizing KCN1 analogs. Some sub-protocol components are not shown.	75
Figure 48. The initial bio-test results for 95 (blue). 2609 (red) was used as a reference..	79
Figure 49. New scaffold for p300/HIF-1 α inhibitor and BW-HIF-84.....	80
Figure 50. Linear regression results for the training sets.....	81
Figure 51. Q ² results for the test set compounds.	81
Figure 52. Synthetic route for the diaryl alcohol analogs.....	81
Figure 53. The structure of curcumin.	87
Figure 54. Structures of curcumin analogs.	88
Figure 55. Sequences of aligned AKT-1(lower row) and AKT-2 (upper row with residue numbers). The residues around the ATP pocket are squared in black.....	90
Figure 56. Top predicted pose from Glide docking of N-protonated 4 to AKT-2.....	91
Figure 57. Top predicted pose from Glide docking of N-protonated 3 to AKT-2.....	92
Figure 58. Top predicted pose from Glide docking of N-protonated 5 to AKT-2.....	93
Figure 59. Top predicted poses from Glide docking of of N-protonated 6 to AKT-2: (a) 6R; (b) 6S.....	94
Figure 60. Top predicted pose from Glide docking of 7 to AKT-2.....	95
Figure 61. Cys311 in the AKT-2 cleft where the substrate binds.....	96
Figure 62. Key residues around the ATP binding sites of various kinases. Residue type and number according to AKT2: 1: Ala179 2: Met229, Glu230, Tyr231, Ala232 3: Met282 4: Glu236 5:Lys181 6:Thr292. Upper panel, >85% inhibition; lower panel, <10% inhibition	97
Figure 63. Constraints imposed in the minimization.	102
Figure 64. Equation for calculating SSD.	102
Figure 65. The GUI for the new NAMFIS program.....	107
Figure 66. Input errors identified by the program.....	109
Figure 67. Substructures with potential liability, stability and reaction concerns	114
Figure 68. An example of transformation file.	116
Figure 69. The core structure for the CA II case study.....	117
Figure 70. Debugging case 1: wrong parameter setting.	119
Figure 71. Debugging case 2: extremely long runtime.....	120
Figure 72. CA II, first iteration demo	129

List of Tables

Table 1. Six arylcyclopropylamine compounds with clogP and IC ₅₀ values.....	50
Table 2. Structures selected to pursue synthesis and the predicted result	58
Table 3. Comparison of predicted results to experimental results.....	59
Table 4. KCN1 and active analogs with multiple experimental measurement repeats	71
Table 5. Result of the FRESH approach for KCN1 analogs.....	77
Table 6. Result of the FRESH program for KCN1 analogs.....	83
Table 7. IC ₅₀ values (unit: μM) of a series of curcumin analogs against various kinases.	89

List of Abbreviations

ADMET	adsorption, distribution, metabolism, elimination, toxicity
BBB	blood-brain-barrier
BPBD	binding pocket burial degree
CNS	central nerve system
FRESH	fragment-based exploitation of modular synthesis by vHTS
HBD	hydrogen bonding degree
HIF	hypoxia inducible factor
HTS	high-throughput screening
FDC	fixed dose combination
GBSA	generalized Born/surface area
GI	gastric intestine
GST	glutathione S-transferase
GSU	Georgia State University
JRE	Java runtime environment
MD	molecular dynamics
MM	molecular mechanics
MTL	multi-target ligand
MW	molecular weight
NAMFIS	NMR analysis of molecular flexibility in solution
NMR	nuclear magnetic resonance
NET	norepinephrine transporter
PD	pharmacodynamics

PK	pharmacokinetics
PSA	polar surface area
PLS	partial least square
QM	quantum mechanics
QSAR	quantitative structure-activity relationship
ROC	receiver operating curve
SERT	serotonin transporter
SNRI	serotonin norepinephrine reuptake inhibitor
SPR	surface plasmon resonance
SQL	structured query language
SSRI	selective serotonin reuptake inhibitor
WCI	Winship Cancer Institute

Part I: Design, Development, Validation and Application of FRESH, a Novel In-Silico High- throughput Screening Program

This part of the dissertation describes the design, development, validation and application of a computational program that is capable of proposing novel, potent and property/ADMET (Adsorption, Distribution, Metabolism, Elimination and Toxicity)-adjusted synthetic candidate structures based on synthetic schemes devised by practicing chemists: FRESH. This program is constructed by the Pipeline Pilot programming platform together with the PilotScript programming language. The application of FRESH in the early stages of a drug discovery project in addition to the traditional medicinal chemistry exploration is expected to encourage and facilitate a closer collaboration between computational and synthetic chemists.

Chapter 1: Introduction

This introductory chapter lays the background for the construction of crucial features of the FRESH program.

1.1. Drug molecules

1.1.1 Available chemical space

The identification of therapeutic targets is often the starting point for modern drug discovery projects. These targets are frequently cellular proteins. Depending on the specific disease, the target can be derived either from the host (human) or the pathogens like viruses, bacteria or fungi. To access the therapeutic effect in early stages, various in-vitro and/or in-vivo cell-based bio-assays are subsequently developed. The majority of therapeutic agents are chemicals, and medicinal chemists play a crucial role at this stage by providing various chemicals, either natural products or synthetic analogs, to the corresponding bio-assay tests to identify bio-active compounds. Multiple pharmacological properties are also tuned-up at this stage to provide suitable clinical candidates for further development.

High-throughput screening (HTS) technology is a milestone of drug discovery history. It has significantly accelerated the screening process compared to the past. Nowadays, thousands of compounds can be screened against a biological target within a modest time-frame. Nevertheless, the available chemical space remains vastly unexplored. According to an estimation by Bohacek et al., even with a constraint of molecular weight (MW) less than 500 amu, the number of possible structures still reaches the scale of $\sim 10^{60}$,

which far exceeds the total number of atoms on the earth.¹ Exploring the entire chemical space is obviously infeasible even with the development of HTS methodology. In addition, medicinal chemists still have to use synthetic methods to access chemical space beyond the immediate commercially available ones, which is time-consuming and labor-intensive in the pre-clinical stage of the drug discovery process.

Therefore, it is more feasible and efficient to identify and focus on small and discrete compound series with possibly enriched potential drug candidates rather than simply searching hopefully in the vast ocean of chemical space.

1.1.2. The concept of drug likeness

Potency is definitely a crucial criterion for a drug molecule. However, since the drug molecule is administered into the body, it also has to be delivered to a specific site and minimize side effects, which usually requires additional pharmacological properties. Ideally, the drug molecule will remain at an effective concentration within the therapeutic window for a period of time after administration to ensure the desired efficacy while minimizing potential side effects. Those pharmacological property factors are also crucial for a molecule to be administered as a drug.

Several retrospective statistical studies on the existing pharmacological space have revealed several contributing physical/ADMET (Adsorption, Distribution, Metabolism, Elimination and Toxicity) factors. Among these, the “Rule of Five” derived by Lipinski for assessing oral availability is still the most widely applied one. Based on ~2,200 compounds that have passed phase I and entered phase II clinical trials, the statistic showed that 90% of the compounds possess no more than 5 total N-Hs and O-Hs (hydrogen-bond donors) and no more than 10 hydrogen-bond acceptor atoms (N and O).

The MW stays within 500, while logP values, a measure of lipophilicity, are generally less than 5.² In 2009, Jorgensen proposed a “Rules of Three” which was based on ~1,700 known neutral drugs.³ The result revealed that 90% of the drugs have solubility (measured by the logarithm of solubility in the unit of mol/L) no less than -6, greater than 30 nm/s Caco-cell permeability and no more than 6 predicted primary metabolites. Compared to the previous Lipinski rule, the Jorgensen rules addressed additional ADMET issues like cell permeability and metabolism. In 2011, Morelli et al. performed a study based on a small dataset (~40) of protein-protein interaction inhibitors.⁴ Complementary to the guidelines of the Lipinski Rules of Five, Morelli’s Rules of Four posits that the average value for MW is larger than 400, logP greater than 4, number of rings above 4 and number of H-bonds more than 4. Accordingly, the new rules have expanded the concept of what constitutes a drug-like molecule.

It is worth noting the review by Leeson et al., which summarized the trends of properties based on the current medicinal chemistry literature.⁵ It reveals increasing trends on MW and lipophilicity, which are likely the consequence of enhanced organic synthesis skills and development of drug formulation technologies. For example, the mean cLogP and MW values are 4.0 and 435 for a list of 1,680 compounds from the recent literature according to Morphy et al..⁶ Oprea et al. also revealed higher cLogP values (> 4.25) and MW (>425) in more than 50% of compounds with high potency.⁷ It appears that the more recent compounds tend to deviate from the traditional drug-like space. Nonetheless, as compounds progress through later stages, the MW and lipophilicity often decline. Developmental selection pressures are likely the contributing factors, as larger and more lipophilic molecules have augmented risks for bioavailability,

solubility, toxicity, synthesis and formulation. The chance of success decreases for these larger and lipophilic compounds as the project proceeds. Therefore, it is still reasonable to adopt the traditional rules during the pre-clinical stage of a drug discovery project.

It is important not to overlook the limitation of these rules. There are exceptions to them. For example, compounds from the category of natural products frequently lie beyond the chemical space of the Lipinski “Rules of Five”, providing successful drugs like Taxol. Thus, these rules should be treated as “rules of thumb” rather than accurately defining the boundary between “drug” and “non-drug”. For a drug discovery project, since exploring the entire chemical space is not an option, focusing on the boundaries of the above-cited rules is more likely to result a successful drug. In other words, these rules prioritize synthetic candidates for medicinal chemists among the vast available chemical space. They are useful selection criteria, especially if the drug discovery project has originated from target-based HTS and focused on small organic inhibitors.

1.1.3. CNS drug likeness

CNS drugs act on the central nervous system. One major difference between an CNS active drug and a peripherally active one is that, in the absence of damaged and leaky brain tissues, the former has to penetrate an additional barrier, the blood-brain-barrier (BBB), before reaching the desired therapeutic target. The transfer of drugs and other materials from the blood stream into the cell and the reverse direction are controlled by endothelial cells inside the BBB. These epithelial cells form tight junctions, possess few pinocytotic vesicles and lack fenestration. Also, even if the molecule manages to cross this endothelial cell layer barrier, it can still be pumped back into the blood by p-glycoprotein dependent active transfer processes. Penetrating the BBB for a molecule

raises additional challenges compared to just the gastric intestinal (GI) membrane. The requirements for physical/ADMET properties for CNS drugs, therefore, are generally more stringent. Designing a CNS drug requires more attention to physical/ADMET properties.

From a general perspective, CNS drugs tend to possess lower MW, higher lipophilicity in terms of calculated logP, less flexibility measured by the total number of rotatable bonds and reduced polarity (measured by polar surface area). According to the study by Levin et al. in 1980, the suggested MW cutoff for CNS drugs is 400.⁸ Another study in 2002 by van de Waterbeemd et al. suggested 450.⁹ Both of these cutoff values are lower than the one suggested by Lipinski Rules of Five, which is 500. Kelder et al. discovered that CNS drugs have lower polar surface area (PSA, cutoff is 70 Å²) compared to non-CNS drugs (120 Å²).¹⁰ Leeson and Davis performed a comparison between CNS and non-CNS drugs in 2004.¹¹ They analyzed ~70 CNS drugs vs ~260 peripherally acting drugs, respectively and concluded that the average percent of polar surface area (PSA) for CNS drugs is 16.3, while the value for all drugs is 21. The average number of rotatable bond for CNS drugs is 4.7, compared to 6.4 for all drugs. Hansch and Leo's quantitative analysis demonstrated that cLogP correlates nicely with LogBB (log ratio of drug concentration in the brain to the one in the blood). However, as stated in the last section, increased lipophilicity can reduce the chance of success as the project proceeds, so the balance of increased LogBB and chance of success in later stages of drug refinement is crucial.¹² The more stringent requirements of physical and ADMET properties for CNS drugs definitely pose additional challenges in the CNS drug discovery process.

1.2. Computer assisted drug discovery

1.2.1 Estimation of ligand-protein interactions *in-silico*

With the development of computer technologies, various computational strategies are available to provide the estimation of ligand-protein interactions, thus enabling the screening of large virtual libraries of molecules for the activity against certain protein receptors. Computational methods like molecular dynamics (MD), molecular mechanics (MM), quantum mechanics (QM) and hybrid QM/MM are employed to assess the ligand-protein interaction. QM and MD can provide more accurate results and dynamic information, but generally consume considerably greater computational resources. MM, on the other hand, is usually less computationally expensive. It can provide excellent molecular geometries. The hybrid QM/MM methodology is a balanced approach between computational cost and accuracy, which applies the more accurate but computationally demanding QM approach to protein regions around the binding pocket and less accurate but relatively fast MM for the rest of the protein-ligand model.

Computational programs like DOCK, FlexX, Glide, GOLD and ICM perform quantitative and semi-quantitative calculations of the ligand-protein interaction *in-silico*. Comparison studies have been performed to evaluate the accuracy of these programs.^{13,14,15,16} Such studies evaluated reproduction of the available crystal structure of protein-ligand complex by extracting the ligands from the complex structure and then docking them back into the corresponding proteins. The comparison studies have revealed that the Glide program, which is included in the Maestro package developed by Schrodinger Inc., remains the most accurate method. Glide had the highest portion of top-ranking poses within 2Å of the original crystal structures, which is comparatively the best overall match.

Further detailed analyses also revealed that Glide showed the best enrichment factors, the least susceptibility to increased ligand flexibility and is less sensitive to hydrogen bonding and binding pocket burial.^{13,15}

Glide adopts a hierarchical protocol for predicting possible poses of ligands. The protein receptors are first prepared by removing water molecules, adding missing hydrogen atoms, optimizing hydrogen bonds and assigning amino-acid residues with appropriate protonation state before performing a ligand docking. A docking grid which constraints the available docking space of the ligand is subsequently selected by users via the specified selection methods and grid size (absolute 3D coordinates, auto-selection centered at a residue, auto selection centered at a molecule, etc). Glide also performs a moderately extensive conformational search for the ligand structures based on the OPLS2005 force-field. The ligands are subjected to a thorough exploration of possible positions and orientations within the designated binding grid of the receptor, followed by energy minimization on pre-computed OPLS Vander Waals and electrostatic grids to generate various docking poses. These poses are ranked by the Glide score, which is an energy-based function. Poses with favored hydrophobic, hydrogen bond or metal-ligation features are prioritized, while those with unfavorable steric clashes and protonation states are penalized.

The enhanced performance of Glide is achieved by several factors. The protein side chains are often flexible. To accommodate this flexibility without consuming large computational resources, Glide adopts reduced vdW scaling parameters (default is 0.85 instead of 1.0), which softens the penalty for unfavorable steric repulsive interactions and consequently increases the likelihood of discovering active ligands. In addition, Glide

implements higher conformational space coverage for the ligands, which may also contribute to discovering active ligands. Perola et al. revealed that relative to ICM and GOLD, the performance of Glide also benefits from post-refinement by OPLS-based energy-minimization.¹³

While Glide docking procedures are generally able to reproduce X-ray crystal structures of the complexes and provide valuable information in the docking of novel ligands, it still needs improvement for predicting binding affinities under some circumstances. The rescoring by physics-based methods like MM-GBSA (Generalized Born/Surface Area, included in the Prime program in the same Maestro package) circumvents some limitations of the Glide scoring functions, especially the ones associated with desolvation and entropy penalties for the ligands upon binding. Guimaraes and Cardozo performed a comparison study on several types of protein receptors. It revealed that, relative to the Glide score, the MM-GBSA score can produce remarkable correlations between calculated results and experimental data under some situations.¹⁷ In the following sections, Glide was chosen to generate the poses of ligands docked in the receptors, while both the Glide scores and Prime MM-GBSA scores are applied to estimate the non-covalent interactions between ligands and receptors.

1.2.2. *In-silico* estimation of physical/ADMET properties

As stated in the previous section, a successful drug should also possess favorable physical/ADMET properties. Developing drug candidates without the consideration of these properties can result in significant failure in the later stages. According to Kubinyi et al, poor bioavailability problems led to 40% clinical failure in the early 90s.¹⁸ The assessments of ADMET have gradually become routine since then, and the failure rate

dropped to 11%. Kola et al. reported that the major causes for drug failure have now shifted to lack of efficacy and toxicity.¹⁹ Experimental determination of these properties is costly and time-consuming, and some properties may never be determined at early stages at all. For this reason, various computer programs are designed to prioritize molecules based on predicted physical/ADMET properties.

The “Qikprop” program developed by the Jorgensen group and provided by Schrodinger Inc. provides estimation of some physical/ADMET properties.²⁰ This program is also available in the same Maestro package as Glide and MM-GBSA. The input molecule requires a 3D structure with explicit hydrogens added, therefore, a ligand preparation job must be performed if the input files are 2D structures. For some basic properties like MW, number of H-bond donors/acceptors, number of primary metabolites and number of rotatable bonds, they can be directly derived from the input structures. Others are estimated based on existing statistical models constructed on various 2D or 3D descriptors and statistical methods (partial least square), like the predicted octanol/water partition coefficient (LogP), solubility (LogS), blood-brain-barrier penetration (logBB), Cell permeability (Pcaco, MDCK), hERG blockage and serum protein binding. In addition to the estimated values for each input structures, QikProp provides the corresponding ranges for the specific property obtained from 95% of known drugs, which is a useful reference for evaluating a new molecule. It also displays warning messages to the users by flagging ~30 reactive functional groups which are known or likely to trigger false positive results in screening tests. Qikprop provides normal and fast mode as two different calculation options. Their speeds differ about 30 fold. For the fast mode, it skips the PM3 calculation so that the dipole moment, ionization potential and electron affinity

descriptors are not available. As a result, some predictions under the fast mode may use different sets of descriptors and, thus, results may vary from the normal mode.

Other commercial entities also provide physical/ADMET prediction programs based on similar principles. In the Pipeline Pilot programming platform developed by BIOVIA (former Accelrys), components for simple property calculations are available like MW and the number of H-bond donors/acceptors. In the 8.5 version release, components are also available for estimating blood-brain-barrier penetration, cytochrome P450 2D6 inhibition, hepatotoxicity and plasma protein binding.²¹ Unlike the previously mentioned Qikprop, it takes 2D structures and use 2D descriptors. The platform also allows users to incorporate additional user-derived data to parameterize for improved accuracy. The software package provided by ACD lab includes estimation of the octanol/water partition coefficient, pKa, solubility, blood-brain-barrier penetration, cytochrome P450 inhibition and hERG inhibition.²² It also allows user-supplied training set data. Some other online programs are listed at <http://www.vcclab.org/online.html>. The estimated results from these programs are routinely referenced in the drug discovery project. They are crucial components in the FRESH program to be discussed in detail in Chapter 2.

1.3. Lead optimization challenges

1.3.1. Multi-target/site therapies

The generally accepted design principle for a drug can be summarized as “the magic bullet”. Alternatively speaking, these drugs are designed selectively for a single therapeutic target. Nevertheless, many diseases remain ineffectively treated with the single magic bullet. Since modulating multiple targets simultaneously can potentially

enhance the efficacy compared to single target drugs, there is an increased interest on designing a “magic shot-gun”, or a “dirty drug” to have effect on several targets simultaneously instead of a “magic bullet” on a single target. This is particularly true for anti-cancer and CNS agents. The serotonin norepinephrine reuptake inhibitors (SNRIs) to be demonstrated in Chapter 4 and curcumin analogs in Chapter 6 are such examples in which multiple protein receptors are likely to be beneficially targeted.

Several strategies are available for developing multi-target drugs in the current drug-discovery atmosphere. The widely accepted “cocktail therapy” in a number of diseases is one example of targeting multiple proteins. It combines more than one therapeutic mechanism and requires two or more individual tablets administered simultaneously. This approach is effective in diseases such as AIDS and various cancers. However, it can suffer from poor patient compliance, particularly in situations where the drug is used against asymptomatic diseases like hypertension or for life-improvement purposes. An improved solution to the cocktail therapy is the fixed dose combination (FDC). In this approach, a single tablet is formulated with two or more agents together to improve patient compliance. Vytarin, which combines ezetimibe (cholesterol absorption inhibitor) and simvastatin (a type of statin, HMG-CoA reductase inhibitors), are examples of the FDC strategy. In addition, for both the cocktail therapy and the FDC approach, the patent life of old drugs can be prolonged. However, these two approaches require administration of multiple components. Variations in the pharmacokinetics/pharmacodynamics (PK/PD) profiles of multiple components require more extensive clinical studies compared to a single component, let alone the discrepancies of relative rates of metabolism between patients which increase the complexity of PK/PD relationships. Additionally, drug-drug

interactions are always a concern when more than one component is administered simultaneously. Another disadvantage of FDC over cocktail therapy is the commercial uncertainty. In practice, clinicians might still prefer the “cocktail therapy” for greater dose flexibility and lower cost in the case of generic drugs.²³

The advantage of using a single component drug against multiple targets is obvious. Like the FDC approach, patient compliance will be less of an issue relative to the cocktail therapy. Additionally, the single chemical component frees researchers from complications like complex PK/PD correlations and drug-drug interactions during the research and development stage. The knowledge-based method, like framework combination (either using a linker for two scaffolds or completely merged scaffold; see below), has also offered interesting drug leads. As illustrated in **Figure 1**, the neuro-protective agent Ladostigil combines the framework of Rasagiline (monoamine oxidase inhibitor) and Rivastigmine (acetylcholinesterase inhibitor). As expected, it demonstrates dual inhibition effect against both monoamine oxidase and acetylcholinesterase.²⁴

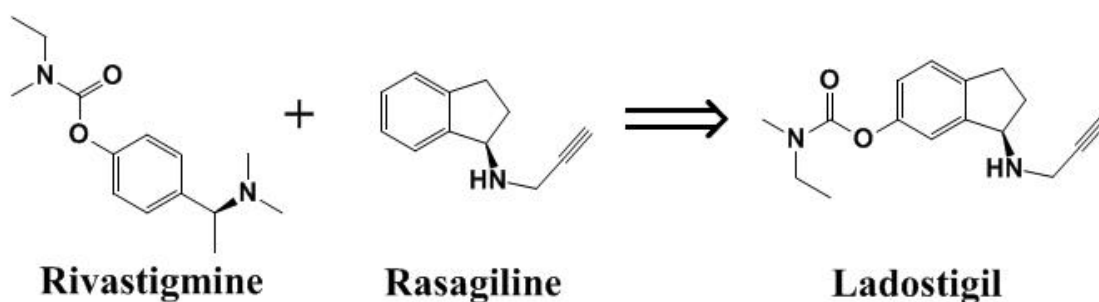


Figure 1. Ladostigil is derived from the framework combination of Rivastigmine and Rasagiline.

1.3.2. Balancing multiple factors

Despite the benefits of using a single agent for multiple targets, one major drawback for developing such multi-target ligand (MTL) drugs is that it is considerably more

complicated to simultaneously gain acceptable activity at two or more targets/sites. In addition, whether an optimal potency ratio exists remains a question, and precisely adjusting the ratio is even more difficult.

Meanwhile, as stated in the sections above, other physical/ADMET factors like lipophilicity, solubility, metabolic stability and BBB permeability (for CNS drugs) also require attention. Some of these factors may act against each other so the drug has to stay in a relatively narrow chemical space in which the “conflicting” factors are all acceptable. Balancing all these factors is already challenging for a single target drug, let alone the MTL drugs.

At this stage, synthetic feasibility is also a concern. Synthetic processes generally consume considerable time and are always labor intensive. This is particularly a big concern for academic institutions with limited funding, labor and resources.²⁵ Thus, synthetic chemists are generally reluctant to invest large amounts of time pursuing just one compound that may eventually fail. Although lead optimization occurs at an earlier and less expensive stage of the entire drug discovery process, balancing multiple factors is nonetheless a challenging task. In this part of the dissertation, I intended to address this problem by developing a novel program called FRESH, which combines knowledge-based rational design and virtual high-throughput screening (vHTS).

Chapter 2: Design and Development of the FRESH Program

2.1. Overall design strategy

The knowledge-driven rational design approach can provide some interesting candidates during the lead optimization stage, but there are certain limitations. The knowledge-driven empirical approach alone generally addresses only one aspect of the requirement, like biological activity, solubility, metabolic stability or selectivity. The situation is more complicated when it is necessary to balance multiple factors. In such a case, the method applied (see FRESH below) should be structured so as to introduce constraints for each of the factors simultaneously consistent with the needs of the therapeutic endpoint.

It is obvious from the astronomical number in Section 1.1.1 that exploring the entire chemical space is not feasible. Thus, medicinal chemists will usually investigate the chemical space around a lead molecule and, hopefully, come up with at least one development candidate with both improved activity and some ADMET properties. However, due to the relatively slow rate of organic synthesis, the coverage of chemical space around a lead by actual organic synthesis is still limited. To increase the number of compounds investigated without significantly compromising the synthesis speed, medicinal chemists frequently pursue modular syntheses based on hit compounds originating from external discoveries as well as low- and high-throughput screening. In this process, a core structure is preserved while exploring one or more substitution patterns by using different building blocks but the same or similar synthetic method. Modular synthesis enables chemists to obtain a series of analogs in a relatively short

period. Nevertheless, the total number of possible synthesis candidates in this context is still well beyond the reach of any academic or pharmaceutical group. On the one hand, medicinal chemists face numerous choices. The chemical space around a drug lead, which is just a tiny portion of the total space, still represents astronomical numbers of possible synthetic candidates. On the other hand, a molecule has to meet various requirements to be considered as a drug. The fundamental design principle for the FRESH program presented here is thus generated: First, construct a diverse and tailored virtual compound library by an *in-silico* modular synthesis. Second, apply a set of prioritizing criteria to identify molecules with improvements in potency and physical/ADMET properties.

2.2. Screening software program selection

The construction of the initial virtual compound library requires combinatorial enumeration over a large dataset. Both Pipeline Pilot (BIOVIA, Dassault) and CombiGlide (Maestro package, Schrodinger) are able to perform combinatorial library enumeration. The fundamental principle behind the two programs is the same. First, supply the program with a core structure with marked attachment points. Second, fragments from previously prepared fragment libraries are covalently connected to the core structure to form the molecular structure. For CombiGlide, the enumeration procedure is already pre-defined and relatively easy to follow. However, it performs poorly in terms of speed. Loading one file of about ~70,000 fragments consumes over 40 minutes. Additional problem also arises when the enumeration scale reaches 7,000. While suitable for processing small sets of data, CombiGlide was not chosen to construct the initial library in the present work.

Pipeline Pilot, in contrast, provides little pre-defined procedures. This platform is actually a workflow programming interface which provides various components. Each component encapsulates codes specifically designed for a task (for example: sorting) at the backend. While the detailed implementation of the backend codes are hidden from the users, the platform still requires considerable knowledge on procedure-based paradigm programming and designing workflows to complete specific tasks. However, its major advantage over CombiGlide is the speed performance. A typical initial fragment processing from a 7 million compound library usually consumes less than one hour. The ability to handle over a billion potential drug candidates per day is easily achieved, demonstrating excellent speed performance and an efficient use of FRESH to probe great swaths of chemical space. In addition, the lack of pre-defined procedures can sometimes become an advantage. It allows greater flexibility when adjusting the existing program to a specific requirement is needed.

Other available components in Pipeline Pilot for relational database management and processing also allow the program to perform other tasks required in FRESH. In addition, during execution, each pipeline displays a concurrent number (**Figure 2**) reflecting the progress of the job. The numbers provide crucial information for estimating execution time, program feasibility, parameter optimization and, most importantly, program debugging. Given all the advantages, Pipeline Pilot was chosen not only to perform the initial library enumeration, but also process the data and present the results of the FRESH program. For the incorporation of calculation results from other programs, FRESH includes the corresponding junctions to accept the result files. To minimize the potential

problem of converting files in different formats, it is conceivable to utilize one shared file format.

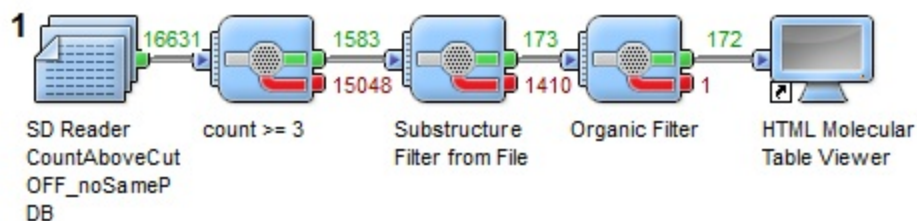


Figure 2. An example of a working Pipeline Pilot program.

The initial combinatorial library generally contains numerous compound structures. The total number frequently reaches the scale of thousands, and it is not uncommon to encounter scaling to hundred million or even over billions. Obviously, it is impractical to include a procedure which requires manual processing of each individual structure. FRESH is designated as a virtual high throughput screening (vHTS) program. Consequently, all the software programs involved in FRESH need to support batch calculations.

As discussed in Section 1.2, Glide and Prime MM-GBSA included in the Maestro package are both valuable tools for estimating ligand-receptor interactions. In addition, both programs support batch processing, so the Glide scores and MM-GBSA scores were chosen for evaluating ligand-protein interactions. For physical and ADMET property estimates, Qikprop was employed in FRESH instead of the ACD program since it is designed to perform batch-mode calculations. All the programs mentioned above support the sdf file format, which can be regarded as a relational database with chemical structure support. Consequently, the sdf file format serves as the “communication language”

between Pipeline Pilot and other involved software programs throughout the entire FRESH program.

2.3. Construction of the FRESH program

The FRESH program consists of four major steps. Since potency is frequently a selection criterion, the program has to establish QSARs based on existing data for predicting the potency of new compounds. The input/training set for QSAR is generally an external library (ChEMBL, BindingDB), which includes published historical testing data. In addition, previously synthesized and tested compounds within the research group can also be added as additional data points. As stated in Section 2.2, the results from Glide and MM-GBSA programs are possible candidates for evaluating potency. In addition, Pipeline Pilot also provides various components to calculate other descriptors or perform statistical analysis. The QSARs generated in this step are applied for potency evaluation in the later stages of FRESH.

The second step is construction of a virtual molecule library, the basis of which is strictly based on a practical modular synthetic scheme derived by a collaborating chemist for a given target series. This step can be regarded as an *in-silico* synthesis mimic corresponding to wet-lab synthesis. Take the intended synthesis of amides as an example. Amides are formed by acid chloride and amine building blocks, among others. In the bench-top synthetic process, building block compounds from commercial vendors are purchased and the corresponding amides are obtained by a nucleophilic reaction. In the FRESH program for virtual library construction, building block structures are queried against a virtual compound library. The source of such a library may be a commercial compound electronic database provided by various vendors like Chem-Navigator, Zinc,

Maybridge etc., a pharmaceutical company's own electronic inventory or a research laboratory's list of all the previously obtained compound and intermediate structures. Thus, the building block structures obtained at this step are regarded as immediately available or easily obtainable ones. In addition, to avoid the drawback of missing possible interesting fragments, the program can be easily manipulated to allow incorporation of additional fragments. For example, users can supply additional members to the fragment library with output from shape and electro-similarity matches by isosteric replacement or any other candidate deemed important for the project team to explore. The fragment structures are then covalently attached to the core by the "Perform Reaction" and "RG reader/writer" components in Pipeline Pilot to obtain the corresponding target molecule structures.

As stated in the previous section, favorable physical/ADMET properties are crucial components of a successful drug discovery campaign. FRESH routinely makes use of these as additional filters in the selection of fragments and molecules. The widely accepted Lipinski "Rules of Five" and Jorgensen's "Rules of Three" are generally included in the selection scheme. Depending on the specific project, the users may add additional rules or adjust the acceptance range for these properties. An example would be a polar surface area (PSA) cutoff specific to the treatment of potential CNS agents. As mentioned in Section 1.1.2, there are specific scaffolds that violate these rules. In case a particular scaffold proves to be an outlier of these rules or the rules remove all possible candidates, the selection criteria based on the corresponding properties can be modified or simply dropped from the FRESH filtering scheme.

The final part of the FRESH protocol is the processing and merging of the calculated results and the selecting of structures of interest, which satisfy the desired properties. The “Merge Data” module in Pipeline Pilot spares users from writing complex structured query languages (SQL) for joining results from different files. The resulting novel target structures can be viewed and stored by various viewer/writer components. The end-game prioritized structures are considered highly interesting synthesis candidates. **Figure 3** demonstrates the general outline of a FRESH program.

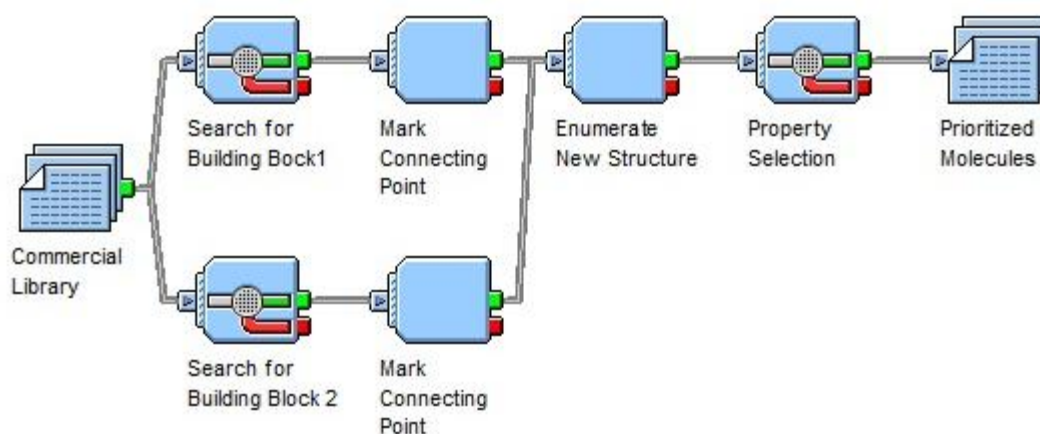


Figure 3. The general outline of a FRESH program.

2.4. Algorithm design and optimization

As stated in the last section, FRESH has to be organized to perform large scale jobs. The chemical space around a drug lead, though not as large as 10^{60} , can still reach an astronomical scale, mainly due to combinatorial explosion. For example, assume that the target molecule structures are obtained from three building blocks from modular synthesis and each building block has over 1000 choices. The total number of possible molecule structures in the initial enumerated library is already above 1 billion. Popular building blocks like aldehydes and amines, depending on the commercial library, can go

over 10,000. Hence, the algorithm for the FRESH program, particularly the construction of the initial screening library, still requires informed selection.

For the optimization of any algorithm, the first round of improvements usually results from resolving the conundrum between the reduction of execution time for a given computational task and the increase in size of a given data set. Time complexity analysis describes the situation mathematically by means of the big O notation. For example, insertion-sort and merge-sort are two algorithms for sorting an integer array. The complexity time-frame for the former is $O(n^2)$ and $O(n \cdot \lg n)$ for the latter. In other words, insertion-sort is a quadratic algorithm while merge-sort is linearithmic, a product of linear and logarithmic functions. Obviously, merge-sort performs better when encountering large scale tasks as it demonstrates reduced time complexity. The fragments or structures in FRESH are not initially sorted. Therefore, the FRESH program requires that all fragments and structures be processed individually, resulting in a complexity of at least $O(n)$. Under this circumstance, no fragment or structure can be skipped by adopting algorithms like binary search. The $O(n)$ is the lower bond of the time complexity for FRESH. Therefore, it is impossible to derive an algorithm for FRESH with reduced complexity such as $O(\lg n)$. Enhancing the performance of FRESH by reducing the time complexity is not a feasible solution.

Alternatively, another optimization option can be exploited by reducing the number of fragments/structures to be processed. The combinatorial enumeration process for constructing the initial library is often the rate limiting step during which the fragments generated from the corresponding building blocks are covalently attached to the core structure according to the modular synthesis route. Reducing the number of fragments

before the attachment step can lower the burden of enumerating the initial library. Thus, a number of fragment selection filters are employed prior to the construction of the initial library. One important first-pass filter depends on the creation of a check-list that identifies and eliminates functional groups and substructures that are reactive, unstable in plasma, prone to cause false-positives in screening assays or induce toxicity.²⁶ In addition, the “Rules of Three” for prioritizing appropriate fragments is incorporated as a valuable second filter.²⁷ A third round of fragment selection depends on the specific synthetic route. Functional groups or substructures that produce regio- or enantio-selectivity environments are presently flagged as low priority fragments and are not passed to the enumeration step. Reducing the fragment numbers before enumeration is the most crucial optimization feature for the FRESH program. Without this optimization step, the FRESH program is extremely susceptible to run time errors due to the available time and system resources.

The speed of the numerous programs utilized in the entire FRESH program varies. For example, Pipeline Pilot can process ~ 1 billion drug-like structures per day on a 3GHz Intel i3 CPU. The estimated daily speed for Qikprop is around several million, while MM-GBSA scoring can only handle several thousands. In situations where one or more steps can't process all structures in a reasonable time-frame, the faster calculation steps are arranged ahead of the relatively slower steps to reduce the work load and improve efficiency. In addition, several numerical identification systems are employed along the processing pathway for easy identification and merging operations of specific structures, and filters for removal of duplicate structures are applied at various points to avoid unnecessary repeated calculations.

Altogether, these optimization efforts ensure the smooth execution of FRESH with large scale input. However, for specific projects, new problems may still arise. The user is advised to frequently monitor the displayed numbers on the pipelines of the program and the execution time to discover and avoid problems like insufficient system resources.

2.5. Advantages of FRESH

The advantages of this novel program are obvious. Compared to traditional lead-optimization strategies, FRESH can cover a much larger chemical space within an acceptable time period and at lower cost due to the enhanced speed. With three years of effort spent on combating technical difficulties, bugs, defects and optimization of the prioritizing scheme, the program can now process up to several billion structures in a single run. Recent modifications in which iterations are incorporated will be discussed in details in the following Chapter. At the output port, instead of producing only a single molecule, a complementary set of structures is generated simultaneously, each of which can potentially serve as backup to the others. The candidates provided by the program are also synthesis-friendly, as the library construction originates from the chemists' synthetic routes, and building blocks are obtained directly from a commercial database or other reliable sources. The problem of multi-dimensional optimization mentioned in the last chapter is also addressed, since bioactivity and drug-like physical/ADMET properties are incorporated into the output structures.

Chapter 3: FRESH Validation Case Studies

3.1. Introduction

To validate the FRESH program, three case studies were performed. The purpose of these cases studies is to demonstrate how FRESH can capture at least one of the literature reported potent compounds. The three cases were chosen based on five criteria: (1) the protein target involved in each case has confirmed or potential therapeutic effect. (2) The data in the literature involved in each case should be recent, preferably within 5 years. (3) The compounds are derived from modular synthesis from a core structure. (4) The starting compound or core structure should not already be a potent ligand. In terms of IC_{50} or K_i value, it must be above 100 nM. (5) The literature contains at least one potent compound with IC_{50} or K_i value less than 10 nM. Additional structured query language (SQL) programming was involved to select possible cases for further investigation.

It is important to acknowledge that all QSAR methods, whether they are ligand-based or receptor-based, have their own limitations. The universal QSAR method which can accurately predict the potency of all the unknown compounds currently does not exist. Considering the shortcomings and limitations of each QSAR method, the FRESH selection scheme for these three cases was designed in two stages. In the first stage, all the possible QSAR scores were used to select the compounds based on consensus voting. The consensus scoring method trades some false negatives with false positives. In other words, some potent compounds may be predicted to be inactive during this stage. However, this is not expected to be a serious concern for FRESH considering the large chemical space it probes (thousands to billions of structures). After obtaining a candidate

compound structure list from the first stage and synthesis of the proposed compounds, if a potent molecule is identified, subsequent information derived from such a molecule will contribute to the next iteration stage of FRESH selection and hopefully recover some of the rejected compounds in the first stage.

All three cases require the evaluation of potency. As stated in Section 2.3, the Glide scores and MM-GBSA scores are possible QSAR parameters. A third approach takes ECFP (Extended Connectivity Finger Prints) as the molecular portrait and Bayesian statistics (an estimate of probability based on the presence/absence of specific features from the binary data input) as the modeling method. Unlike Glide or MM-GBSA, this descriptor is a 2D-based one. Intuitively, a 2D-based method should be less accurate. However, it has been suggested that sufficient information may already be available in a 2D structure to enable extraction of a useful QSAR and compete favorably with traditional 3D approaches.^{28,29} In addition, one significant advantage for the 2D-based method is the speed of calculation. The ECFP descriptor is generated orders of magnitude fold faster than Glide or MM-GBSA. Unlike some 3D-based methods (COMFA, Field-based QSAR in Maestro), it does not require pre-alignment of 3D structures. This descriptor is particularly useful in the FRESH program when probing a large chemical space. QSAR selections based on ECFP are placed ahead of the more computationally expensive Glide or MM-GBSA programs to prioritize candidate structures.

Data points for historical compounds which serve as the training and test sets for the QSAR in the FRESH program are collected from the ChEMBL database for the specific protein target.³⁰ The data in ChEMBL is extracted from the primary published literature on a regular basis, then further curated and standardized. It currently contains over 1

million compounds for 5.4 million bioactivity measurements against 5200 protein targets. The database used here to acquire building blocks in the library enumeration step is the “Zinc bb now” database, which is a collection from several suppliers and claimed to be building blocks immediately available.³¹ All chiral molecules were excluded because of the potential concerns for the uncertainty of the chiralities of the compounds in the database, the inability of ECFP descriptors to differentiate chirality and the synthesis challenge.

3.2. Case I: Phosphoinositide 3-Kinase, α isoform. (PI3K α) – Homology receptor model case

3.2.1. Case Background

Phosphatidylinositol 3-kinase (PI3K) phosphorylates the 3-hydroxyl group of the inositol ring of phosphatidylinositol and converts phosphatidylinositol (3,4)-biphosphate (PIP₂) to phosphatidylinositol (3,4,5)-triphosphate (PIP₃). PI3K is a crucial component in a number of pathways, which regulates cell proliferation, survival, chemotaxis and differentiation.^{32, 33, 34} Amongst all the isoforms, PI3K α is an interesting cancer therapeutic target. It has been recognized that the up-regulation of PI3K signaling pathway promotes angiogenesis and is associated with development of human cancers. In addition, it has been implicated in conferring resistance to conventional therapies.³⁵ Currently, there is no FDA approved anti-PI3K drug available.

The present PI3K-based case study originates from the literature by Kim et al. in 2011.³⁶ They have explored the side chains of the particular scaffold in **Figure 4**. The initial hit compound (R₂ = H) is a 360 nM (IC₅₀) agent that emerged from scaffold screening at

a concentration of 10 μ M. Various aromatic substituents on the R₂ group were tested. The R₂ group has been modified by Suzuki coupling.

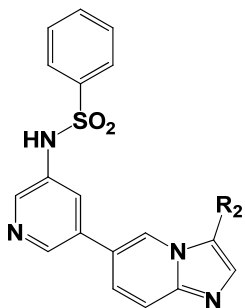


Figure 4. PI3K inhibitor scaffold.

3.2.2. Design of FRESH program, 1st round

All the PI3K α inhibitors in the ChEMBL database were employed to generate training and test sets for establishing the QSAR in the FRESH program to predict potencies of novel chemical structures. To make the retro-study more realistic, two filters were used in this section of FRESH. These excluded all compounds with either the same structure as those in the Kim paper or those that appeared after 2010. Thus, the QSAR derived from the ChEMBL database is based strictly on the available data points at the time of the project.

The performance of a given scoring function is evaluated by the receiver operating characteristic (ROC) curve. The latter curve is constructed by plotting the rate of true positives against the rate of false positives. It demonstrates how well a particular scoring function (For example: Glide score) can differentiate active compounds from the inactives. The performance is measured by the area under the curve or AUC. The closer AUC is to 1, the better. An AUC under 0.6 is generally considered unacceptable.

10 nM was chosen as a cutoff value to differentiate “Active” and “Inactive” analogs for constructing the ligand-based Bayesian model using ECFP as the descriptor. The selected ChEMBL database compounds were divided by a 2:1 ratio for training and test sets to build the QSAR. The training set contains ~500 compounds with ~70 actives and the test set contains ~250 compounds with ~30 actives. The resulting AUC of the ROC curve (**Figure 5**) for the test set compounds is 0.93, which is excellent. Thus, the Bayesian score with ECFP as the descriptor is one of the filters for evaluating activity.

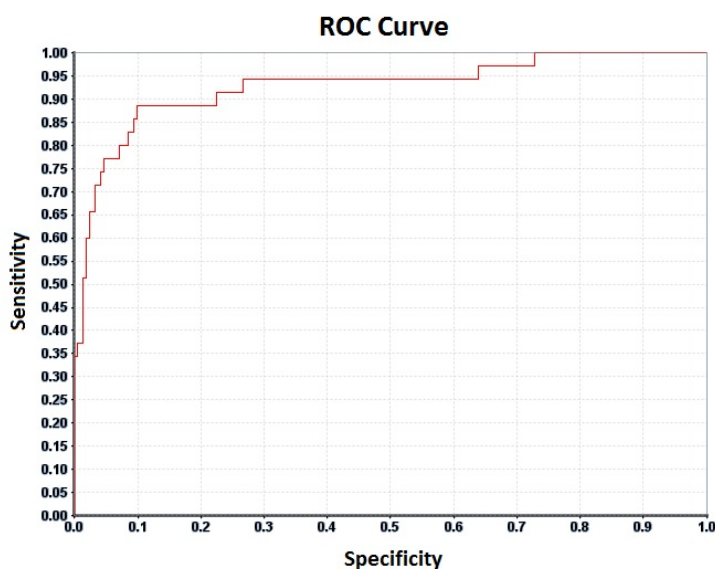


Figure 5. ROC curve for the ECFP method. AUC = 0.93

At the time of this project, there was no crystal structure of PI3K α available. However, the authors mentioned a homology model based on PI3K γ , another isoform of PI3K with 41% sequence identity and 51% similarity at the kinase domain. The original homology model is not available, so an Emory homology model using the same template (PDB code: 1E8Z³⁷) was constructed and used consequently for receptor-based docking analysis. The ChEMBL database compounds were prepared by LigPrep in the Maestro package and processed for both Glide and MM-GBSA scores. The corresponding AUC values of ROC

curves (Figures 6 and 7) are 0.64 and 0.72 (acceptable), so the Glide and MM-GBSA scores were also used as complementary criteria for potency evaluation.

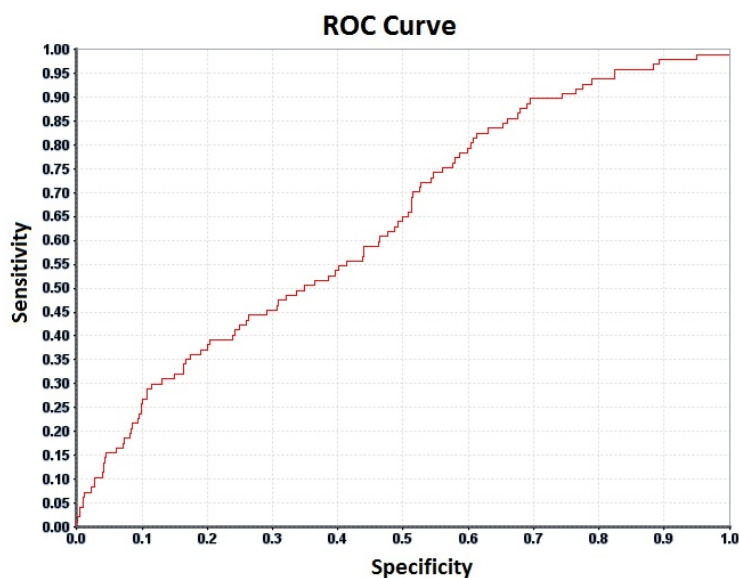


Figure 6. ROC curve for the Glide Score. AUC = 0.64

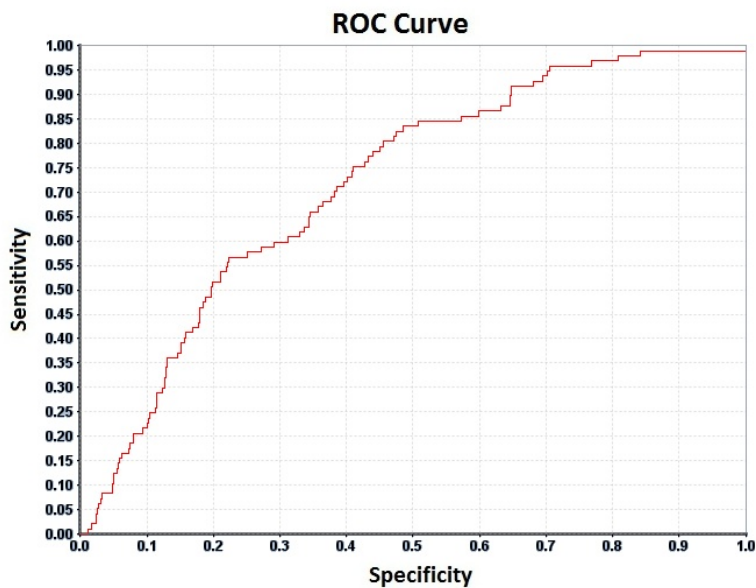


Figure 7. ROC curve for the MM-GBSA score. AUC = 0.72

According to the synthesis route, the aromatic R_2 group is attached to the core structure by Suzuki coupling in which the arylboronic acids (or bromides) are the building blocks.

To obtain the fragment library for enumerating virtual structure libraries, the building block structures were queried against the “Zinc bb now” commercial compound database, and ~44, 000 fragments were initially obtained. The R₂ fragment was first filtered by “Fragment Rules of 3”, groups with potential liability and reactivity concerns as described in Section 2.4 for the crucial algorithm optimization step. The remaining fragments were then covalently connected to the core structure in **Figure 4** to generate structures of all possible PI3K α inhibitors. Subsequently, a series of widely-accepted drug-like filters (Rules of Five, Rules of Three) was applied and structures with desirable drug-like properties were established for further processing. Finally, the Bayesian, Glide and MM-GBSA scores were obtained for the remaining structures. Using the corresponding scores of the hit compound (R₂ = H, 360 nM) as a reference, structures with all three scores better than the hit compound were allowed into the final list for this round. Only ~40 structures remained from this round of selection.

Because the ROC curve of the Bayesian score has the highest AUC, it was used to rank and prioritize the remaining ~40 structures. Upon examination of the prioritized structure list, the 3rd structure appeared as compound **19d** (**Figure 8**) in the literature with an oxadiazole ring at the R₂ group. With an IC₅₀ value at 2 nM, **19d** is a potent inhibitor against PI3K α . It is also featured in the Table of Contents graphic of the paper.

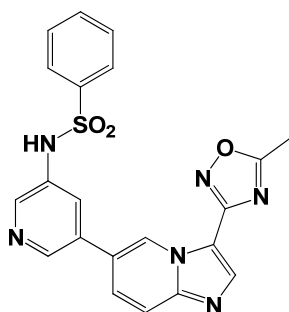


Figure 8. Compound **19d**

3.2.3. Design of FRESH program, iterations

After initial identification of **19d**, the FRESH program then incorporated this newly-available information in order to perform the second round of virtual screening (iteration step). **Figure 9** demonstrates the docking pose of **19d** into the PI3K α kinase domain. From the docking pose, the aryl-side chain points towards the solvent accessible area. This may also explain the improved activity from phenyl (IC_{50} is 720 nM) to the oxadiazole at the R_2 position. Thus, one direction for further exploration is to increase the number of heteroatoms on the 5-membered ring. The tetrazole group with 4 heteroatoms is a reasonable synthetic target that would be a potentially fruitful direction for the medicinal chemist. This structure turned out to be another of the reported potent compounds with IC_{50} of 0.8 nM (**19f**, **Figure 10**).

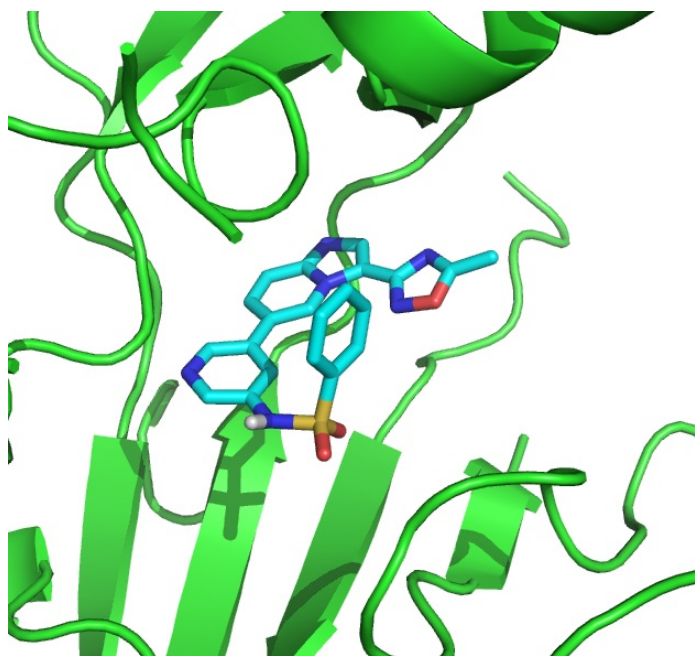


Figure 9. Docking pose of **19d** on PI3K α

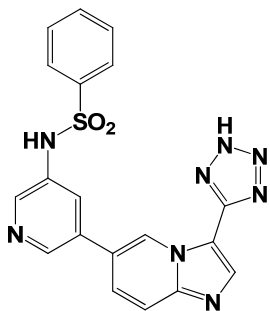


Figure 10. Compound **19f**

However, the polar surface area of **19d** is very close to 120 \AA^2 , and this may explain the relatively low oral bioavailability (24%). Additional heteroatoms are likely to further decrease the cell-permeability as supported by the cell-based anti-proliferation cell-based assay results of **19f** ($IC_{50} > 10 \text{ \mu M}$). An alternative direction to explore for the second round of FRESH is to reduce the number of heteroatoms in the ring. As depicted in **Figure 9**, the docked structure demonstrates that the aryl ring is exposed to the solvent accessible area, so at least one heteroatom should be included at this location in order to retain some activity. Therefore, the decision to pursue a second round of FRESH iteration seeks to investigate the aryl ring with only one heteroatom. This will reduce the polar surface area, while hopefully retaining some activity. The FRESH program was thus designed to identify aryl rings with only one heteroatom. These structures were then ranked for similarities with **19d** and higher similarity structures are prioritized. After a round of iteration, the 5th structure in the list with 4-pyridine substitution at the R_2 group proved to be another potent reported compound (**19k**, **Figure 11, left**) with an IC_{50} value of 4 nM. It is reasonable to believe that in a real drug discovery project, the 3-pyridine compounds would also be pursued along with the 4-pyridine compound, which is another potent reported compound (**19j**, **Figure 11, right**) with IC_{50} of 7 nM.

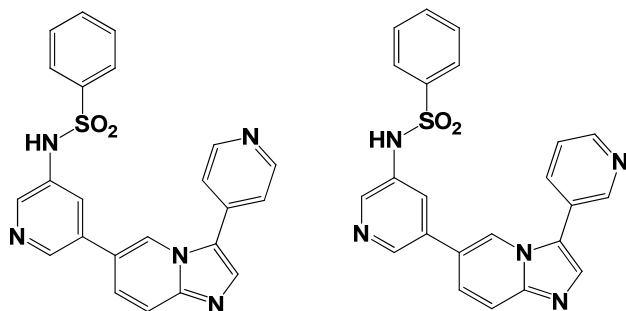


Figure 11. Compounds **19k** (left) and **19j** (right)

In summary, the application of the FRESH program in this case study identified potent **19d** in the first round, while further iterations led to **19k**, **19f** and **19j**. FRESH located both **19d** and **19k** among its top 5 candidates.

3.3. Case II: Carbonic Anhydrase 2 (CA II) – Crystal structure model case

3.3.1. Case Background

Carbonic anhydrases (CAs) are ubiquitously expressed in all organisms. This family of enzymes catalyzes the reversible hydration of CO₂ to bicarbonate and a proton. CAs are categorized as metalloenzymes, as the catalytic center contains a zinc ion. They are involved in many physiological processes, and their inhibitors are explored clinically for various therapeutic effects such as anti-glaucoma, anti-convulsants, anti-obesity, pain-killer and anti-tumor activities. The physiologically dominant isoform, CA II, is one of the most extensively studied proteins among all known protein targets.³⁸

Pacchiano et al. has investigated CA inhibitors with the ureido-benzenesulfonamide scaffold shown in **Figure 12**, which is derived from the popular sulfonamide scaffold.^{39,40}

The R₁ group contains at least one phenyl ring. The corresponding building blocks for the R₁ group are iso-cyanates or acid chlorides.

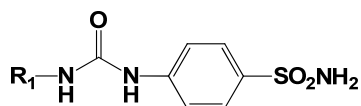


Figure 12. CA II Inhibitor Scaffold

3.3.2. Design of FRESH program, 1st round

The construction of a FRESH program for the 1st round is similar to the previous case study on PI3K α inhibitors. The training and test sets are also obtained from the ChEMBL database. Filters are placed to ensure the QSAR derived from the ChEMBL database is based strictly on the biological data points at the time of the projects. This case study involves two publications; thus, the publication cut off year was selected to be the earlier one (2010).

For the Bayesian model using ECFP as descriptors, the training set contains ~1100 compounds with ~350 actives (10 nM is the cutoff for active) and the test set contains ~560 compounds with ~170 actives. The resulted AUC of the ROC curve (**Figure 13**) for the test set compound is above 0.88, indicating an excellent separation of actives and inactives. The Bayesian score with ECFP as the descriptor was selected as one of the filters for evaluating activity.

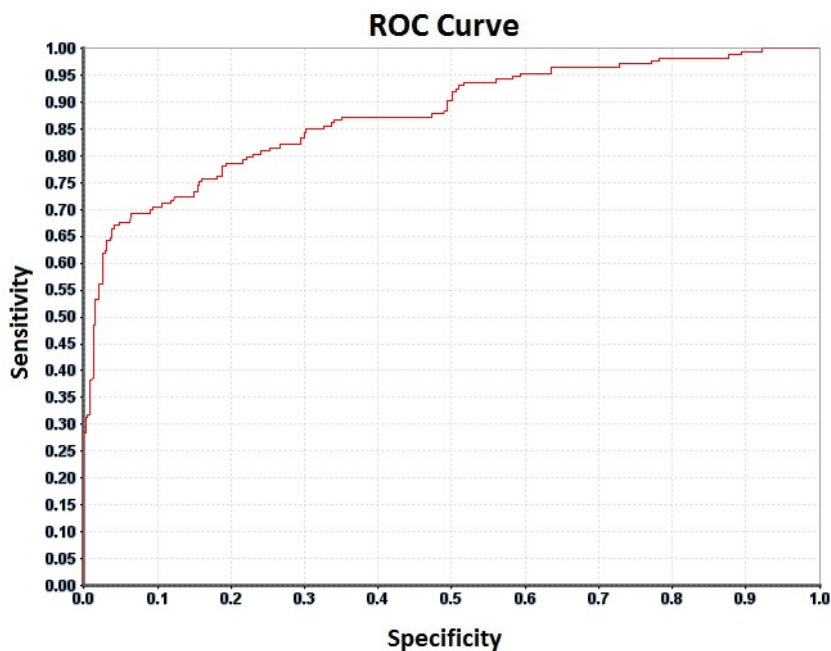


Figure 13. ROC curve for the ECFP method. AUC = 0.88

Unlike the PI3K α case, this protein has many crystal structures available for receptor-based docking/modeling. Therefore, the crystal structure with the publication year prior to 2010 with the highest resolution (PDB code: 1LUG) was chosen to be the receptor for estimating Glide and MM-GBSA scores.⁴¹ In the Glide docking, restraints were imposed to enforce the crucial ligand-receptor interactions with the zinc atoms and Residue 198 within the catalytic pocket. The Glide score gave an acceptable AUC of 0.63 (**Figure 14**), while the MM-GBSA score fails (AUC < 0.6). Therefore, only the Glide score was included in the FRESH program to assess potency.

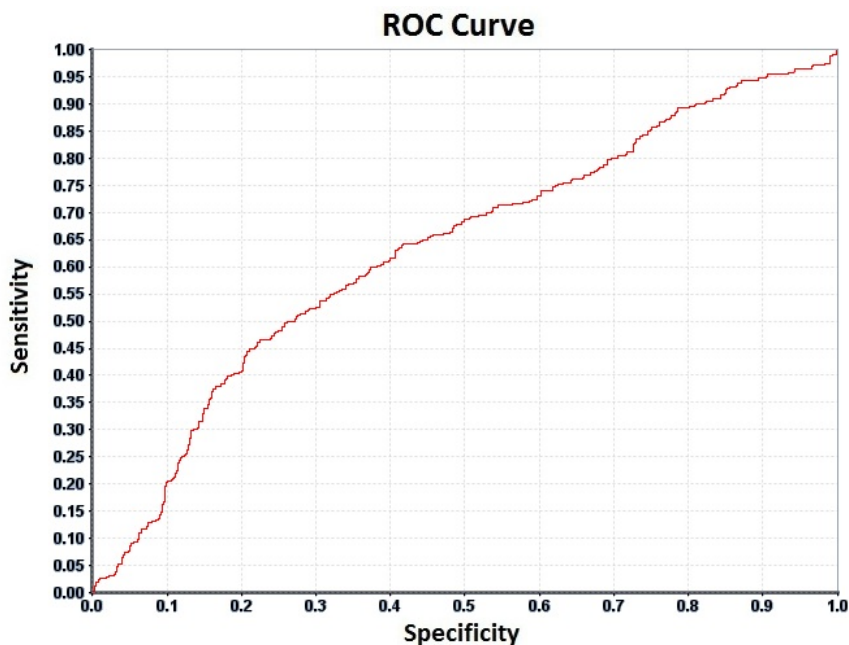


Figure 14. ROC curve for the Glide score. AUC = 0.63

The two publications involved in this case study did not explicitly specify a starting compound to be used as a reference when comparing QSAR scores. However, since the authors mainly investigated the phenyl substitution effect on activity, the un-substituted phenyl compound (K_i is 3730 nM) was selected as the reference structure. The FRESH program was designed to retain structures with both improved Bayesian scores and Glide scores compared to the reference compounds. The same selection scheme for physical/ADMET properties in the previous case study was also applied.

Starting from ~1300 R_1 fragments, only ~40 structures survived the selection scheme. Similar to PI3K α (highest AUC), the Bayesian score was chosen to rank and prioritize the final list of molecules. Examination of the latter for the 1st round of FRESH revealed that **30 (Figure 15)** with a K_i value of 8.9 nM (2011 paper) is the 4th structure in the final list. The program has successfully captured a potent CA II inhibitor.

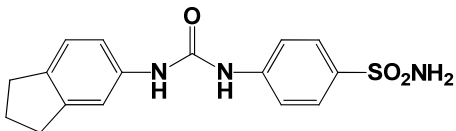


Figure 15. Compound **30**.

3.3.3. Design of FRESH program, iteration steps

The 1st FRESH iteration was constructed on the basis of structure **30**. Compared to the phenyl group, the indan-5-yl group increases the hydrophobicity. From the docking pose shown in **Figure 16**, the indan-5-yl group is directed towards a hydrophobic pocket surrounded by Phe130, Leu140, Leu197, Pro201 and Leu203. A hypothesis was formulated that retaining the hydrophobicity at R₁ group can lead to additional potent compounds. Thus, the iteration was designed to identify structures with retained hydrophobicity (measured by logP of the R₁ group) while retaining similarity to **30**. After examining the search list for this iteration, the 3rd structure was identified as **3** (**Figure 17**) with an IC₅₀ of 3.3 nM as in the 2010 publication. The crystal structure of **3** was obtained by the same research group. As shown in **Figure 18**, the R₁ group points to the hydrophobic pocket as expected.

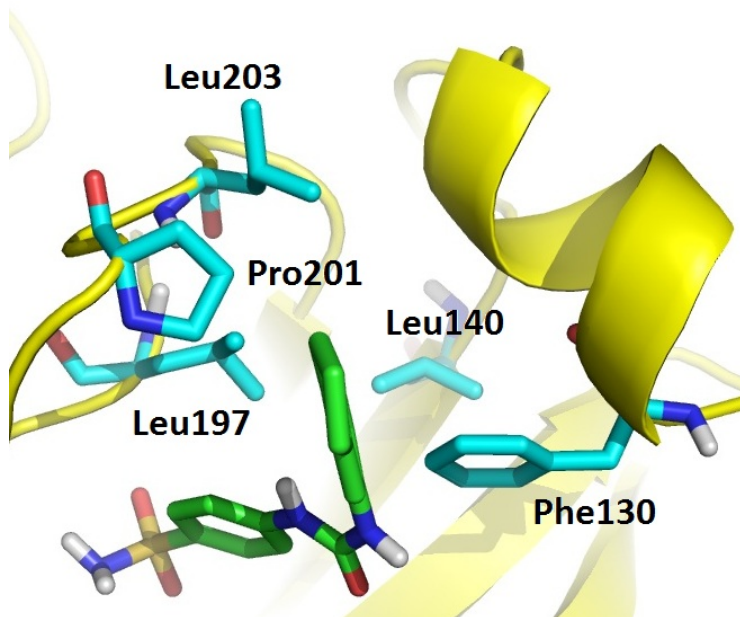


Figure 16. Docking pose of **30** at the catalytic pocket of CA II.

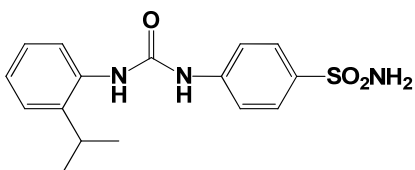


Figure 17. Compound **3**.

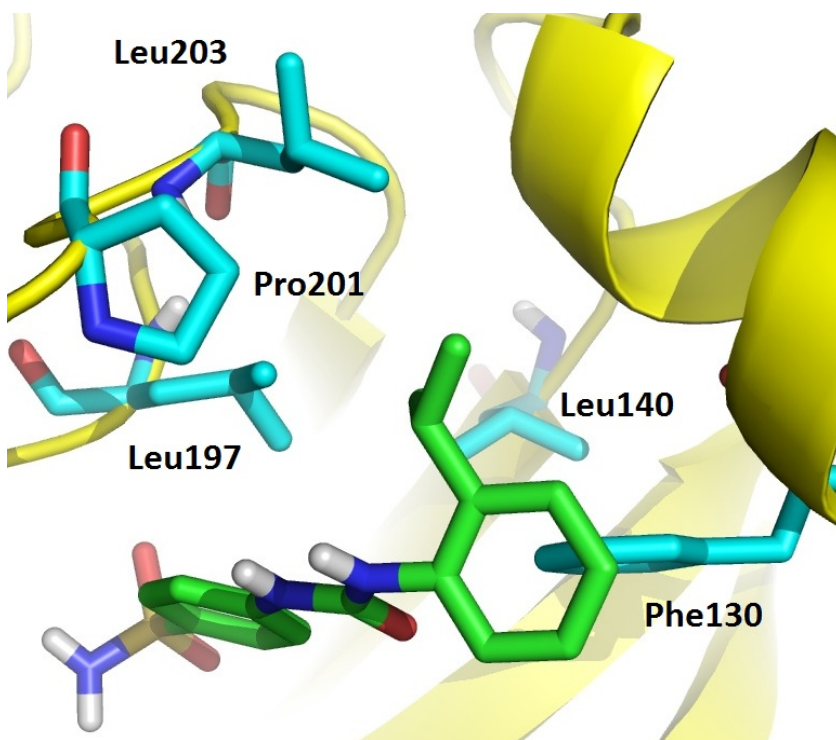


Figure 18. Crystal structure of **3** at the catalytic pocket of CA II.

A second search iteration was performed with the information from **3**. From the crystal structure shown in **Figure 18**, the 2-substitution appears to be the most favorable position to pick up hydrophobic interactions. Thus, all the 2-substituted structures were put through the filtering scheme and consequently ranked by similarity with **3**. The 5th structure on the list is **24** (**Figure 8**) with an IC_{50} of 9.7 nM in the literature.

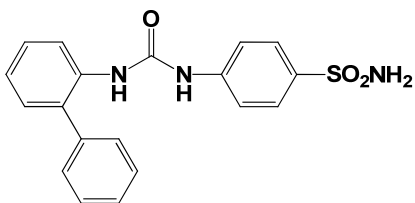


Figure 19. Compound **24**.

In summary, the application of FRESH in this case study has identified a potent compound **30** in the first round, while further searching led to the additional potent agents **3** and **24**. The ranks of all three compounds generated by FRESH analysis are again among the top 5.

3.3.4. Further experiments

To further test whether the FRESH program can provide additional potent inhibitors previously undiscovered by the research group, a collaboration project was initiated. Together with lab members Dr. Thomas Kaiser and Dr Zackery Dentmon, we selected 10-15 additional compounds for synthesis from the output list of FRESH analysis. Compounds which appear in the ChEMBL or PubMed database were excluded to guarantee novelty. The original author of the two publications (Dr. Supuran, University of Florence, Italy) was also engaged in this project to perform the bio-testing using the same method described in the paper.

3.4. Case III: Histone Deacetylase 1. (HDAC 1) – Ligand-only example

3.4.1. Case Background

HDAC enzymes catalyze the removal of acetyl groups from acetylated lysine amino acids on a histone. This allows histones to be more tightly wrapped around the DNA and consequently regulates DNA expression. HDAC enzymes fall in four different classes. Among these, Class I and II isozymes have been associated with uncontrolled tumor growth.⁴² An example knockdown study on HDACs performed by Glaser et al. suggested that HDAC 1 is essential to the proliferation and survival of mammalian carcinoma cells.⁴³ In 2006, FDA approved suberoylanilide hydroxamic acid (SAHA,

vorinostat, **Figure 20**) for the treatment of cutaneous T-cell lymphoma.⁴⁴ This validated HDAC inhibitors as a strategy for cancer therapy.

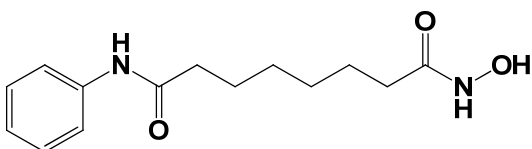


Figure 20. FDA approved HDAC inhibitor SAHA Inhibitor Scaffold

This particular case study is based on the publication by Wang et al. in 2010.⁴⁵ The authors explored the urea scaffold shown in **Figure 21**. Like the SAHA molecule, the right part of this scaffold is a hydroxamic acid group, which is designed to bind to the zinc dication within the catalytic pocket. The length of the aliphatic chain linker varies and the left part of the scaffold consists at least one aromatic ring. The corresponding building blocks for the R_1 group are aryl acid chlorides.

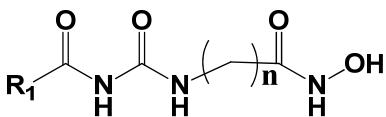


Figure 21. HDAC1 Inhibitor Scaffold

This particular case was chosen as a “ligand-based-method-only case” to demonstrate the usefulness of FRESH under some difficult circumstances where only limited information is available. In addition to the R_1 group, the FRESH program was constructed to vary the linker length from 1 to 7 simultaneously with variation in the R_1 group to intentionally worsen the situation.

3.4.2. Design of FRESH program, 1st round

Construction of the FRESH program for the 1st round was similar to the previous two case studies on PI3K α and CA II inhibitors. The ChEMBL database served as data source

for the training and test sets. The rules of “no presence of identical literature structures” and “no future structures” were strictly enforced. Since only ligand-based methods are allowed in this case study, the Glide and MM-GBSA scores were not applied. For the Bayesian model using ECFP as descriptors, the training set contains ~580 compounds with ~50 actives (10 nM is the cutoff for active) and the test set contains ~290 compounds with ~30 actives. The resulted AUC of the ROC curve (**Figure 22**) for the test set compound is 0.87; again excellent. The Bayesian score with ECFP as the descriptor will be the only filter for evaluating activity. Other filters such as drug-likeness, potential liability groups and Fragment Rules of Three remained the same.

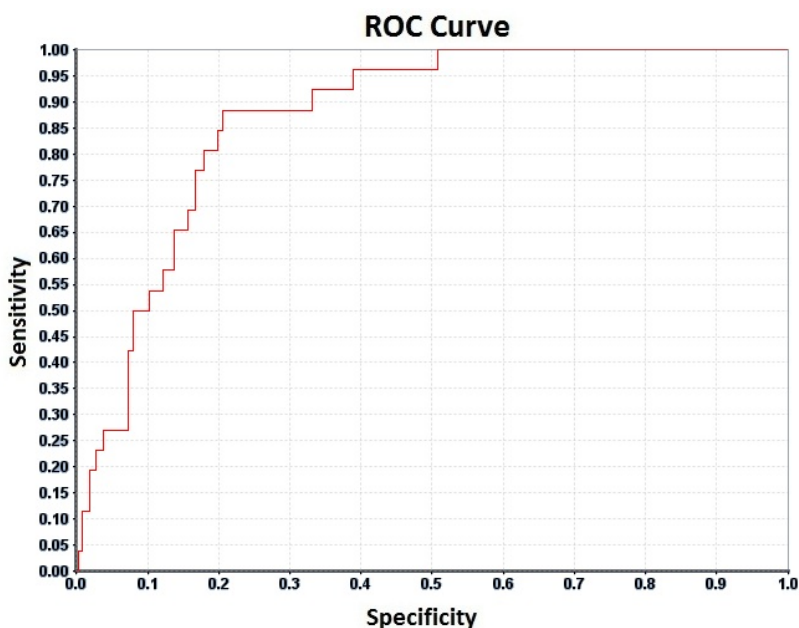


Figure 22. ROC curve for the ECFP method, AUC = 0.87

About 1,200 R₁ fragments were generated from the corresponding commercial building blocks. The value of n varies from 1 to 7, so the total number of possible structures generated by combinatorial enumeration was around 8,400. The structure list for the 1st round was selected by physical/ADMET properties and then ranked and

prioritized by the Bayesian score. The 3rd structure in the list is **5n** (Figure 23) with a reported IC_{50} of 6 nM. **5n** is the most potent one in this publication.

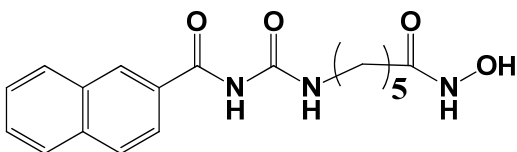


Figure 23. Compound **5n**.

3.4.3. Design of FRESH program, iterations

With the hint from **5n**, structures with 5 carbon atoms as the linker should receive higher priority. Similar to the situation for the PI3K inhibitors discussed in Section 3.2.3, the PSA of this scaffold is already very high (close to 120 \AA^2), so structures with no additional N or O atoms should receive higher priority to maximize the oral bioavailability. The next iteration of the FRESH program for this case was thus designed to collect structures with $n=5$ and without additional N or O atoms. This round led to the identification of **5f** and **5g** (Figure 24) with IC_{50} values of 26 nM and 13n, respectively. They are 4th and 5th on the list. Although **5i** (Figure 25) with an IC_{50} of 8 nM is the 15th on the list, in a real project it would be prioritized after **5f** and **5g** were acquired and tested, since it would complete the heavy halogen series.

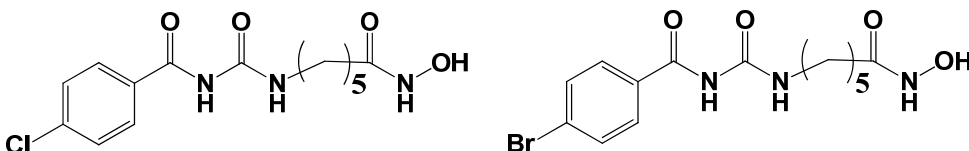


Figure 24. Compounds **5f** (left) and **5g** (right).

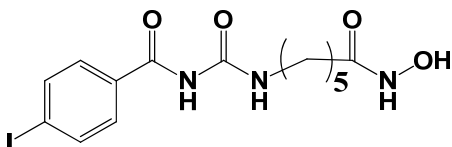


Figure 25. Compounds **5i**.

In summary, the application of FRESH in this case study has identified potent **5n** in the first round. Further iterations led to additional actives **5f** and **5g** with the inclusion of **5i** in the final step. The ranks of all three compounds uncovered by FRESH analysis are all among top 5.

3.5. Conclusion and Future Work

As demonstrated by all three case studies, the FRESH workflow is able to provide a set of synthetic candidates containing at least one potent compound. Iterations based on the information from the newly-identified potent compounds subsequently led to additional potent agents. The ranks of all reported potent compounds are among the top 5 candidates derived by FRESH. One additional finding is that the QSAR approaches based on 2D-descriptors like ECFP can perform equally well or even better than 3D methods.

It is worth noting that there are still certain situations FRESH are unable to handle. In the case of a project where a receptor structure is lacking and not even a single compound has been made previously, there is no way to construct a QSAR for direct incorporation into FRESH. While using bio-datasets for ligands bound a structurally/functionally similar receptor may offer a solution, it introduces additional uncertainties. In this situation, FRESH can only provide candidates with favorable predicted physical properties, which may not be so attractive at this stage. Traditional medicinal chemistry exploration is still needed before a valid QSAR can be generated. In this situation, rather than completely replacing the traditional med-chem exploration, FRESH should be used as a complement to it.

In conclusion, the FRESH program is a useful scheme in assisting drug lead optimization. In all the three case studies, a potent molecule among the top five highly

predictions was identified in the 1st round of virtual screening. Furthermore, iterations flowing from the first identified molecule proved to be valuable starting points for discovery of other potent candidates. The program also proposed directions for improving physical properties. Further work on constructing a universal FRESH program template with ease of use for the synthetic chemist is in progress.

Chapter 4: Application I: Designing novel SNRIs

4.1. Project background

Chronic pain initiated by disease or injured nerves remains a prevalent problem in the US. Johannes et al. performed an Internet-based statistical study and revealed that the prevalence of chronic pain is more than 30% based on ~30,000 adults.⁴⁶ Half of these adults experienced daily pain, and the three-month-averaged pain intensity was severe (≥ 7 on a scale ranging from 0 to 10) for 32%. Lower back pain is the most common form of chronic pain, followed by osteoarthritis pain, Rheumatoid arthritis and migraine headaches. Though the pain itself is generally not lethal, the chronic pain condition obviously reduces life quality and sometimes leads to a debilitating life-style and loss of ability to work. In the US, an estimated \$61.2 billion per year vanishes as a result of lost productivity from chronic pain. In addition, treating patients who experience such debility adds an annual cost of ~\$6,000.^{47,48} Therefore, chronic pain remains a large health, social and economic burden on the US that requires effective medical solutions. Such therapies will not only improve the quality of patients' life but also exert an enormous impact on the entire society.

Repair of damaged nerves continues to be a challenging medical condition, as there is no definite clinical treatment currently available to reverse the damaged tissue.⁴⁹ However, as the condition is generally not life-threatening, medications that alleviate severe pain syndrome are feasible alternative solutions. Opioid therapy, which inhibits the release of neurotransmitters and subsequently blocks the corresponding pain signaling pathway, has been adopted and remain an effective therapy against severe pain. However,

addiction and drug abuse are major problems for the opioid-based therapy. In addition, the potential threat from easy access to prescription opioids for illegal use is not trivial. Medical needs for alternative chronic and severe pain therapies beyond opioids are still unmet.

Monoamine transporters like the norepinephrine transporter (NET) and the serotonin transporter (SERT) are another class of potential therapeutic targets against neuropathic pain. As shown in **Figure 26**, these transporters pump the corresponding monoamine neurotransmitters in the synaptic clefts back to the presynaptic neurons.⁵⁰ By blocking these transporters, the amount of neurotransmitters in the synaptic clefts increases. This consequently induces stronger signals transmitted to the postsynaptic neurons and subsequently suppresses the pain pathways.⁵¹ Previous studies have revealed that NET inhibitors are effective against neuropathic pain. In addition, the simultaneous inhibition of SERT appears to enhance the efficacy, although the SSRIs (selective serotonin reuptake inhibitor) alone are not effective against pain.⁵² A therapy targeting both NET and SERT may offer an alternative solution to the opioid-based therapy.

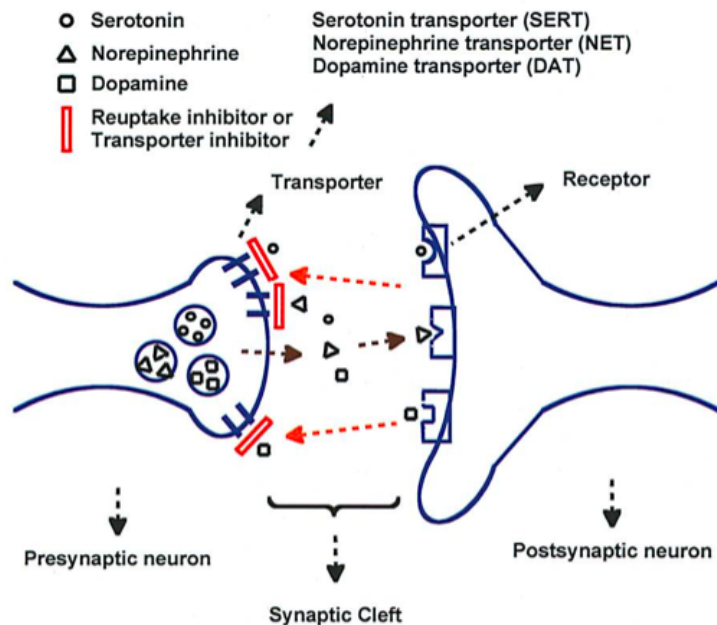


Figure 26. Mechanism of monoamine transporter/reuptake inhibitors.

Administering separate inhibitors for NET and SERT by a cocktail therapy, as discussed in Section 1.3.1, can be problematic due to different PK/PD properties of each component and the potential drug-drug interaction threat. Therefore, a dual target ligand, which inhibits both NET and SERT simultaneously, would be more favorable. Several dual NET/SERT inhibitors (referred to as serotonin norepinephrine transporter inhibitor or SNRI below) have demonstrated efficacy against chronic pain, one of which is milnacipran (**Figure 27**).⁵³

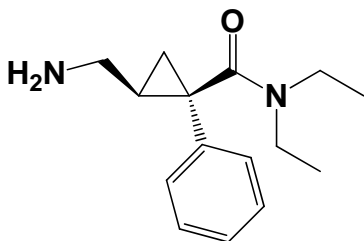
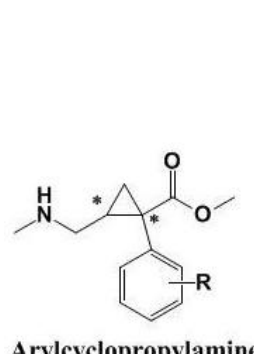


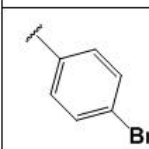
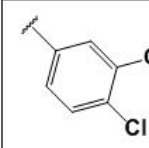
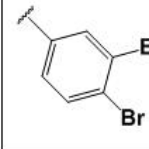
Figure 27. Milnacipran.

However, whether an optimal ratio of NET to SERT potency exists has not been determined yet. Dr. Davies' group at Emory University has synthesized a series of arylcyclopropylamine compounds (**Table 1**) based on the milnacipran scaffold.⁵⁰ These analogs demonstrated significant potency for both NET and SERT. The intended plan is to use them as guide to obtain additional analogs to probe the effect of potency ratio. The FRESH program was applied in this project to provide synthetic collaborators with specific arylcyclopropylamine candidates with suitable variations of the potency ratio of NET/SERT while retaining the overall potency.

Table 1. Six arylcyclopropylamine compounds with clogP and IC₅₀ values.



Arylcyclopropylamine

R group	Name	Chirality	MW	cLogP	NET Ki (nM)	SERT Ki (nM)	SERT to NET ratio
	1	SS	298	2.2	29.9	630	21.1
	2	RR	298	2.2	1.66	31	18.7
	3	SS	288	2.8	5.3	27	5.1
	4	RR	288	2.8	0.2	1.6	8.0
	5	SS	377	3.0	3.3	13	3.9
	6	RR	377	3.0	0.45	1.22	2.7

4.2. Challenges in the lead optimization step

Substitutions on this particular arylcyclopropylamine scaffold readily remain within the chemical space of drug-likeness. For example, the MW of compound **1** is ~300 and the value of logP is 2.2. There is still sufficient chemical space available to accommodate

additional modifications before the structure reaches the “drug-likeness” threshold. However, the main focus of this project at the moment is the potency (measured by K_i) ratio of antagonists bound to SERT and NET receptors. The team is interested in a series of potent compounds that probe a wide variation of potency ratios in order to identify an “ideal” ratio for treating neuropathic pain. The knowledge-based rational design method would experience difficulty in predicting compound structures with specific ratios. By contrast, the FRESH program implements systematic and focused screening that can readily identify an easily synthesized library of novel analogs with the ability to scan a wide range of SERT/NET ratios.

4.3. Receptor-based QSAR models of NET and SERT

At the time of this work, no crystal structure of NET or SERT was available. An alternative strategy to obtain a receptor structure for the Glide and MM-GBSA programs is to utilize homology models. Ravna et al. established such homology models in 2009 for both NET and SERT based on a Leucine transporter.⁵⁴ The sequence similarity between NET and the Leucine transporter is 39% and for SERT 35%. These two receptor structures provided by the homology models, together with the preliminary biological data listed in **Table 1** provide a good starting point for a localized QSAR model. Initial attempts to develop such models were made by Dr. Spandan Chennamadhavuni in the spring of 2011 before his Ph.D. dissertation defense. However, the modeling work had not proceeded to the development of a quantitative QSAR model suitable for FRESH.⁵⁰

The homology modeling study performed by Ravna et al. revealed two possible binding sites (Site 1 and Site 2) on both the NET and SERT homology models. Site 1 can

accommodate relatively small molecules like cocaine, while Site 2 generally accommodates larger tricyclic compounds. Site 1 was chosen to initiate the modeling work.⁵⁴

However, an initial attempt to dock the ligands to the Site 1 pocket using the standard version of Glide docking program failed to yield any pose. The flexibility of the receptor structure provided by the default scale-factor of 0.85 of the Glide program was not enough to perform the docking. Induced-fit docking program included in the Maestro package was considered to provide the protein receptor side chains with additional flexibility. However, the induced-fit docking program is a computationally expensive method. While it may be useful for a specific data set of structures, it is definitely less ideal for the vHTS step involved in the FRESH program. In addition, the use of induced-fit docking also introduces a technical problem: the induced-fit docking output pose file contains only one receptor and one ligand and manual separation steps are required before subsequent application of other programs. No manual operation is suitable for the highly automated FRESH program operated on large scale data sets.

With all these considerations, a decision was made that induced-fit docking would be used to generate docking receptors only once. After this step, the resulting receptor structure was submitted for standard Glide docking. This two-step strategy captures a degree of protein flexibility without sacrificing library processing speed required by the FRESH program. Among the ligands in **Table 1**, the two bulkiest ones, **5** and **6** were chosen for induced-fit docking to maximize the flexibility. For each compound, only one pose was requested. The receptor structures derived from the resulting poses were

extracted and subsequently used as the receptors for the standard Glide docking procedures.

The poses generated by the Glide docking were then rescored by the MM-GBSA program. The top pose for each ligand was submitted for QSAR analysis. Construction of a linear regression based on the equation $\Delta E = -RT \ln K_i$ was attempted. Note: the ΔE term represents the MM-GBSA score and the K_i term is the experimental K_i values.

Figures 28 to 31 illustrate correlations of ΔE to the experimental K_i values. The comparison between **Figures 28 and 29** (data points for the NET receptor) reveals that the receptor structure derived from **6** delivers better correlation (judged by the R^2 value) than that from **5**. Similar conclusion can also be drawn by comparing **Figures 30 with 31**. Consequently, receptor structures derived from the induced-fit docking of Compound **6** were employed for subsequent analysis in the FRESH program. It is worth mentioning that these correlations were based on preliminary results in which only six ligands are involved. Nevertheless, they provided a good starting point for generating prospective compounds for further synthesis.

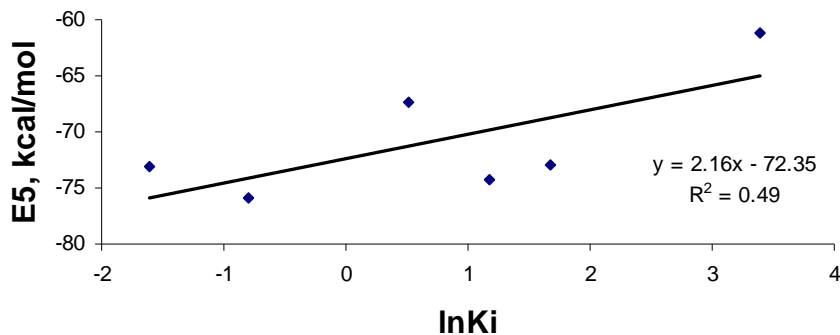


Figure 28. Correlations of estimated binding NET affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of **5**.

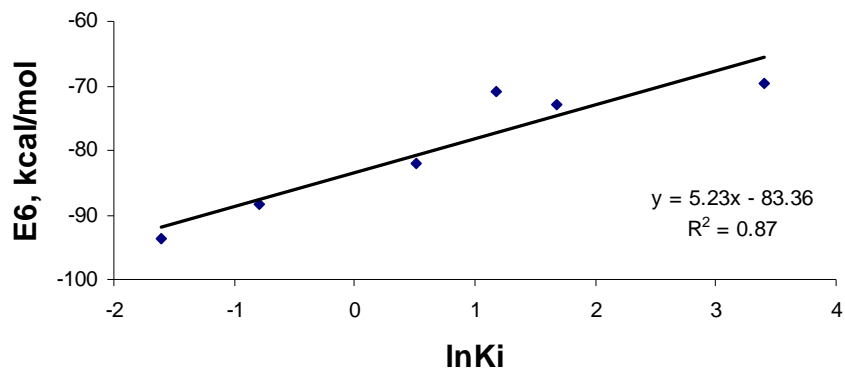


Figure 29. Correlations of estimated binding NET affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of **6**.

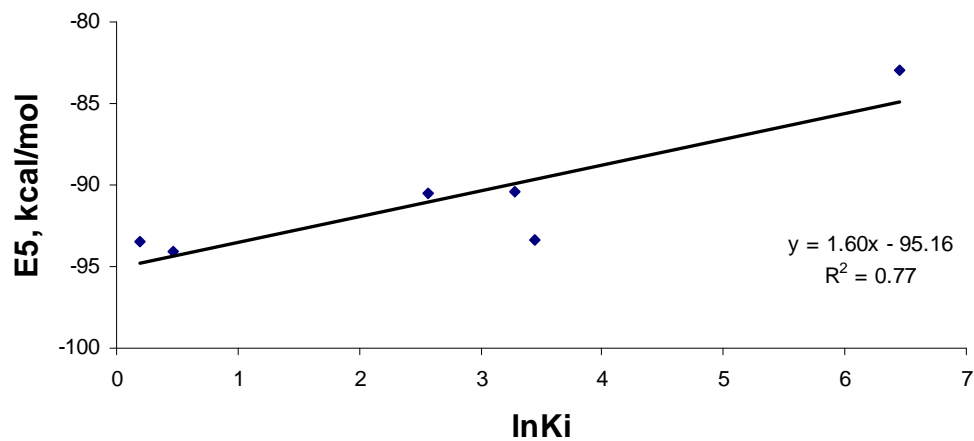


Figure 30. Correlations of estimated binding SERT affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of **5**.

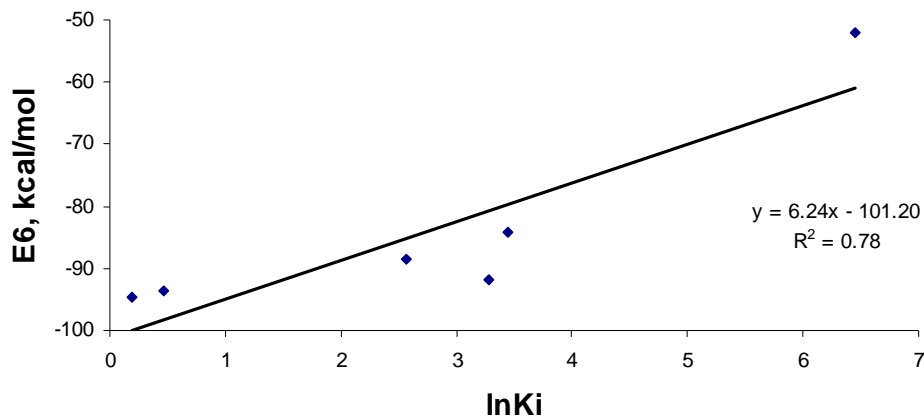


Figure 31. Correlations of estimated binding SERT affinity with experimental K_i for six ligands. Docking receptor was generated from induced fitting of **6**.

4.4. Application of the FRESH program

4.4.1. The FRESH Program design

The FRESH program for this project was constructed in accordance with the synthetic route in **Figure 32** provided by Dr. Davies' group. The R_1 and R_2 groups originate from a carbene precursor which contains an aromatic ring (R_1) and an ester group (R_2), while the R_3 group is from the amine building block. Before the project was stalled by lack of federal funding, the primary interest was the bio-effect of R_2 . Therefore, R_2 was chosen to serve as the only structural variable in this round. Accordingly, the corresponding building blocks with phenyl or pyridine rings were screened. The results in **Table 1** demonstrate that the RR ligands show better activity. Thus, these analogs were assigned a higher priority and investigated first. The corresponding core structure in this round of FRESH is RR.

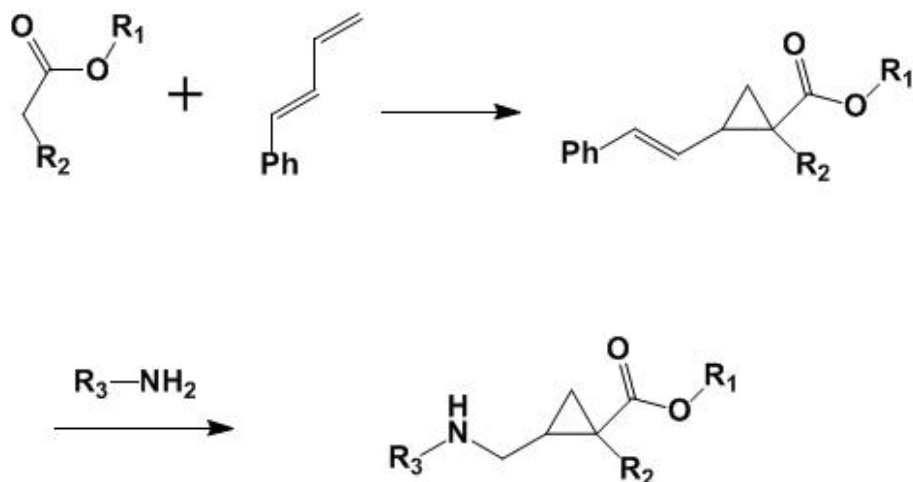


Figure 32. Synthetic route for the arylcyclopropylamine analogs.

Figure 33 demonstrates the general outline of the FRESH program. Starting from the commercial library, the corresponding building blocks were screened and the R_2 fragments were extracted. As stated in Section 2.4, a sub-protocol was placed before the enumeration of the initial library to prioritize fragments without unfavorable functional groups. Similarly, physical/ADMET property selections were also applied in the program to prioritize molecules with drug-like properties. The K_i values for potential NET and SERT antagonists were predicted using the correlations in Section 4.3. At this stage, the cutoff for the predicted K_i value for both SERT and NET was set at 50 nM.

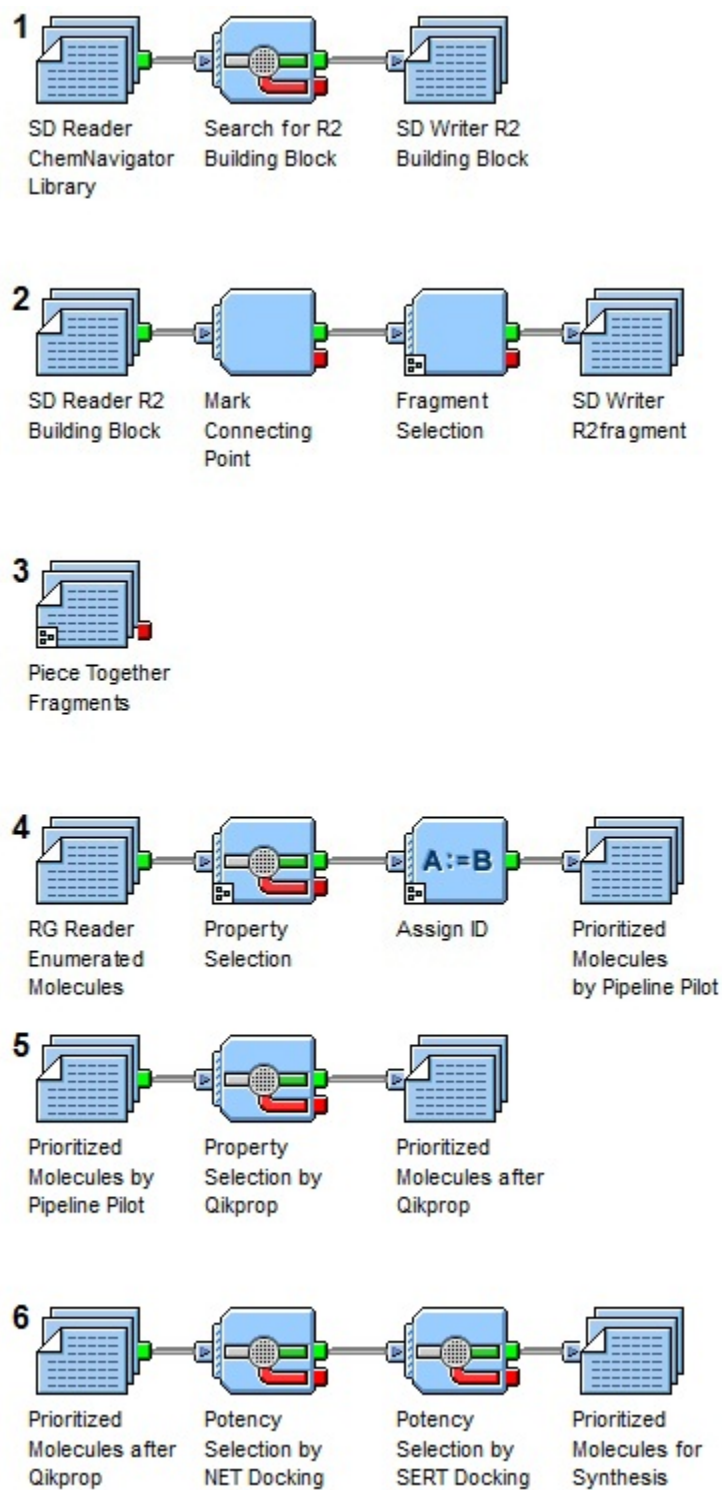
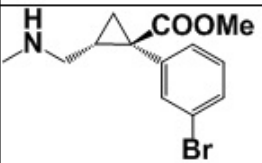
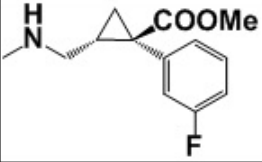
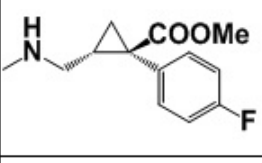
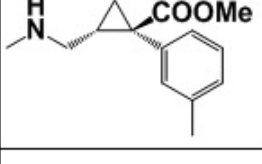
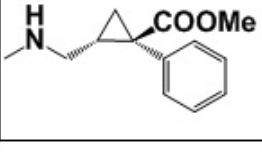


Figure 33. The FRESH protocol (main interface) for prioritizing arylcyclopropylamine analogs. Some sub-protocol components are not shown.

4.4.2. Resulting structures

FRESH identified ~30 output structures at the final output port. After discussion with the collaborators in Dr. Davies' laboratory, five structures were chosen as synthetic targets. **Table 2** depicts the selected structures with the corresponding predicted SERT/NET ratio. The predicted SERT to NET ratio values were around 3 which is at least 10 time higher than the corresponding ratio of Milnacipran.

Table 2. Structures selected to pursue synthesis and the predicted result

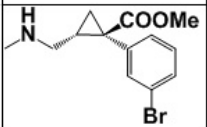
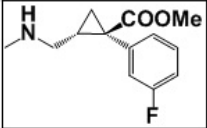
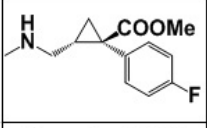
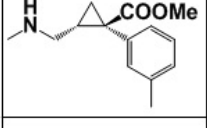
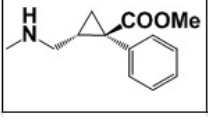
Structure	Name	Predicted NET Ki (nM)	Predicted SERT Ki (nM)	Predicted SERT to NET ratio
	8	1.2	2.5	2
	68	2.0	3.7	2
	69	1.5	6.0	4
	140	2.7	8.1	3
	155	2.2	12.9	6

4.4.3. Test result and comparison

The experimental K_i values for both the NET and SERT were later obtained and listed in **Table 3**. Among the five compounds in the table, **69** is the best match with both

predicted K_i values and the predicted ratio within 2-folds of the experimental ones. **68** also matches the experimental values reasonably well, providing both K_i values within 10-fold of the experimental ones. **8**, **140** and **155** deliver either one or two K_i values outside the 10-fold prediction window. Nonetheless, the predicted ratio remains within 10-fold that of the experimental values.

Table 3. Comparison of predicted results to experimental results.

Structure	Name	Predicted NET K_i (nM)	Experimental NET K_i (nM)	Predicted SERT K_i (nM)	Experimental SERT K_i (nM)	Predicted SERT to NET ratio	Experimental SERT to NET ratio
	8	1.2	15.4	2.5	4.8	2	0.3
	68	2.0	13.7	3.7	11.0	2	1
	69	1.5	2.9	6.0	5.2	4	2
	140	2.7	34.7	8.1	114.0	3	3
	155	2.2	48.3	12.9	39.5	6	1

4.5. Conclusion and future direction

The FRESH program has provided at least two desired candidates out of 5 for this round of structure prediction. Considering the fact that the QSAR models were constructed on just six historical data points, the target receptors are rigid homology models and a revised “induced-fit” method was adopted, the performance of this particular FRESH analysis is unexpectedly successful.

Further work on the FRESH program is needed to improve performance, mainly on the QSAR model component. Additional QSAR models based on alternative methods might well reduce the chances of “false positives”. For example, given the fact that the K_i values are sensitive to the substitution patterns on the phenyl ring, 2D ligand-based methods, like those involving Hammett sigma constants, are worth investigating. 3D ligand-based methods like COMFA, which use calculated electrostatic and steric terms as molecular descriptors are also possible candidates for inclusion in the FRESH program. Additional biological data can also be included in the QSAR training/test sets as derived from other compounds. For example, the ChEMBL database contains several thousand data points for both SERT and NET.

Additional structural modifications of the ligands are also required to minimize the appearance of potential problems in the context of oral bioavailability, although intrathecal administration (IT, injection into the spaces surrounding the spinal cord or brain) was the route of administration at that time when the project was active. Two moieties in the current lead candidates probably need further modification before the compounds qualify as potential orally-available drugs. Thus, the ester bond can be hydrolyzed in a living organism resulting in formation of a carboxylic acid with a low capacity to cross the BBB, an undesired outcome. The amine moiety, which is basic, also presents a barrier for BBB penetration. This usually requires additional FRESH program for other side chain modifications to probe additional chemical space.

Chapter 5: Application II: Identification of novel KCN1 analogs to block the p300/KCN1 interaction

5.1. Project background

Hypoxia is a condition in which tissues suffer from insufficient oxygen supply. It is a prevalent phenomenon in solid tumor tissues due to inadequate development of the vascular blood supply.⁵⁵ The hypoxia inducible factor (HIF) pathway is found to be crucial for tumor cell growth under hypoxic conditions, and hypoxic tumors are associated with resistance to chemo- and radio-therapies.^{56,57} Along the HIF pathway, the HIF-1 α subunit associates with the CH1 domain of cofactor p300 to form a functional transcription factor.

The p300/HIF-1 α complex, therefore, has become a potential therapeutic target against hypoxic cancer. Our collaborators at Georgia State University (GSU) have synthesized a series of aryl sulfonamide analogs, which demonstrate sub-micromolar potency against the p300/HIF-1 α interaction.⁵⁸ These analogs were derived from the lead compound KCN1 (**Figure 34**). FRESH was applied in this project to perform lead optimization. Potential synthesis candidates from the program are anticipated to have improvements in potency against the p300/HIF-1 α interaction and to introduce ADMET characteristics into the target compounds. Since one of the targeted cancers is glioblastoma, it is desirable that molecules also possess favorable CNS drug-likeness properties.

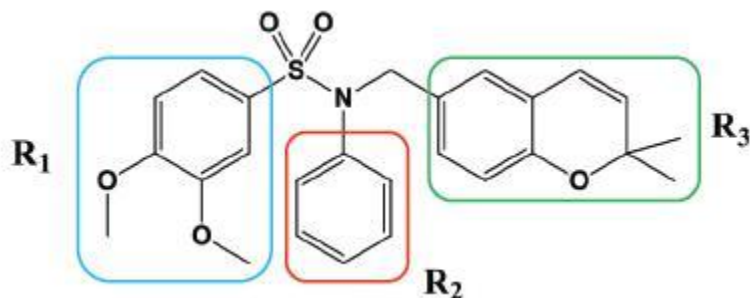


Figure 34. Structure of KCN1 with three highlighted substituent groups

5.2. Challenge in the lead optimization step

KCN1 was the lead molecule at the time of the project. However, if evaluated by the “Rules of lead-likeness”, it is clearly a violator. The MW of 465 is already close to the suggested cutoff of 500 from the Lipinski Rule of Five. This MW value is significantly higher than the suggested cutoff of 350 for lead-likeness, let alone the general trend for CNS drugs to possess a smaller MW.⁵⁹ Therefore, in terms of ligand efficiency, KCN1 is not an ideal lead with capacity for alternative substituent decoration in the lead optimization process.

The experimental IC₅₀ value for KCN1 is approximately 650 nM.⁵⁸ If the targeted IC₅₀ range is 5 nM or less, it requires an improvement of ~100 fold. Even considering the 500 MW threshold from the Lipinski rule, there is little chemical space available for exploration of additional heavy atoms. For some small molecules it is possible reach an approximate 1.5 kcal/mol ligand efficiency. However, as the molecular size increases, the likelihood of encountering one such “magic methyl” decreases rapidly.⁶⁰ Thus, there is a remote possibility that by simply adding extra groups, the IC₅₀ value can be improved by ~100 fold while the MW still remains within the 500 amu window.

In addition, as pointed out in Section 1.1.1, the lipophilicity tends to increase during the lead optimization process.⁵⁹ A desired cLogP cutoff for a lead molecule is therefore suggested to be 3 to ensure the final clinical candidate can stay within the threshold of 5.⁵⁹ However, the cLogP value for KCN1 is already close to 5. In fact, previous attempts to use structure-based methods alone failed to provide structures with a predicted potency increase while retaining the required properties and ease of synthesis.

5.3. Receptor-based QSAR approach

5.3.1. Binding receptor selection¹

The FRESH program requires a QSAR scoring function to estimate the potency of novel structures. Initial attempts were made to develop a receptor-based model. In idea is that KCN1 can interrupt the p300/HIF-1 α interaction by binding separately to p300 or HIF-1 α , or by interfering with the p300/HIF-1 α complex formation to attenuate its function. De Guzman et al. has performed an NMR study on the two proteins and revealed that p300 is able to maintain its 3D-architecture without the presence of HIF-1 α , while HIF-1 α is disordered when uncomplexed with p300.⁶¹ Therefore, a reasonable hypothesis is that the binding partner of KCN1 is p300.

A series of experiments by collaborators at the Winship Cancer Institute (WCI) at Emory and GSU were conducted to test the hypothesis.⁶² **Figure 35** demonstrates the results from an affinity pull-down analysis (Dr. Erwin van Meir and colleagues; WCI). The input lane and the – lane were control groups: The former used a fraction of the cell extract before being pulled down to verify protein expression; the latter employed

¹ This section introduces the work from my collaborators and is directly from my master thesis.

uncoupled beads for non-specific binding (- lanes). The figure illustrates that while KCN1 can pull down a fraction of p300, no detectable interaction between KCN1 and HIF-1 α is observed.

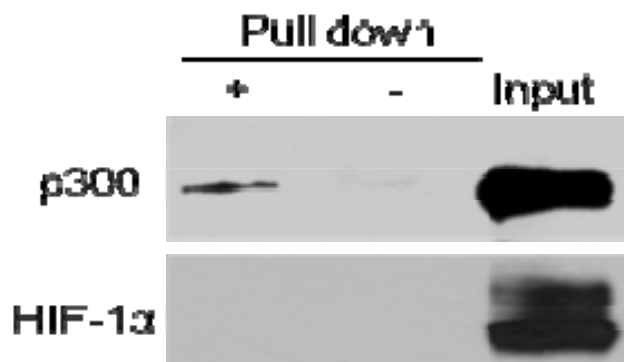


Figure 35. Affinity pull-down analysis of p300 and HIF-1 α proteins using KCN1-coupled agarose beads.

A radio labeling experiment has also been performed. Recombinant fusion peptides, which contain Glutathione S-transferase (GST) and the CH1 domain of p300 (GST-p300-CH1) were incubated with ^{14}C -KCN1. In **Figure 36** (left), the bound activity for GST-p300-CH1 was shown to be significantly greater than that obtained with GST-only peptides. The size of GST-p300-CH1 was verified by Coomassie stained gel as shown in **Figure 36** (right). Once again, this experiment demonstrates the direct interaction between KCN1 and p300.

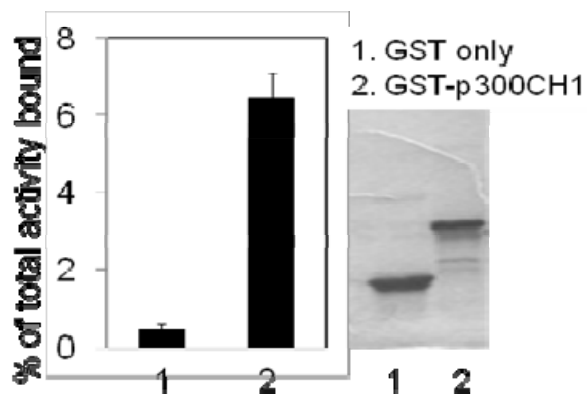


Figure 36. ^{14}C -KCN1 binding experiment result.

Another experiment based on surface plasmon resonance measurements was performed by the GSU portion of the team under the guidance of Dr. Binghe Wang to certify the direct KCN1-p300 interaction. KCN1 was attached covalently to a gold surface (**Figure 37**). The p300-CH1 peptides were streamed over the surface in a series of concentrations as illustrated by **Figure 38**. SPR signals show response to the p300-CH1 peptides and changes with variations of concentration. Analysis of the curve shapes and concentration dependence results in a K_d value of ~ 345 nM for KCN1 binding to p300-CH1 which is comparable for the bio-assay value.⁵⁸ Altogether, these findings have supported the hypothesis that KCN1 can bind to p300-CH1. Based on these results, a binding model of the p300-CH1/KCN1 complex was initiated.

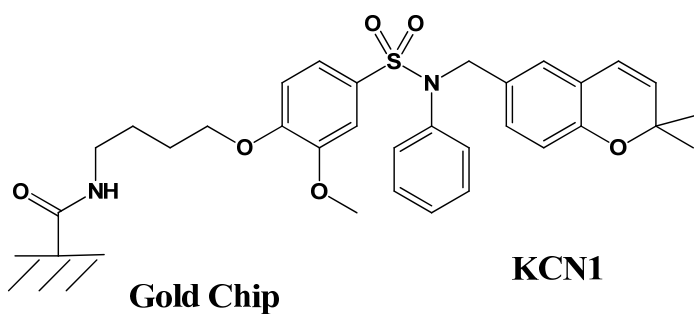


Figure 37. KCN1 attached to the gold surface.

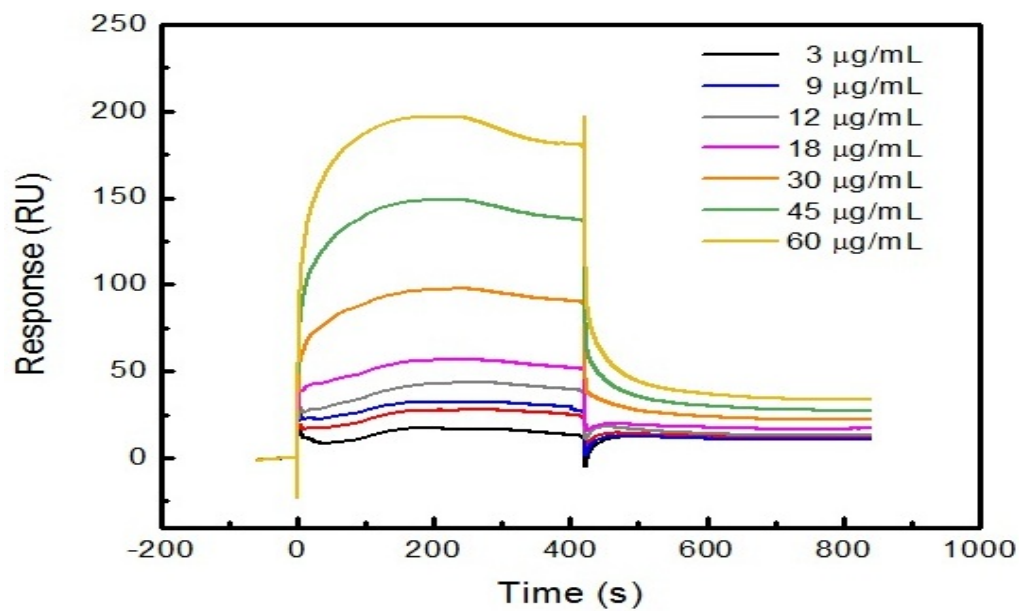


Figure 38. SPR sensorgrams for KCN1 binding to p300.

5.3.2. Binding site selection

Although the results in the previous section support the hypothesis that KCN1 binds to p300, they do not provide additional binding site information. In addition, no crystal structure of a small molecule in complex with p300 is available in the PDB database at the time. Therefore, the possible binding sites were derived from the structure of p300/HIF-1 α complex (PDB code: 1L3E) and a previously reported mutagenesis study.^{63,64}

Figure 39 illustrates the structure of the p300-CH1 domain extracted from the p300/HIF-1 α complex. The p300-CH1 structure consists of three major helices which are nearly perpendicular to each other. Four residues on p300 (Leu344, Leu345, Cys388 and Cys393) were determined to be crucial for forming the complex with HIF-1 α according to the random mutagenesis study by Gu et al.⁶⁴ Among the four residues, Cys388 and Cys393 form a zinc finger and coordinate to a zinc ion. Consequently, disruption of the p300 zinc finger interrupts the p300 structure and consequently prevents the p300/HIF-1 α

interaction. Leu344 and Leu345 are located in the region where the three helices are adjacent to one another. Therefore, four possible clefts, at which KCN1 can interact with at least 2 helices, were investigated (**Figure 41**, left). Each was subjected to KCN1 Glide docking followed by Prime MM-GBSA rescoring. **Figure 41** (right) depicts the top two sites with the best predicted binding energy values. These two sites are hypothesized to be the possible binding sites for KCN1.

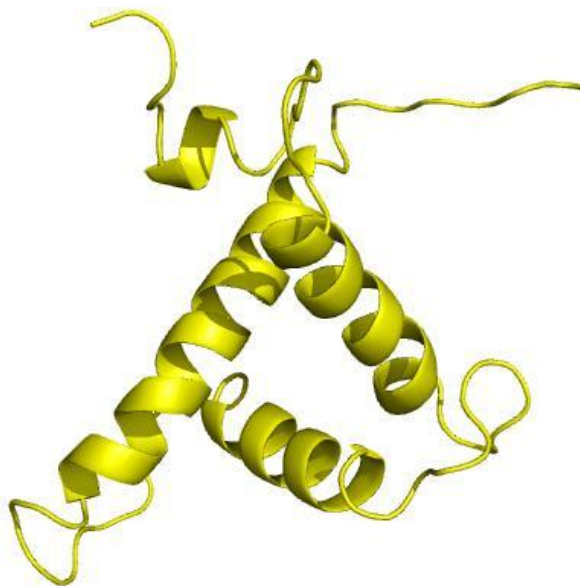


Figure 39. p300-CH1 extracted from the complex.

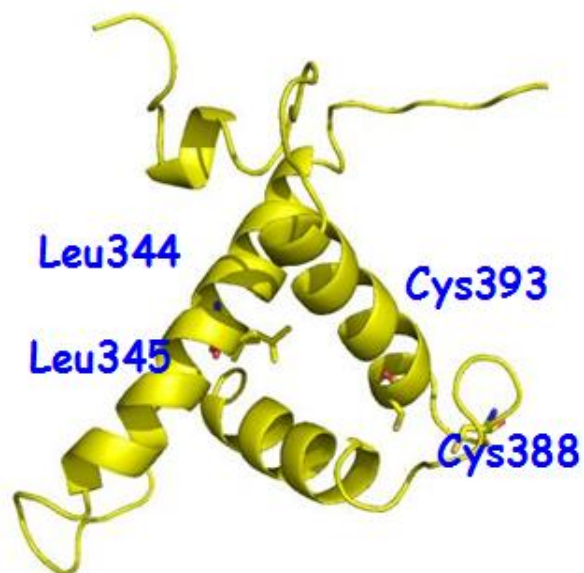


Figure 40. Crucial residues Leu344, Leu345, Cys388 and Cys393 on p300 CH1. Leu344 is hidden under the helix behind Leu345.

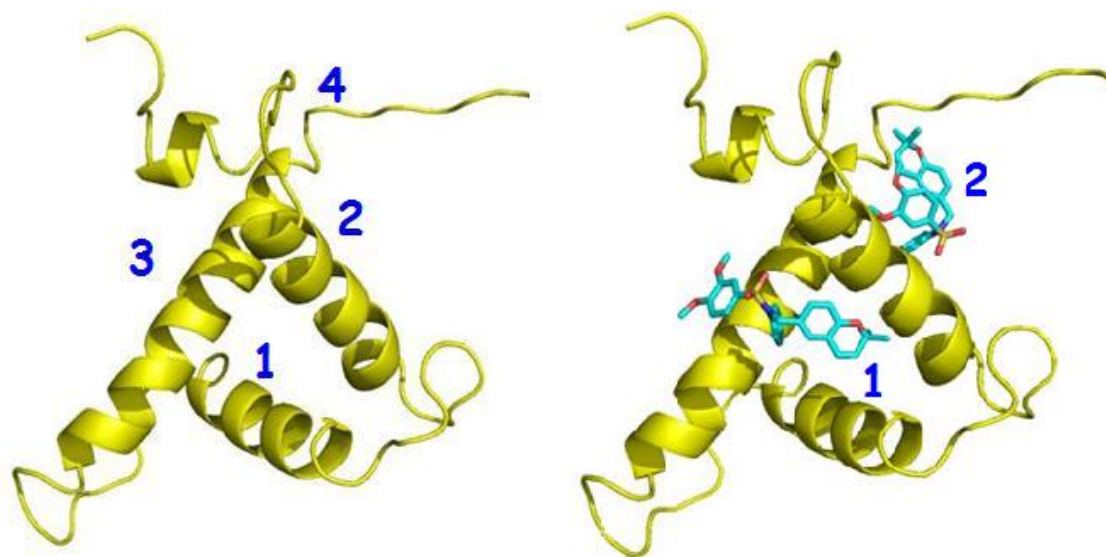


Figure 41. Four clefts chosen for the docking sites(left) The top two sites with docked KCN1 (right)

Remarkably, the top two best-scoring sites (Sites 1 and 2) are coincident with the binding locus of the two HIF-1 α helices (**Figure 42**). Furthermore, the mutagenesis study revealed four crucial residues on the HIF-1 α part, namely Leu795, Cys800, Leu818 and Leu822.⁶⁴ Coincidentally, Leu818 and Cys822 are located on helix A, while Leu795 and Leu800 are found on helix B (**Figure 43**). The coincidence provides further confidence

that KCN1 is likely to bound to p300-CH1 at these two sites. Accordingly, these two centers have been employed in the following receptor-based docking study.

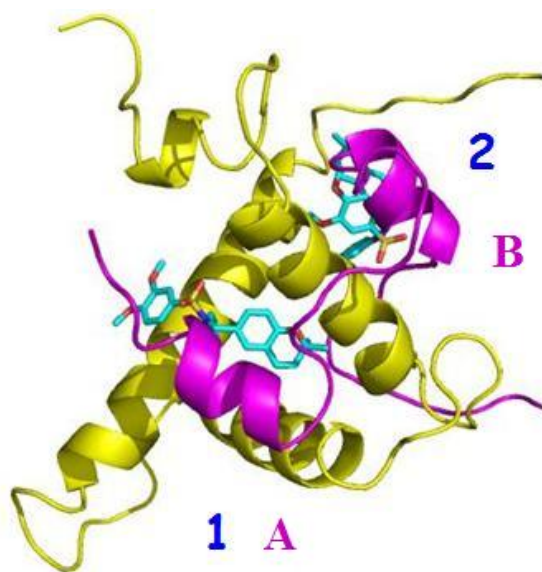


Figure 42. Two helices on HIF-1 α (purple) superimpose with docked KCN1 at Site 1 and Site 2

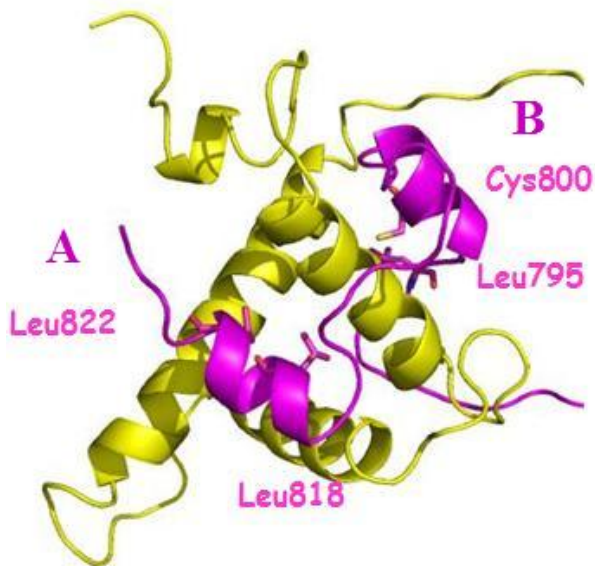


Figure 43. Crucial residues on HIF-1 α , namely Leu795, Cys800, Leu818 and Leu822

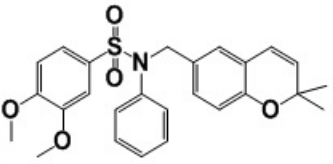
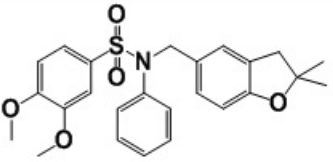
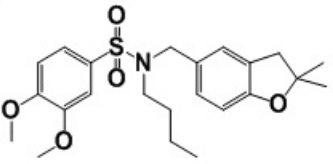
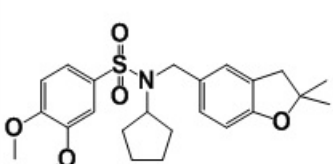
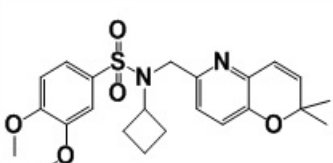
5.3.3. Validate the scoring functions

At this point in the project, there were ~30 KCN1 analogs with the corresponding IC₅₀ values available (structures and IC₅₀ values are included in the **Appendix I**).⁵⁸ It is worth noting the structure-activity relationship of these compounds appears “flat”. Nevertheless they were used in the event they might provide some preliminary guidance. An initial QSAR construction was attempted and each analog were subjected to Glide docking at each of the two proposed binding sites (**Figure 41**, Site 1 and Site 2) followed by energy rescoring with Prime MM-GBSA.^{65,66}

Most of the experimental IC₅₀ values were from a single run and no average value was available. To minimize possible biological variability over time and different cell batches, the corresponding KCN1 IC₅₀ value determined in the same run was used as a reference in order to uniformly scale the IC₅₀ value for each analog. Instead of directly applying the IC₅₀ value, the ratio of the IC₅₀ values of the analog and the same plate KCN1 sample was used as a measurement of activity ($\Delta\text{IC}_{50} \sim K$) and applied to the linear regression analysis. However, no acceptable linear regression result could be obtained for the 30 compounds. Alternative approaches were considered to construct a QSAR-like scheme for prioritizing the structures generated by FRESH.

Thus, receiver operator curves (ROCs) for the binding sites were obtained to assess whether the estimated binding energy values at Site 1 and 2 two sites might be useful for predicting enrichment of the active compounds. The compounds listed in **Table 4** which possess averaged IC₅₀ values no greater than KCN1 were defined as actives. **Figures 44** and **45** illustrate the ROCs at Sites 1 and 2.

Table 4. KCN1 and active analogs with multiple experimental measurement repeats

Structure	Name	No. of repeats	IC50 (nM)
	KCN1	42	648
	1601	6	306
	1604	6	478
	1606	6	378
	2609	30	280

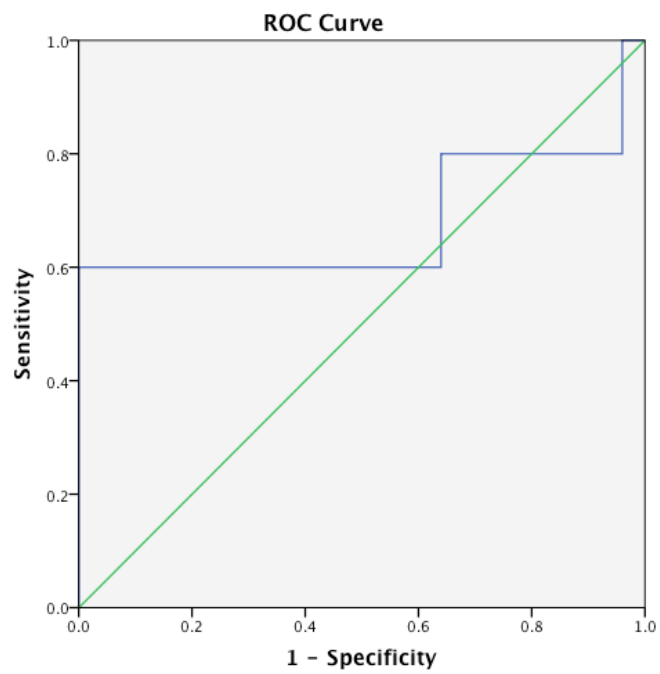


Figure 44. ROC at Site 1. AUC = 0.68

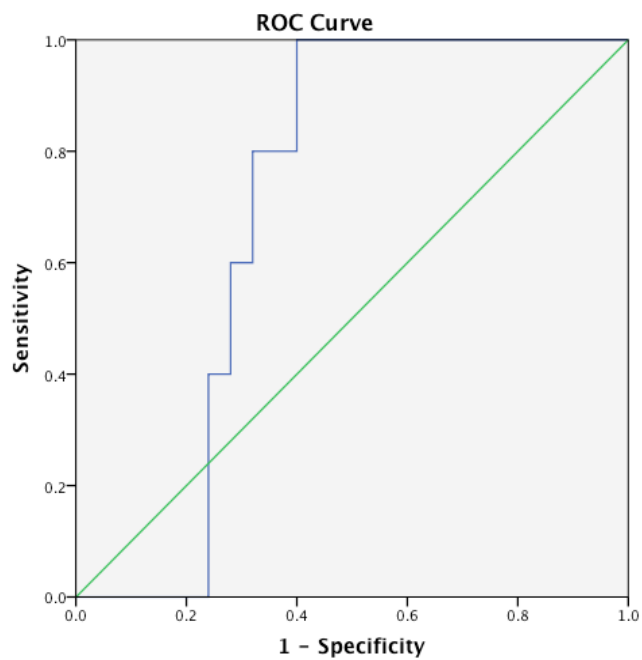


Figure 45. ROC at Site 2. AUC = 0.70

The AUCs for both sites are around 0.7, indicating that both the estimated energy values at Sites 1 and 2 can differentiate active compounds from inactive analogs to some extent. Therefore, the intended plan for the FRESH program was to incorporate the MM-GBSA scores at both sites for prioritizing. A test run for this plan was performed on the existing 30 compounds, with the MM-GBSA scores for KCN1 as the cutoffs (Site 1: -26.5 kcal/mol, Site2: -27.5 kcal/mol). Structures with better MM-GBSA scores at both sites were retained. Only 3 compounds remained: KCN1, **1601** and **2609**, all of which are true positives and included in **Table 4**. It is noteworthy that **1601** and **2609** were the two most potent analogs at this point in the project. This result further supports the plan to incorporate the MM-GBSA at both sites into FRESH.

5.4. Application of the FRESH program

5.4.1. The FRESH program design

The FRESH program was designed according to the synthetic route in **Figure 46**. The R₁ group originates from a sulfonyl chloride building block, the R₂ group from an amine and the R₃ group from an aldehyde. For this particular FRESH program, the R₁ group was kept constant (the same as KCN1). The fragments for the R₃ group were only chosen from the compounds in **Table 4**. The R₂ groups were obtained from the corresponding building blocks in the commercial library.

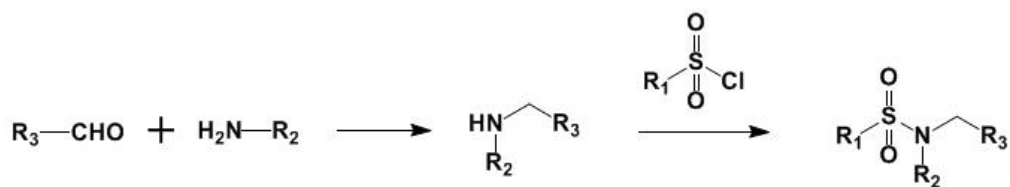


Figure 46. Synthetic route for KCN1 and its analogs

Within the context described above, the FRESH program for KCN1 analogs based on **Figure 46** was thus designed as shown in **Figure 47** (illustration interface only, detailed sub-protocol components are hidden). Starting from the commercial library, the corresponding building blocks for R_2 (amine) were screened and collected. For this particular step of the entire program, all the amides also appear in the list. However, according to the synthesis route, amides cannot serve directly as the desired building blocks, so the fragments arising from amides should not be included. Additional components were thus incorporated in the pipeline to treat and eliminate amides. Like all other FRESH programs, fragments were prioritized by filtering with potential liability, stability or synthesis concerns. Other properties such as the Rules of Five were also incorporated. The polar surface area cutoff was set at 90 \AA^2 and the log of BBB was set at -0.5 to ensure the desired CNS drug-likeness. For the potency criteria, as discussed in the last section, the threshold for the MM-GBSA values of the two docking sites was -26.5 kcal/mol (Site 1) and -27.5 kcal/mol (Site 2).

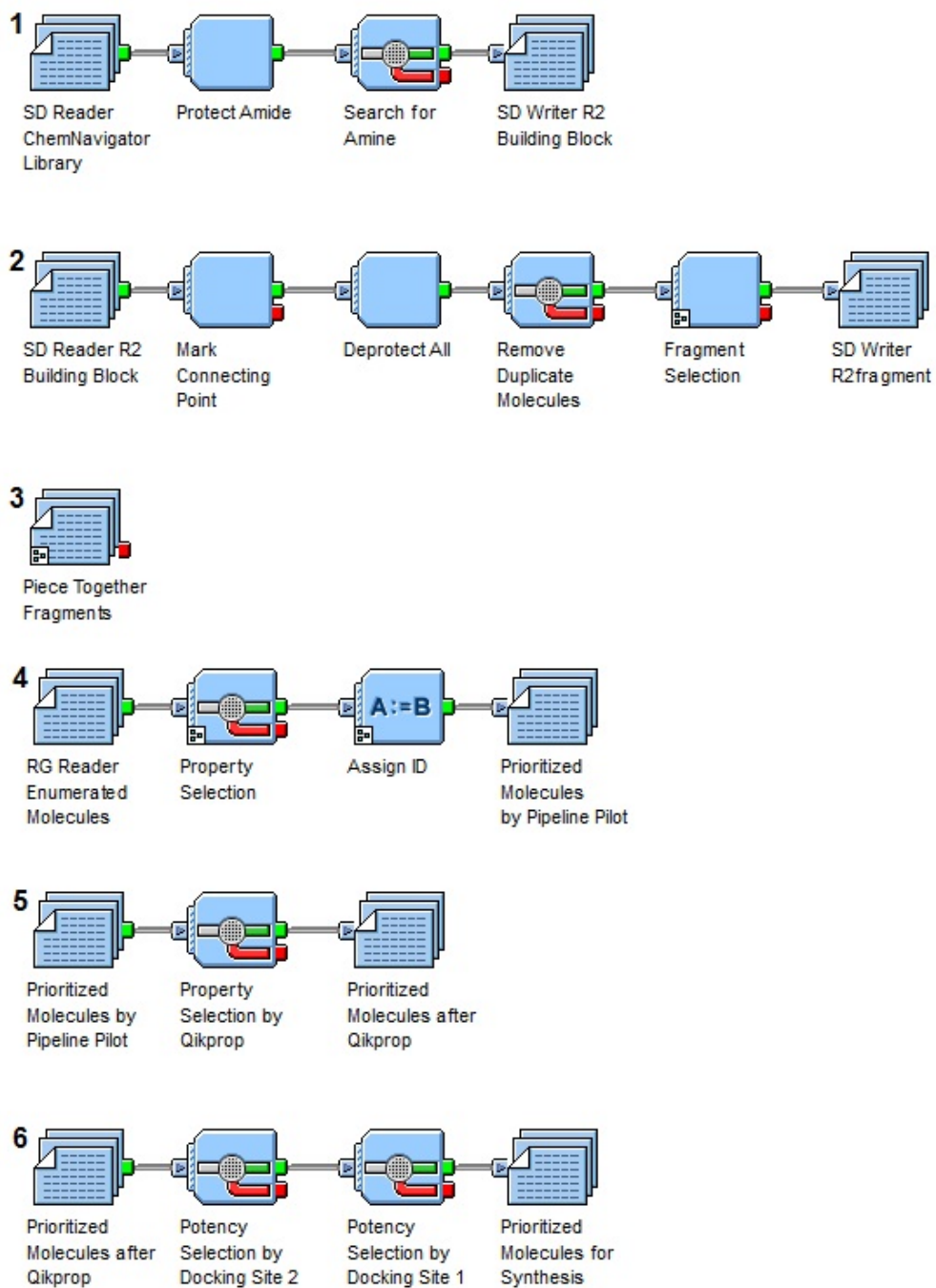


Figure 47. The FRESH program (illustration interface only) for prioritizing KCN1 analogs. Some sub-protocol components are not shown.

5.4.2. Resulting structures

Table 5 illustrates the final list of novel structures output by FRESH. All six compounds are chiral except **270**. Upon further investigation on the Internet for the commercial vendors, the building blocks for **95** and **98** are available both as racemates and optically-pure enantiomers. The decision was made to pursue the synthesis and testing of racemic mixtures considering the cost of the building blocks.

Table 5. Result of the FRESH approach for KCN1 analogs

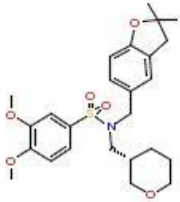
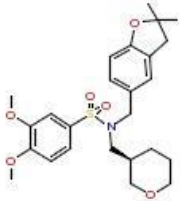
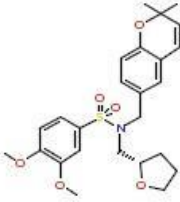
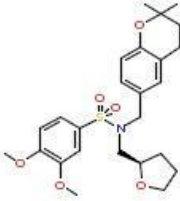
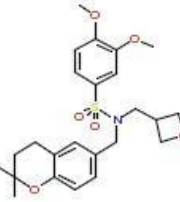
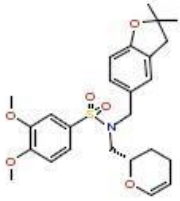
Molecule	Name	MMGBSA_Site1	MMGBSA_Site2
	93	-28.7	-27.9
	94	-28.8	-30.2
	95	-29.9	-33.3
	98	-28.2	-30.3
	270	-27.8	-31.4
	280	-30.3	-33.0

Figure 48 demonstrates the bio-assay results for the racemic mixtures, the single predicted structure so far tested. Compared to the smooth diving curve of **2609**, **95** demonstrates an abnormal wave-shaped curve. The compound is likely to be a false positive.

It is worth noting that for all previous FRESH cases, the false positives were always present. Nevertheless, other active compounds were identified successfully. Since the only result available at the time of the project was a racemic mixture, it was too early to simply reject this particular QSAR. Additional compounds still need to be synthesized and tested before commenting on these particular QSAR selection criteria. However, as the project proceeded, the team decided to abandon this particular scaffold and receptor model as new evidence and discoveries emerged. For example, the GSU collaborators found that **1609** does not bind to p300 as expected by SPR experiments, suggesting alternative receptor targets may be operating. A new scaffold has now been identified which is discussed in the next section.

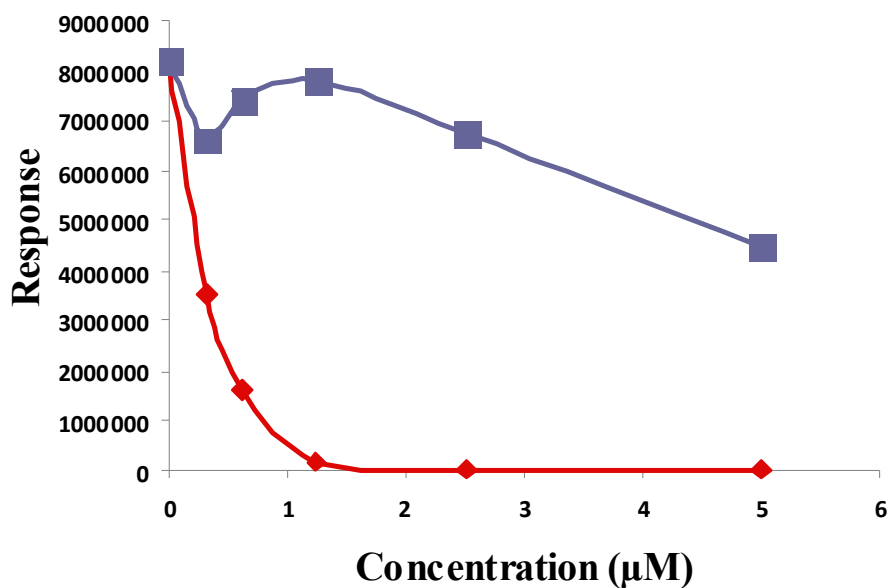


Figure 48. The initial bio-test results for **95** (blue). **2609** (red) was used as a reference.

5.5. Application of FRESH to the new scaffold

5.5.1. New compound scaffold

As the project progressed, the GSU collaborators discovered a new diaryl alcohol scaffold shown in **Figure 49** (left). The lead compound **BW-HIF-84** (to the right of the figure) has an IC_{50} of 300 nM, which is comparable to **1609**. Compared to the original lead KCN1, this scaffold is more “lead-like”. For example, the MW of **BW-HIF-84** is 344, which is more than 100 amu lower than KCN1. The predicted logP value of the new lead is around 3.6, which is 1 log unit lower than KCN1. The relatively low MW and logP values offer a larger chemical space to explore. The FRESH program discussed below will focus on this scaffold.

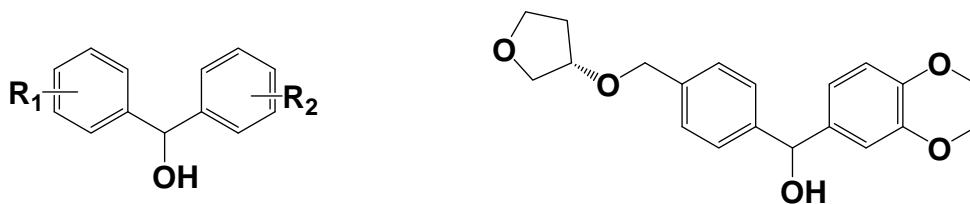


Figure 49. New scaffold for p300/HIF-1 α inhibitor and **BW-HIF-84**.

5.5.2. Construction of QSAR

As stated in the last section, the team found contradictory evidence for the p300-based receptor model, suggesting one or more unknown protein targets. Therefore, a ligand-based model independent of any particular receptor was pursued. This round of QSAR took ~50 structures with their corresponding IC₅₀ values as the biological data and divided them roughly by a 2:1 ratio for training set and test set compounds. Linear regressions using various descriptors were attempted, and only the ECFP molecular descriptor discussed in Chapter 3 delivered acceptable correlations. **Figure 50** demonstrates the linear regression result for the training set with an R² value of 0.83. The predicted IC₅₀ values for the test set compound were plotted against the corresponding experimental values and a satisfactory predictive Q² value of 0.73 was obtained (**Figure 51**). The QSAR line in **Figure 50** will be employed in the FRESH program.

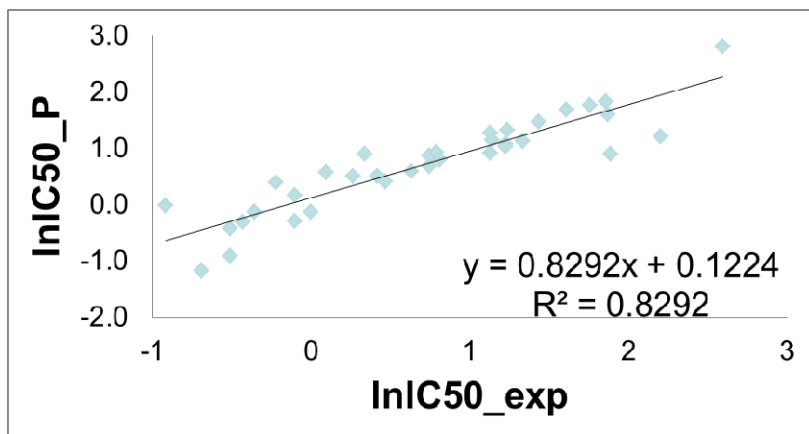


Figure 50. Linear regression results for the training sets.

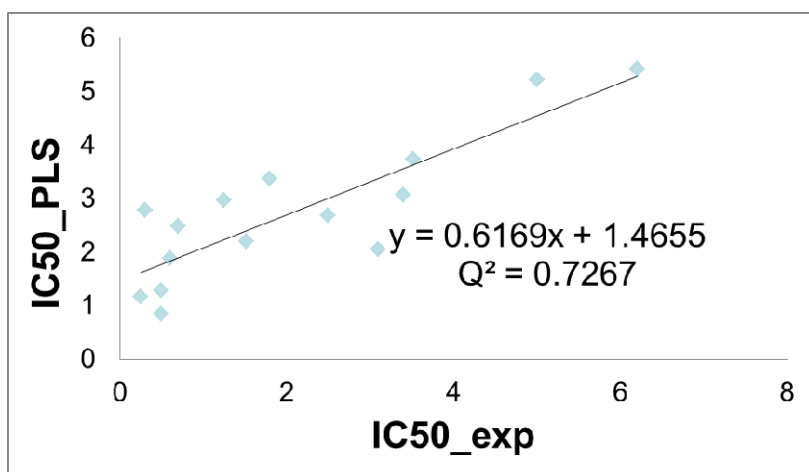


Figure 51. Q^2 results for the test set compounds.

5.5.3. The FRESH program and the resulting structures

The FRESH program was designed according to the synthetic route in **Figure 52**. The R_1 group originates from a bromide building block and the R_2 group from an aldehyde. Both R groups were screened against the commercial library.

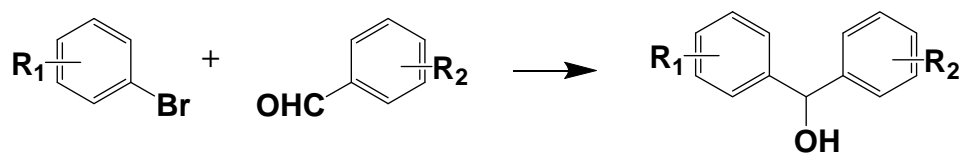
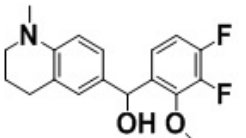
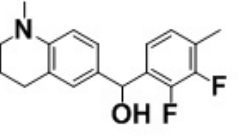
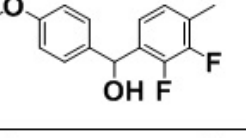
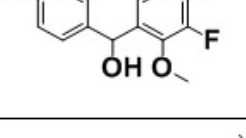
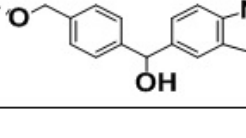


Figure 52. Synthetic route for the diarylalcohol analogs

Within the context described above, FRESH applied to the diaryl alcohol analogs was designed in a manner similar to that for the old scaffold. The estimation for potency applied the QSAR line in the last section. It is worth noting that at the time of the project, no compounds has a predicted logBB value greater than 0 which is desirable for blood brain barrier penetration. Compounds with KCN1-similar IC_{50} 's but significantly improved logBB values would definitely be more attractive. As stated in Section 5.5.1, this particular scaffold offers larger chemical space to explore. For this round of FRESH, the logBB cutoff was determined to be 0 and the IC_{50} cutoff was set at 2 μ M following discussion with GSU collaborators. After obtaining the final list, another meeting was held with GSU collaborators and together the Emory and GSU collaborators selected the compounds in **Table 6** for synthesis and testing. Currently, the GSU chemists are working on the syntheses.

Table 6. Result of the FRESH program for KCN1 analogs

Structure	Name	logBB	Predicted IC ₅₀ (nM)
	2131810	0.1	330
	1812170	0.1	340
	2131810	0.1	350
	1094161	0.1	350
	2751578	0.1	1400

5.5.4. Future directions

Two possible future directions are proposed depending on the testing results of the compounds in **Table 6**. If at least one compound demonstrates improved IC₅₀ as expected, further iterations will be performed based on this result in a manner similar to the case studies described in Chapter 3. Additional structures with high similarity will be prioritized for synthesis.

It is also possible that all the compounds in **Table 6** are inactive and, thus, they are all false positives. Were this to be the outcome, it may still be possible that additional compounds from the FRESH program will prove to be active and able to penetrate the BBB. Nonetheless, it would be more practical to consider revising the QSAR for potency

selection. One reason to suspect the failure of the current QSAR line in Section 5.5.2 is that it most likely lacks sufficient bio-information to discriminate between actives and inactives (compounds with IC_{50} values higher than the experimental cutoff, currently at 5 μM). Inactives were not included in the training set due to the lack of precise IC_{50} values. At this time of the project, over 150 compounds are available with ~40 compounds having IC_{50} values lower than 5 μM . A Bayesian model with ECFP as the descriptor, which is able to incorporate inactive analogs with imprecise bio-values in the training set, should be considered for the new QSAR.

Part I: Conclusions and Future Directions

Chapter 4 and Chapter 5 describe how the FRESH approach is able to provide a list of structures for a specific lead-optimization task which satisfy multiple requirements including predicted potency, physical properties and ADMET aspects. For the SNRI dual inhibitor project in Chapter 4, the program has successfully provided at least two valuable compounds out of a total of five using a relatively rough receptor-based model. Additional work on the QSAR part is needed to better eliminate false positives. For p300/HIF-1 α project in Chapter 5, the program provided interesting candidate for synthesis and additional testing work has to be performed before drawing a conclusion.

Part II: Monocarbonyl Curcumin Analogues: Heterocyclic Pleiotropic Kinase Inhibitors That Mediate Anticancer Properties

In this part of the dissertation, a series of monocarbonyl curcumin analogues are investigated for their anti-inflammatory and anticancer properties. Mechanism has been examined by exploring kinase inhibition trends. In particular, among the 50 screened kinases relevant to many forms of cancer, the binding of curcumin analogues to AKT-2 were analyzed in detail by molecular modeling at the kinase ATP pocket. In addition, the most extensively studied compound (**EF31, 4**) was further investigated for potency variations against a series of kinases by protein sequence comparisons.

Chapter 6: Curcumin analogs as Pleiotropic Kinase Blockers

6.1. Project background²

Curcumin (**Figure 53**) is the major component of the root powder of *Curcuma longa*. It is consumed world-wide as a spicy flavor ingredient (curry powder). It has also been used as a food coloring agent due to its distinctive turmeric yellow color. In addition, curcumin plays a crucial role in traditional medicine for its therapeutic effects against rheumatoid arthritis, indigestion, liver disease like jaundice and insect bites.⁶⁷ Curcumin has attracted the attention of medicinal chemists recently due to its anti-tumor activity, relatively low toxicity and its pleiotropic properties, which fits the “magic shot-gun” drug design strategy. Unfortunately, low potency, poor bioavailability and fast metabolism have limited its clinical application.

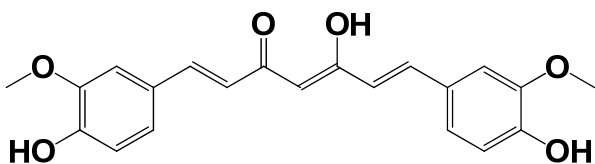


Figure 53. The structure of curcumin.

In attempts to improve the solubility, bioavailability and stability of curcumin, our research group previously prepared a series of curcumin analogs by modifying both the central and terminal moieties of the molecule. The central keto–enol functionality of was replaced by a monocarbonyl group embedded in a heterocyclic six-membered ring conjugated with a pair of flanking C=C bonds. The terminal oxygenated aromatic rings were exchanged for fluorophenyl and pyridine moieties. A representative set of analogues

² This background section is from both my master thesis and my publication

are portrayed by structures 3–7 in **Figure 54**. These analogs have demonstrated enhanced anti-tumor activity relative to curcumin while being tolerated to the cell.^{68 67}

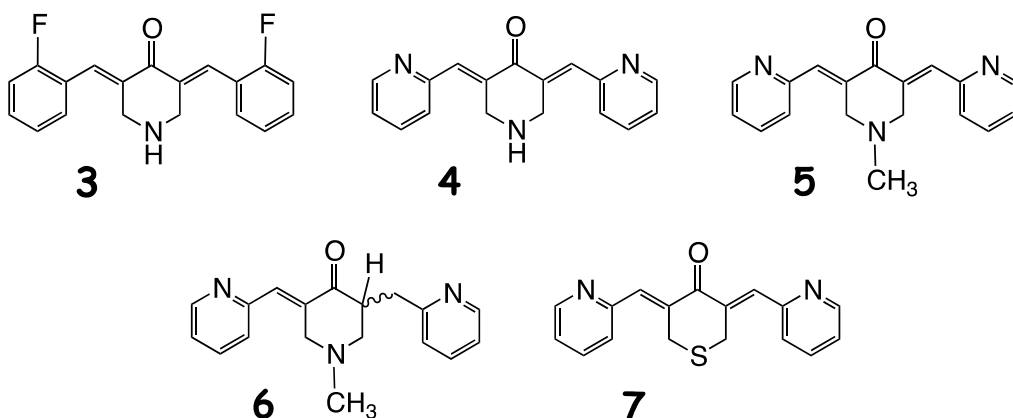


Figure 54. Structures of curcumin analogs.

Our collaborators at WCI (Professor Shim and Dr. Andrew Brown) have explored the mechanism of action for the analogs depicted in **Figure 54**.⁶⁹ The compounds were evaluated against a series of kinases that appear crucial for various tumor cell signaling pathways. According to the results listed in **Table 7**, AKT-1 and AKT-2 appear to be the most inhibited kinases as reflected by **4** with an IC_{50} value of 20 nM against both AKT-1 and AKT-2. Molecular modeling studies described in the next section were performed to understand the variation of the ligand IC_{50} values.

Table 7. IC₅₀ values (unit: μM) of a series of curcumin analogs against various kinases.

Kinase	3	4	5	6	7
AKT1 (PKB α)	0.78	0.02	2.6	>100	3.6
AKT-2	0.72	0.02	1.9	>100	3.3
IKBKB (IKK β)	72	1.1	4	15	34
RPS6K1	20	1.4	6	>100	12
AMPK (A1/B1/G1)	46	1.5	2.5	10	26
RAF1 *cascade	24	1.6	6.5	>100	8.5
MEK1 *cascade	12.8	1.1	9	>100	30
ERK2	13	4	27	>100	20
NEK1	77	0.5	3.5	83	14
KDR (VEGFR2)	0.77	0.66	1.3	>100	6.5
MAPK14 (p38 α)	92	35	>100	>100	>100
MAPKAPK2	>100	6.7	9.8	>100	16

6.2. Modeling of curcumin analogs³

6.2.1. Comparison of AKT-1 and AKT-2

The kinase domains of AKT-1 and AKT-2 demonstrate 82% identity and 93% similarity. In addition, comparison of the residues surrounding the ATP pockets for AKT-1 and AKT-2 reveals that the residues are identical (**Figure 55** squared in black). It is therefore not surprising that the IC₅₀ values of AKT-1 and AKT-2 are almost identical. Additional analysis for **4** by Dr. Andrew Brown (WCI) has revealed that competition with ATP dominates drug action against AKT-2.⁶⁹ Accordingly, the crystal structure of the corresponding kinase domain of AKT-2 (PDB code 3E88⁷⁰) was selected as the receptor structure for ligand docking. Once adjusted by the Maestro Protein Preparation Wizard, the ATP binding pocket of AKT-2 was subjected to Glide docking by **3–7**.

³ This section is directly from my publication⁶⁹

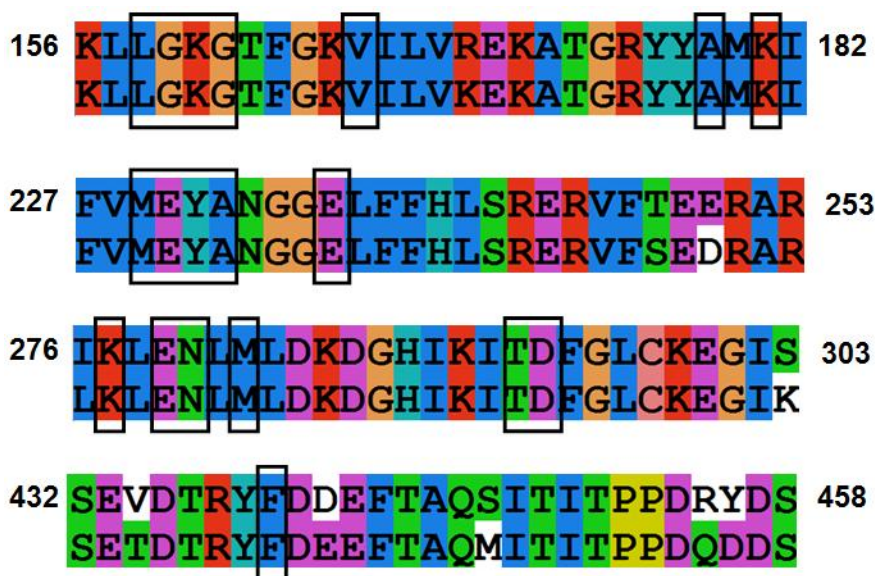


Figure 55. Sequences of aligned AKT-1 (lower row) and AKT-2 (upper row with residue numbers). The residues around the ATP pocket are squared in black.

6.2.2. Docking pose analysis

Figure 56 illustrates the best-scored docking pose for protonated **4** in the ATP binding pocket of AKT-2 in which several non-covalent interactions anchor the ligand to the protein. Hydrogen bonds are established between one of the pyridine nitrogen atoms and side chain Thr292 on the receptor and between the ligand carbonyl group and Lys181. A salt bridge between Glu236 and the protonated nitrogen within the central ring of the ligand is evident. In this project, it is assumed that the curcumin mimics with pKa values around 6.5-7.0 are protonated at physiological pH, a phenomenon consistent with creation of a ligand-Glu236 salt bridge. In addition, the docking models for **4** sustain a pyridine ring in a relatively hydrophobic pocket circumscribed by Ala179, Met229, Glu230, Tyr231, Ala232, and Met282.

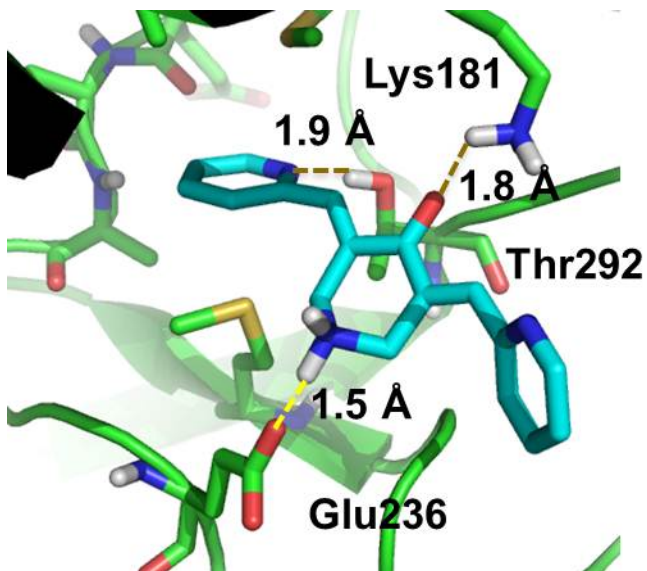


Figure 56. Top predicted pose from Glide docking of N-protonated **4** to AKT-2.

For fluorinated **3**, the top Glide pose is similar to that of **4** (**Figure 57**). H-bonds with Glu236 and Lys181 are maintained, and one of the phenyl rings resides in the same hydrophobic pocket. However, consistent with the higher IC_{50} values of **3**, the favorable hydrogen bond with Thr292 is lost. To some extent the latter is compensated by electrostatic association between the aromatic fluorines and the proton of the axial NH^+ bond in the central ring, although the associations are weak ($r(\text{aro}) F(\delta) \cdots H\delta(N)$ of 3.1 Å).

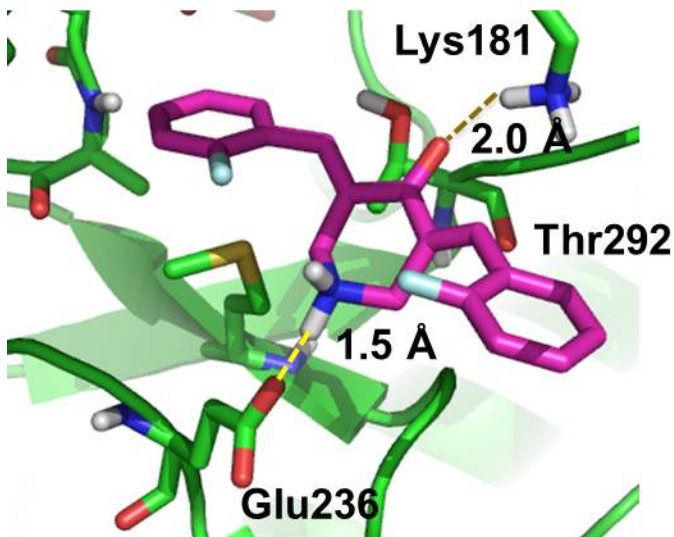


Figure 57. Top predicted pose from Glide docking of N-protonated **3** to AKT-2.

Figure 58 shows that the *N*-methyl analogue of protonated **5**, like **4**, forms a hydrogen bond between one of the pyridine nitrogen atoms and Thr292. The *N*-methyl group on the central ring assumes an equatorial conformation, causing the NH^+ to form a salt bridge with Glu236 from the axial position. The latter obligates the ligand to move toward Glu236 and away from Lys181 by comparison with **3** and **4**, essentially deleting the hydrogen bond interaction with this residue. In addition, two hydrogen atoms in **5** are separated by only 2.0 Å (black stippled line), somewhat below the sum of van der Waals radii (2.4 Å) and thereby introducing internal ligand strain energy. The same $\text{H}\cdots\text{H}$ distance for **4** (2.3 Å) is at the minimum acceptable van der Waals contact. The diminished hydrogen bonding and ligand strain energy can explain the relatively higher IC_{50} values of **5** relative to **4**.

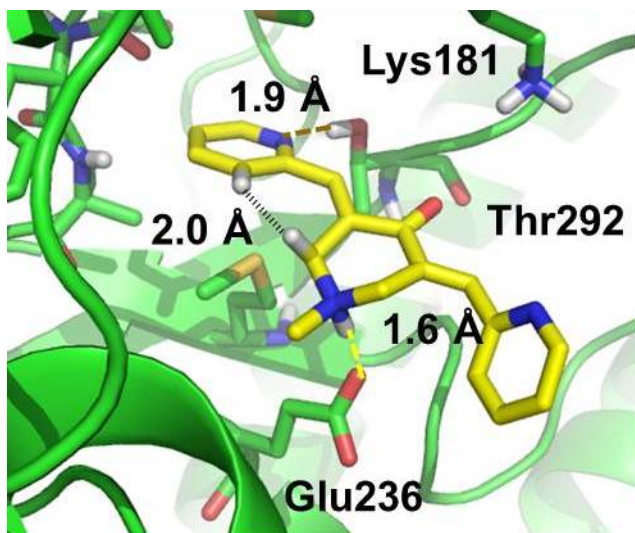


Figure 58. Top predicted pose from Glide docking of N-protonated **5** to AKT-2.

Partially saturated **6** exists as two enantiomers, **6R** and **6S**. As shown in **Figures 59a** and **59b**, respectively, the left half of each ligand and the equatorial *N*-methyl orientations are identical to those of **5**. While the stereoisomer poses retain the non-covalent interactions with Thr292 and Glu236, the hydrogen bond with Lys181 is lost as observed for **5**. Critical new contacts arise, however, because the C=C saturation in **6** requires the CH₂-pyridine moiety to be relocated, placing the relatively hydrophobic edge of the pyridine ring into a polar sector of the protein's glycine-rich loop. Furthermore, in **6R** the pendent CH₂-pyridine group fits into the pocket only by adopting a near eclipsed conformation with the adjacent C–H bond of the central six-membered ring. Thus, the loss of a key H-bond, the predicted placement of the benzylic pyridine rings into unfavorable regions and in one case in an unfavorable conformation appear to contribute to the significantly reduced activity of the enantiomers of **6** on AKT-2.

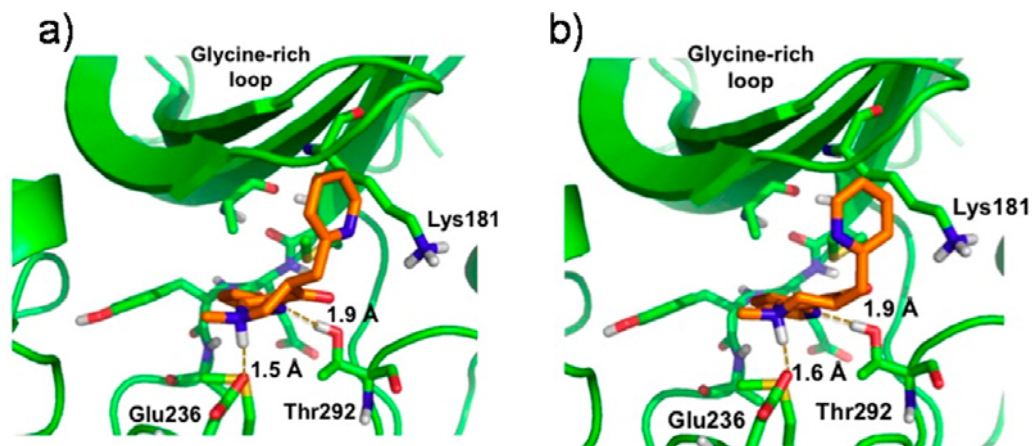


Figure 59. Top predicted poses from Glide docking of of N-protonated **6** to AKT-2: (a) 6R; (b) 6S.

Analog **7** possesses a sulfur atom in the central ring instead of nitrogen. In the most favorable pose depicted by **Figure 60**, the hydrogen bond shared between ligand and Thr292 remains, while association between C=O and Lys181 NH increases to 2.7 Å, weakening the H-bond significantly. The NH to S replacement naturally not only eliminates the salt-bridge with Glu236 but also causes the ligand to retreat as a consequence of the unfavorable electrostatic contact between the bulky electron-pair bearing sulfur atom and Glu236. The lack of two anchoring H-bonds and the electrostatic disconnect are believed to contribute to the increased IC₅₀ values of **7** compared to **4**.

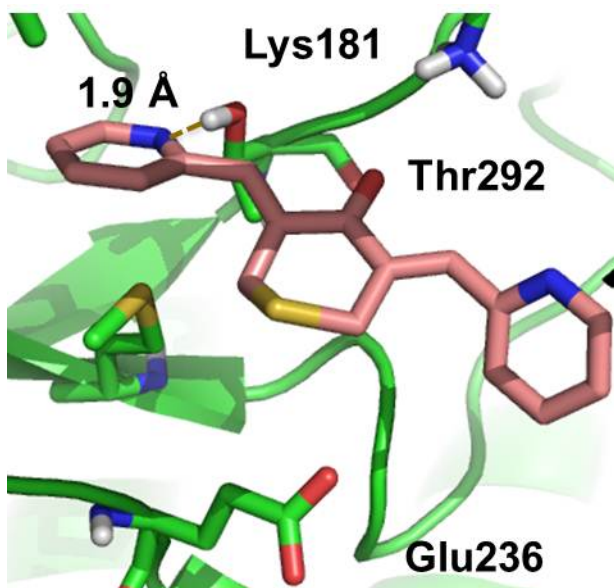


Figure 60. Top predicted pose from Glide docking of 7 to AKT-2

It is worth noting that all the curcumin analogues investigated herein carry an α,β -unsaturated ketone Michael acceptor. Thus, reversible covalent bonds might be formed between kinase cysteines and these compounds. For example, **Figure 61** shows that a cysteine (Cys311 AKT2) is located near the substrate binding site. Consistent with the design of the bioassay, the formation of a covalent bond between ligand and cysteine would most likely interrupt substrate binding, making the target peptide reagent susceptible to cleavage as monitored by subsequent separation of the FRET pairs. Such an event can hinder quantification of the non-covalent binding affinity of the ATP competitive inhibitor unless this mechanism is a small contribution to the overall blockade. However, as mentioned in the last section, the binding analysis by Dr. Brown suggests that the competition with ATP dominates action against AKT-2.⁶⁹ This is consistent with the fact that the docking pose at the ATP-binding site provides a semi-quantitative explanation of the various IC_{50} values for AKT-2 inhibition.

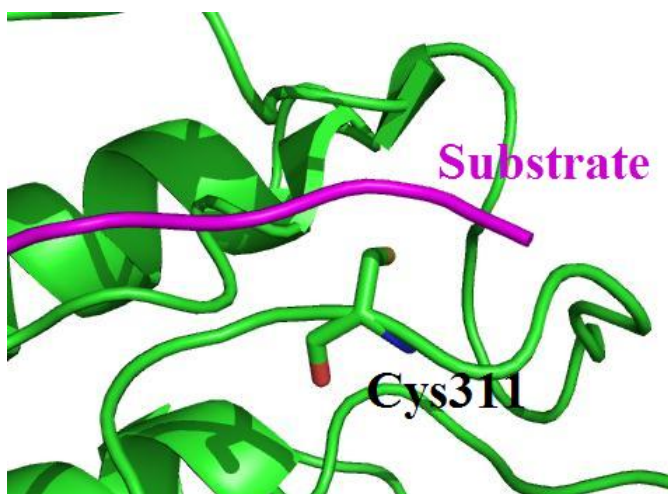


Figure 61. Cys311 in the AKT-2 cleft where the substrate binds

6.2.3. Sequence Comparison of Kinase Binding Sites

Further analysis to understand the pleiotropic aspects of **4** by sequence alignment and identity/similarity comparison of the residues around the ATP binding sites for various kinases was performed. The 3D structures of the proteins were aligned using the protein structure alignment tool in the Schrodinger Maestro software package. By employment of the docking pose of **4** in AKT-2, key residues within the ATP binding sites were selected for comparison with the corresponding residues in other kinases: Ala179, Met229, Glu230, Tyr231, Ala232, Met282, Glu236, Lys181, and Thr292.

Figure 62 demonstrates the sequence comparison result. It reveals that most of the active kinases share >55% identity and >75% similarity with these residues in the AKT-2 binding environment. KDR is an exception to this rule, i.e., 33% and 67%, respectively. Further examination of the KDR binding center reveals two cysteines, Cys919 and Cys1045, which can form covalent Michael addition to the α,β -unsaturated ketone moiety. It is conceivable that covalent binding dominates the binding affinities in this case and thereby rationalizes the similar KDR IC₅₀ values for the four curcumin analogues **3**, **4**, **5**,

and 7 (Table 7). However, examination of other kinases with low activity in Figure 62 also reveals similar cysteines at the ATP binding sites (example, KIT), suggesting that the corresponding cysteines alone do not contribute the majority of the observed activity of KDR. Further work is still required to understand why exceptions like KDR exist.

Kinase	1	2	3	4	5	6	%Identity	%Similarity
AKT2	A	MEYA	M	E	K	T	100	100
AKT1	A	MEYA	M	E	K	T	100	100
IKKB	A	MEYC	V	D	K	I	56	78
NEK1	V	MDYC	F	D	K	G	33	56
RPS6KB1	A	LEYL	M	E	K	T	78	89
AMPK	A	MEYV	L	E	K	A	67	78
PRKCB1	A	MEYV	M	D	K	A	67	78
CHEK1	A	LEYC	L	E	K	S	56	89
KDR	A	VEFC	L	N	K	C	33	67
MAPKAPK2	A	MECL	L	E	K	T	67	78
PIM1	A	LERP	L	D	K	I	33	67
IRAK4	A	YVYM	L	S	K	S	33	56
CDK7	A	FDFM	L	D	K	A	22	67
KIT	A	TEYC	L	D	K	C	44	66
ACVR1B	A	SDYH	L	S	K	A	33	56
CDK2	A	FEFL	L	D	K	A	33	67
EGFR	A	TQLM	L	C	K	T	33	56
PLK1	A	LELC	F	S	K	G	33	44
ERBB2	A	TQLM	L	C	K	T	33	56

Figure 62. Key residues around the ATP binding sites of various kinases. Residue type and number according to AKT2: 1: Ala179 2: Met229, Glu230, Tyr231, Ala232 3: Met282 4: Glu236 5:Lys181 6:Thr292. Upper panel, >85% inhibition; lower panel, ≤10% inhibition

6.3. Conclusion

Molecular modeling at the ATP binding site of AKT-2 provides a qualitative explanation for the observed ligand activity consistent with the fact that competitive inhibition dominates the action against AKT-2. At the same time, reversible Michael addition can possibly occur between the curcumin analogs and the cysteine residues on the kinases.

The sequence comparisons revealed that most of kinases which are active to **4** have a relatively high similarity with AKT-2, although exceptions exist.

Part III: Development of 2nd Generation NAMFIS

Software Program by Java

The NAMFIS (NMR Analysis of Molecular Flexibility in Solution) program was originally developed by Cicero et al. using Python 2.4.⁷¹ It takes a conformational pool of an organic ligand molecule and its NMR spectrum as the input and provides useful conformation information for the ligand. In this part, the author redesigned the entire NAMFIS program in Java which is less susceptible to the cross-platform issue. The 2nd generation of NAMFIS program also provides a user-friendly GUI input and incorporates additional validation procedures to handle exceptions and capture the user input errors. The new NAMFIS also provides solutions to the “mathematical infeasible” situation which is untreated in the old Python version.

Chapter 7: Design and Improvement of the NAMFIS Program

7.1. NAMFIS background

7.1.1. The Problem of Force Field Methods

The conformation of a ligand, either bound or free in solution, is useful information for structure-based ligand-protein inhibitor design. Assuming the protein-bound bio-active conformation of a ligand molecule is known, when modifying the lead structures medicinal chemists can constrain the ligand molecule into that bio-active conformation while still in solution. In theory, this can reduce the entropic penalty upon binding and hopefully improve binding affinity. Obtaining an x-ray crystal structure of the ligand-protein complex provides a way of acquiring the conformational information, but unfortunately protein crystallization remains a challenge for many proteins of therapeutic interest.

Systematic *in-silico* conformational search of a ligand molecule with a single or combined force fields provides an approach to obtain considerable conformational information. However, the output of such a search can include several hundreds to thousands of conformers. One possible solution to limiting the task is to rank the conformers by the force-field-based energy calculations and select the top ones for further investigation. This solution appears completely feasible and logically implemented. However, different force fields can assign different partial charges to polar groups. As a result, force field equations remain somewhat semi-empirical. The calculated energies are actually force field dependent and the variation between force fields differences can lead to different relative energy values and rankings for the same

set of conformations. This observation was made by Lakdawala, Snyder et al. in 2001.⁷² During the investigation of seven taxol conformers, neither molecular mechanics nor semi-empirical quantum chemical methods presented a consistent energy ranking.

7.1.2. Mathematical Background for NAMFIS

Small molecule NMR spectroscopy provides another approach to extract conformational information in solution. Two elements of geometry are the key. On the one hand, is the proton-proton coupling constant, $^3J_{\text{H-H}}$, which is related to the corresponding atom types and dihedral angles by various Karplus equations. A second valuable source of information from NMR analysis is the distance (within 5 Å) between protons, which is acquired by the NOESY (Nuclear Overhauser Effect Spectroscopy). Taken together, the dihedral angles and the inter-proton distances provide a powerful set of parameters for determining ligand conformation.

The NAMFIS program was developed to obtain conformational information from the experimental $^3J_{\text{H-H}}$ and NOESY distance values. NMR signals for most organic molecules at room temperature and below is an averaged signal from all the rapidly interconverting conformations. Instead of assigning a single conformer, the NAMFIS program provides a set of conformers to fit the averaged NMR signal. NAMFIS requires two input files. One contains the experimental NMR information, while the other is a conformational ensemble of ligand conformers generated by a conformational search. For each input conformer, the NAMFIS program calculates the corresponding proton-proton distances using the included 3D atom coordinates, and $^3J_{\text{H-H}}$ values from the conformer input geometries by the corresponding Karplus equations.⁷³ Consequently, a constrained minimization is performed to vary the mole fractions of each conformer, calculate the

averaged coupling constants and distances, and compare these to the corresponding experimental input. The constraints are listed in **Figure 63**. A_i^{exp} stands for the experimental input and ΔA_i represents the experimental errors. The first constraint requires the averaged calculated value A_i^{calc} to remain in a window defined by A_i^{exp} and ΔA_i . The second and third constraints specify that the mole fraction must be non-negative and the sum of the mole fractions should equal to 1.0. The minimization target function (a Sum of Square Differences, SSD) is shown in **Figure 64**, where the sum of relative errors is added for each selected parameter. The J terms (J coupling constants) are assigned a weight value $w(j)$ with a default value of 1.0.

$$A_i^{exp} - \Delta A_i \leq A_i^{calc} \leq A_i^{exp} + \Delta A_i$$

$$x_j \geq 0$$

$$\sum_j x_j = 1.0$$

Figure 63. Constraints imposed in the minimization.

$$SSD = \frac{1}{2} \left(\sum_{i=1}^m [(d_{calc}(i) - d_{expt}(i))/d_{err}(i)]^2 \right) + \frac{1}{2} \left(\sum_{j=1}^n [w(j) \times (J_{calc}(j) - J_{expt}(j))/J_{err}(j)]^2 \right)$$

Figure 64. Equation for calculating SSD.

7.1.3. Problem with the Current Version of NAMFIS

The current version of the NAMFIS program was written in Python 2.4 which is already 10 years old. Unfortunately, it is not compatible with the latest release of Python version, even in syntax. The NAMFIS program requires several external libraries to run,

which are also upgraded and not back-compatible. In addition, Python is a scripting language. It is not capable of compiling all the dependency libraries together and then executing. Each time when the NAMFIS program is executed, all the dependent libraries must be present at the designated locations. The user therefore must resolve the dependencies of all the old versions of libraries before executing the program. This renders the program extremely susceptible to the cross-platform problem. The inconvenient result is that program-sharing can require considerable work to establish a running version.

The current version of the NAMFIS program also lacks necessary procedures to capture user input errors. In most of the situations sustaining an error, the program runs normally, but the calculated results are not the user-intended ones. The input file for the experimental NMR data is a text file in a pre-defined format. The user inputs the empirical values and generates a new input file. The latter is susceptible to user input errors like wrong atom mappings, duplicate input, wrong atom numbers and illegal formats. For example, the J coupling constant requires atom input in the order H_1, C_1, C_2 and H_2 . If the user accidentally writes H_1, C_2, C_1 and H_2 instead (already happened to one on-going project and previously undetected), the program still runs normally, but the calculated dihedral angles are obviously not the user-intended ones, and the final result is consequently affected. Input errors like this can be difficult to identify and should be captured at an early stage.

While the above-mentioned issues can possibly be avoided by the user, the following problem requires further attention by the developers. If the input NAMFIS case is mathematically infeasible, or in other words there is no feasible space to the specific

NAMFIS case under all the constraints, the program should detect this and warn the user about the situation. For example, assume the input conformational pool consists of 5 conformers with calculated J values of 4.0, 4.1, 4.2, 4.3 and 4.4. If the upper bond derived from the experimental input is 3.5, there is no solution to this NAMFIS case under all the constraints in **Figure 63**, because the averaged calculated J values from the 5 conformers can never go below 4.0. However, the current version of NAMFIS lacks this function. It still produces an output with an SSD value just like a normal run, but clearly the result will violate at least one constraint and thus compromise reliability. In this situation, the user should be informed by the program in order to decide whether to re-run the NMR experiment to obtain a new experimental value for this particular parameter or improve the conformational search result. However, without proper acknowledgement from the current version of the program, the user can rarely discover that the input case is mathematically infeasible. The program needs improvement on dealing with mathematical infeasible cases.

7.2. New Generation of NAMFIS

7.2.1. Java VS Python

The new generation of NAMFIS program is being developed using Java programming language instead of Python. Unlike Python, which is susceptible to the cross-platform problem, the Java program is platform-independent as long as the proper Java Runtime Environment (JRE) is properly installed on the machine. Java is also a programming language which requires compiling before execution. The compiling step packs all the dependent libraries together for runtime use, avoiding end users having to resolve the

dependency issue by searching all the required libraries online. In addition, various Java libraries are also available for different tasks which minimizes the job of “re-inventing the wheel” during the development period. For this version of NAMFIS program, Java was chosen as the programming language.

7.2.2. GUI and Backend Design

Instead of modifying an old input file to generate a new one, the new program uses a GUI as the input. This avoids the problem of ill-formatted input text file and simplifies the capture of various user input error. The GUI was constructed by using the Swing components of the Java library. It complements a similar GUI written by Aaron Padwa in our laboratory during a summer internship.

Figure 65 demonstrates the GUI for the new NAMFIS program. At the data input regions, four tabs are available for inputting atom indexes, permutations, experimental NOE distances and J coupling constants. A table is placed besides the input area to list the added inputs and a button is available to select and remove wrong inputs. The input parameters can be saved to a file which is compatible with the old program’s format and then later reloaded for calculation or modification. Conformational files (currently only sdf format is supported) are chosen by clicking the button and select the desired input files from the pop-out dialog box. After completing the experimental input and selecting the conformational file, the NAMFIS job can be executed by clicking the “Execute NAMFIS” button. The text area below serves as a notification board to provide end users with the current status, progress, results and error messages.

At the backend of the program, the atom mapping, permutations, various experimental input and other required parameters are stored in a specific class “NAMFISAllData” which includes regular data structures like ArrayList, Array, HashMap or other defined classes. The data inside the class are updated simultaneously with the tables on the GUI. When executing a NAMFIS calculation, the program reads directly from this class and processes the data instead of reading from an input file. The external library for processing chemistry-related information is JChem developed by ChemAxon.⁷⁴ The library contains classes for structure import/export, geometry calculation, structure comparison, sdf file modification and so on. The library performing the constraint non-linear minimization is discussed in the following section. After the calculation completes, if a feasible solution with the best fit (lowest SSD) is identified, the corresponding set of conformers and their mole fractions are extracted into a new sdf output file.

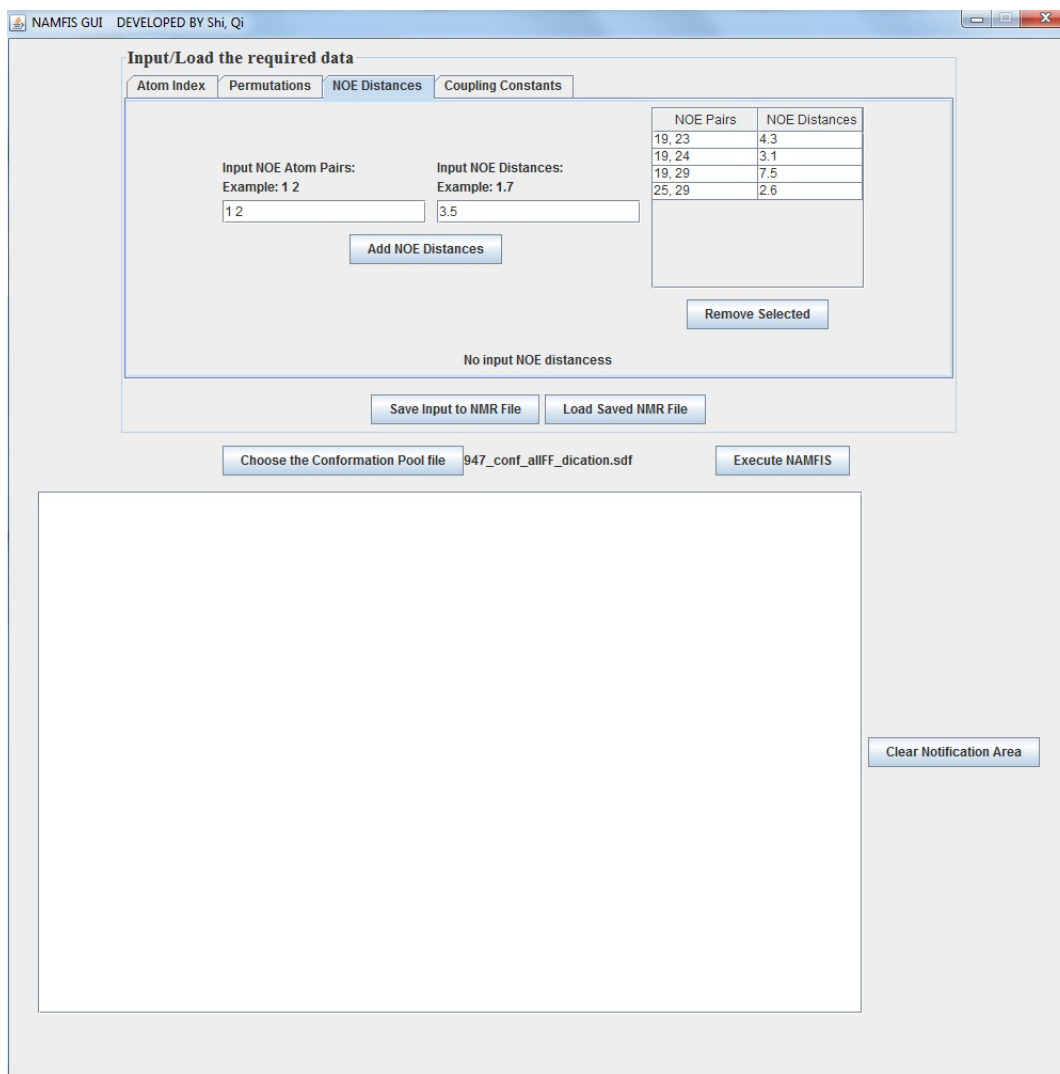


Figure 65. The GUI for the new NAMFIS program.

7.2.3. Minimization Step

The most important and challenging part of designing this program is implementing the constrained minimization of SSD. This task is executed using the Java interface from the IPOPT library.⁷⁵ In this program, the “NAMFISOptimizer” class extends the abstract “IpopT” class and then overrides the required abstract method by supplying the target minimization function $f(x)$ (in this case SSD), the gradient matrix of $f(x)$, the constraint

functions $g(x)$ (in this case the inequality and equality constraints in **Figure 63**), the Jacobian matrix of $g(x)$ and the Hessian matrix for the Lagrangian function $f(x) + \lambda * g(x)$. The output for mole fractions is provided by an array, and the index of the array element corresponds to the conformer index in the input conformation file.

7.2.4. Exception Handling

A significant portion of this program consists of exception handling and error checking. At the four input regions of the GUI, each input must be valid before being added to the table. For every input of atom index, the program checks for and then rejects duplicates. For the input of permutations, NOE distances and J couplings, the corresponding atom index must be a valid atom index (included in the “atom index” tab). In addition, no duplicate or negative experimental value is allowed. There are also codes for dealing with the wrong formatted input files with hints specifying the locations of the errors.

When the “Execute NAMFIS” button is pressed, the program will first check to see if it lacks any required input parameter. Then, the first structure from the conformation pool input file is extracted and used as a reference to check for mismatching input J coupling types or atom order. The input error mentioned in the second paragraph of 7.1.3 (wrong order of atom set for dihedral angle calculation) can be identified at this step. The program will also ensure the other structures in the conformation pool file are the same molecule. Any error captured during this process will terminate the execution of NAMFIS and provide the end user information on how to correct it. **Figure 66** demonstrates an error detected at the input tab region (duplicated input) and another in the notification text area when trying to execute the NAMFIS. For the first error, as shown in **Figure 66**, the user has put the NOE coupling pairs “19 23” again, although the

NOE distance value is different. The program captures the duplicate and reminds the user. For the second error, the correct order for that input J coupling atom set should be 23 9 10 25. The NAMFIS execution is discontinued.

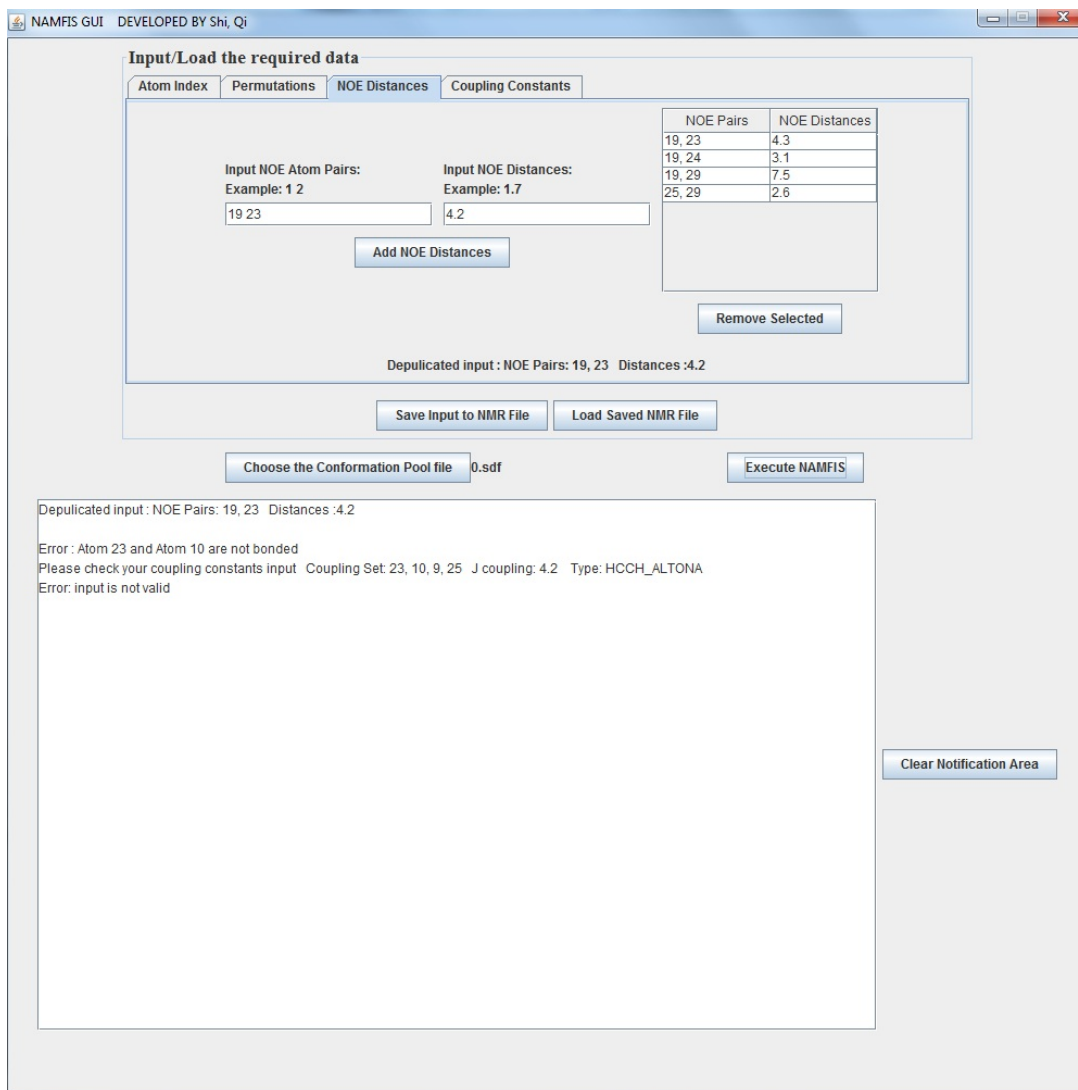


Figure 66. Input errors identified by the program.

As stated in Section 7.1.3, the old version needs improvement on handling the “mathematical infeasible” situation. This is often caused by the tight constraints imposed by at least one experimental parameter. In the new Java version of NAMFIS, the user will notice this in the notification area if the “Optimization Status” is not equal to 0. Before

executing NAMFIS, the program will also perform a check to see if there are some parameters where all the calculated values are already off the boundaries. Use the same example in Section 7.1.3. Assume the input conformational pool consists of 5 conformers with calculated J values of 4.0, 4.1, 4.2, 4.3 and 4.4. If the allowed upper bond derived from the experimental input is 3.5, the program will detect this obvious violation even before running the minimization and list the information at the notification text area. In addition, the user will also know which constraints are the possible problems by examining all the parameters when the calculated values are precisely at the allowed boundaries. The experimental data can then be re-examined easily.

7.3. Conclusion and Future Work

The second generation of NAMFIS program written in Java avoids several pitfalls in the old python version, which can result in run-time errors or incorrect results. The cross-platform ability for this version of NAMFIS is improved significantly due to Java's excellent cross-platform ability. In addition, it successfully captures some previously unnoticed user input errors in the old NAMFIS input files. It also provides solutions to the mathematical infeasibility issue, which is not addressed in the old version of the program.

Some features still need to be implemented before release to the public. For example, the Java version currently does not support the input of chemically equivalent atom sets (example: three protons on a methyl group). Implementing this function requires changing the fundamental data structures and additional exception handling codes. The permutation feature, which is currently turned off in the Python version, is also not

available in the Java version. Additional work may also be needed for the instructions and layout of the GUI. Nevertheless, the newer version provides a friendly user interface that is less susceptible to user input errors and cross-platform issues. The author expects a wider usage of the new NAMFIS program once published.

Chapter 8: Experimental Details

This chapter provides a detailed description of the techniques, software, codes and algorithms employed for the computational procedures in this dissertation. Since the procedures are essentially identical for a specific task across different projects, rather than repeating the same description for each project, this chapter is focused mainly on providing general but applicable descriptions of specific tasks. Each captures enough detail for a properly-trained person to perform the computational manipulations.

8.1. Pipeline Pilot Tasks

This section covers the critical settings for Pipeline Pilot to execute FRESH applications.

8.1.1. Filter for desirable fragments/compounds by substructure matching

One frequently executed task in the FRESH program is substructure matching. For example, when querying for building blocks in a commercial database, searching for groups with potential liability, stability or reaction concerns, excluding literature structures as described in the FRESH validation case studies and obtaining the rank of literature structures, an input file is queried against one or multiple structures.

Substructure matching is accomplished with the “Substructure Filter from File” component in Pipeline Pilot. Two parameters need setup before the execution of the program. First, the file containing the queried structures (building blocks, literature structures etc.) must be specified by the “Source” parameter. The queried structure file can be manually prepared by ChemDraw in 2D format and saved as an sdf format file. Second, depending on the specific task, the “MatchType” option has to be correctly

selected. The default “AllQueries” option is actually never used in FRESH. For building block collection, literature compound identification or any other tasks which require saving matched structures to be further processed, the correct option is “AnyQuery”. On the other hand, for tasks like excluding literature compounds or removing structures with liability concerns, the “NoQueries” option should be used. **Figure 67** lists the substructures with potential liability, stability and reactivity concerns that are used in all the appropriate FRESH actions.

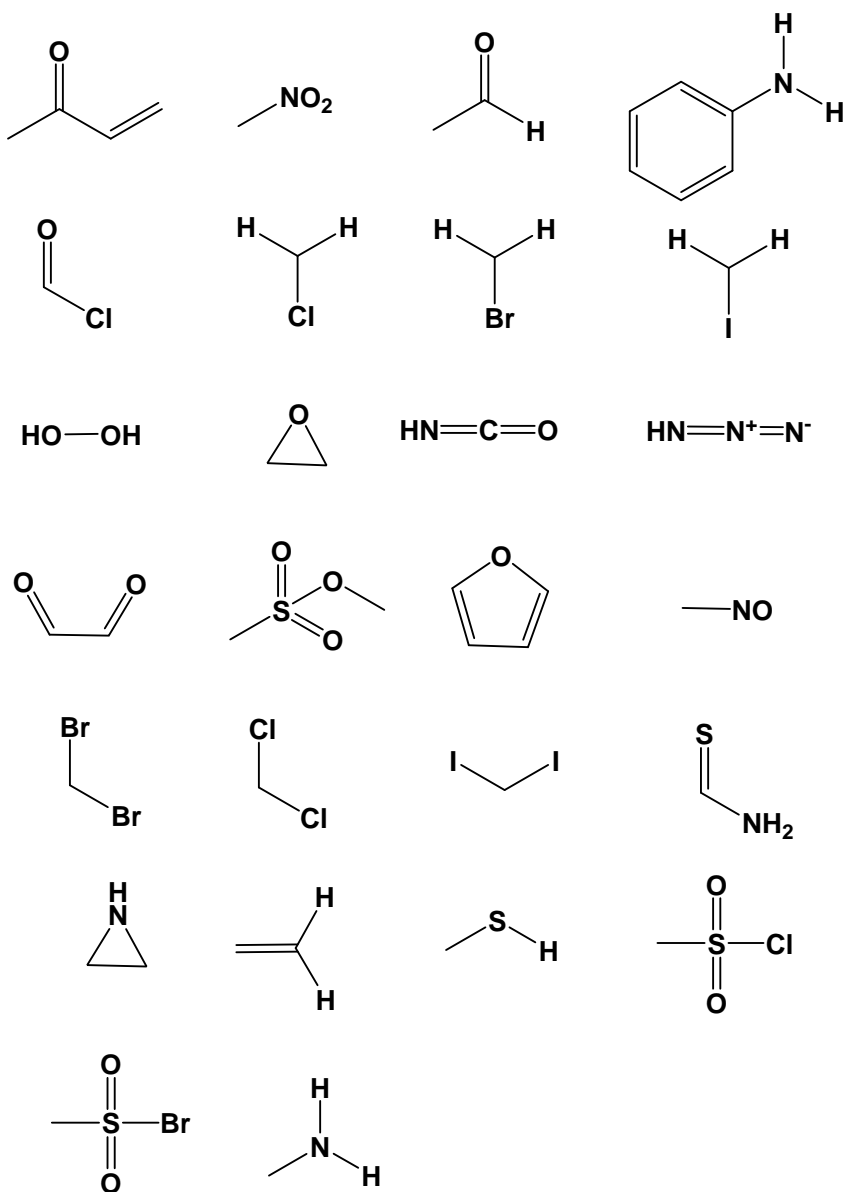


Figure 67. Substructures with potential liability, stability and reactivity concerns.

8.1.2. Select fragments/compounds according to physical/ADMET property values

Selections based on particular property values are the most frequent operations in the FRESH program. A selection based on Rules of Thumb (3, 4, 5 etc.), calculated physical/ADMET properties or QSAR scores is implemented by this operation. When establishing QSARs, the grouping of training and test sets also requires this operation.

The “Property Range Filter” and “Property Value Threshold Filter” are two possible components to employ for this task. However, the use of PilotScript in a “Custom Filter (PilotScript)” component provides a more flexible and convenient approach. For example, if the property to be selected is molecular_weight (less than 500) and logP (between 3.0 and 3.5, both inclusive), the PilotScript can be formatted as: “molecular_weight < 500 and logP >= 3.0 and logP <= 3.5;”.

The general fragment selection filter is expressed by the following PilotScript language: “molecular_weight < 301 and (N_Count + O_Count) <= 3 and Num_H_Donors <= 3 and Num_positiveatoms == 0 and Num_negativeatoms == 0 and alogP <= 3;”, which implements the Fragment Rules of Three and requires no permanent charges. The building blocks containing bridghead atoms and spiro atoms are usually costly and frequently require “made on request”, so additional filters are added by “Num_BridgeHeadAtoms == 0 and Num_SpiroAtoms == 0;”.

The general compound selection filter is expressed by the following PilotScript language: “Molecular_weight < 501 and (N_Count + O_Count) <= 10 and Num_H_Donors <= 5 and AlogP <= 5 and Num_rotatablebonds <= 10 and Molecular_solubility >= -6 and i_qp_qplogPotow <= 5 and r_qp_PSA <= 120 and i_qp_nummetab <= 6 and r_qp_qplogS >= -6;”, which covers the Lipinski Rules of Five, Jorgensen Rules of Three and polar surface area. They are routinely incorporated in FRESH unless one specific term removes all the structures or uses another range required in the project. The property name for the calculated blood brain barrier penetration is “r_qp_qplogBB”, which is preferably greater than 0 for CNS drugs.

8.1.3. Process the commercial library to obtain synthetic fragments

Obtaining synthetic fragments from a commercial library requires marking the connecting point at the specific site according to the synthetic scheme provided by the chemists. As required by FRESH, no manual operation step should be involved at this point. The “Perform Reaction on each Molecule” component is designed to accomplish this task.

Like the “Substructure Filter from File” component, it requires a source file to specify how to transform the structure from a building block to the corresponding fragment. **Figure 68** demonstrates an example (from HDAC1 case study) of how to specify the transformation from a phenyl acid chloride to the required fragment. This transformation can be prepared in ChemDraw and then saved as an .rxn format file. The Z1 specifies the attachment point to be connected to the core structure (see the section below). The parameter for “IfMultipleReactionsPossible” should be set as “PerformEachReaction”. For a symmetrical molecule, this setting may generate duplicate fragments, which are removed afterward (see section 8.1.5).

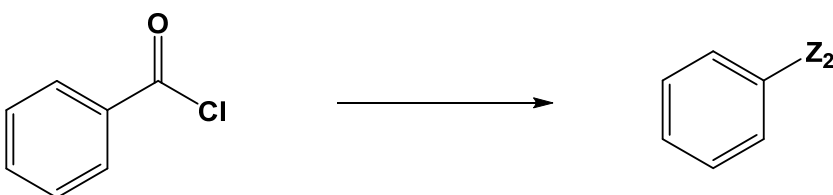


Figure 68. An example of a transformation file.

8.1.4. Covalently attach fragments to the core structure

The attachment sites on the core structure are marked with R₁, R₂, R₃ etc.. They correspond to the attachment points for fragments marked with Z₁, Z₂, Z₃ etc.. The core

structure is usually prepared directly with ChemDraw and saved as an .sdf file. **Figure 69** illustrates the core structure for the CA II case study.

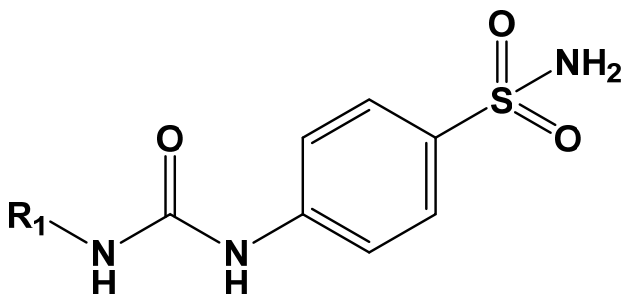


Figure 69. The core structure for the CA II case study.

The covalent attachment of fragments to the core structure is performed by the “RG Writer” and “RG Reader” components. First, the core structure and all fragments are stored together in an .rg format file by the “RG Writer” component at the specified location. Then, the “RG Reader” component takes the saved rg file and enumerates the combinatorial library. The parameter for “WhatToOutput” should be specified as “EnumeratedMolecules” to obtain the desired molecule structures. The initial library construction is completed once this step is finished, and the generated sdf output file (each structure is properly numbered) is applicable to various computational software programs or filters further down the pipeline.

8.1.5. Remove duplicate structures

Duplicate structures should be removed routinely along the workflow to avoid unnecessary computational cost. In addition, if one structure contains multiple result entries and only one should be kept for further processing (example: select the best docking pose among all 5 poses), it requires removal of other entries with the same structure. The removal of duplicates can be achieved by the “Remove Duplicate Molecules” component. This component employs a short PilotScript to detect and remove

duplicates. The default judgment criterion for duplication is the calculated “Canonical_Smiles”, which is a unique 2D string representation of a specific structure. However, the criterion can be adjusted to other terms like the unique molecule ID number implemented in the FRESH program.

8.1.6. Merge data

At some point in the FRESH workflow, the calculated results for the same molecule from various software programs need to be merged. This task is accomplished by the “Merge Data” component. The most crucial parameter for this component is “MergeUsing”, which specifies the shared property name between different files. Incoming data records are merged into a single data record if they have the same value for the specified “MergeUsing” property. Similar to the “Remove Duplicate Molecules” component, the “MergeUsing” can be set to use the calculated canonical_smiles or the ID numbering system implemented in FRESH.

8.1.7. Program debugging examples

As stated in Section 2.2, the numbers displayed along the pipeline provide crucial information for program debugging. Two debugging examples are provided in this section to demonstrate the usefulness of the displayed number. The two examples are the most frequent errors a user will encounter during the construction of a FRESH workflow for a particular project. It is worth mentioning that this section only covers a tiny portion of the debugging options. An end user of FRESH is advised to participate in the professional Pipeline Pilot training or acquire procedure-based paradigm programming experience to be prepared.

As illustrated in **Figure 70**, the intended task is filtering of molecule fragments and removing those with unfavorable substructures specified in the FlaggedGroup.sdf file. However, in this example all structures went to the “fail” port of the Substructure Filter as demonstrated by the red number “2670”, which is extremely unlikely. Upon further examination, the “MatchType” parameter was incorrectly set at the “AllQueries”, which is the default option. The correct setting, as stated in Section 8.1.1, should be “NoQueries”.

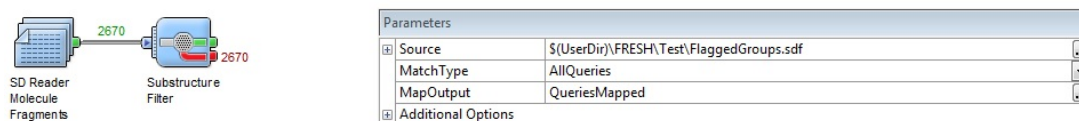


Figure 70. Debugging case 1: wrong parameter setting.

Figure 71 illustrates a more tricky case involving an undesirable runtime without obvious symptoms. The workflow attaches three fragment pieces to the core structure and then enumerates all molecule structures to perform some calculations. The workflow appears to be running smoothly. Nevertheless, a serious problem can be identified based on the information provided by the numbers. The estimated number of structures to be enumerated is approximately $3 * 10^{10}$ derived from the product of 1133, 9831 and 2440. In this example, roughly 10^3 structures have already been processed in 86 seconds, so the daily processing speed is approximately 10^6 structures. However, the $3 * 10^{10}$ target structures would require ~30,000 days (longer than a typical life span of a man) to complete the program, which is obviously erroneous. Runtime errors like this should be avoided at the early stage. Possible solutions include performing fragment filtering before enumeration, arranging other components with fast calculation speed ahead of the slow one and clustering of fragments.

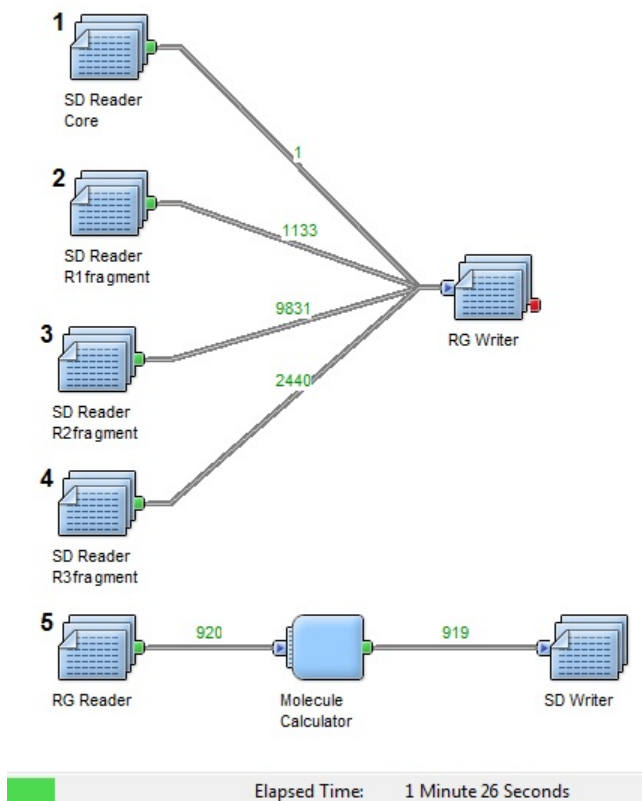


Figure 71. Debugging case 2: extremely long runtime.

8.2. Glide Task

8.2.1. Prepare protein receptor for docking and sequence comparisons

The X-ray crystal structure from the PDB database requires preparation work before it can be employed by Glide or MM-GBSA for evaluation of ligand-protein interactions. The preparation is performed by the Protein Preparation Wizard workflow included in the Schrodinger Maestro package. The corresponding PDB structure is downloaded directly from the PDB database or imported from the local file system. All water molecules are deleted. For PDB structures with multiple identical chains, only one chain is kept for further processing. The structure is then pre-processed by adding hydrogens, replacing missing side chains and assigning bond orders using the default options. Optimization of

the protein's hydrogen bond network and assigning the ionization state of the protein side chains are then performed by the H-bond refinement tool using the default options. A final minimization step is performed with OPLS 2005 force field. The resulting structure is now suitable for Glide, MM-GBSA and sequence comparisons.

8.2.2. Generate receptor docking grids around ligand binding sites

The receptor docking grid specifies the location for the ligands to be docked. Generation of the grid is completed by the "Receptor Grid Generation" program embedded in Glide. The ligand included in the crystal structure is then selected for the "enclosing box" option under the "Site" tab. The "Centroid of Workspace ligand" and "Dock ligands similar in size to the Workspace ligand" options are selected. For the CA II FRESH validation case (Section 3.3.2), additional constraints were added under the "Constraints" tab. These required an H-bond with the NH backbone of Thr198 and the nitrogen of the sulfonamide to coordinate with the zinc atom at the catalytic site. All the remaining options used the defaults.

8.2.3. Set ligand docking parameters to control output

The setup of Glide ligand docking parameters is the last step before executing a Glide docking. The generated receptor docking grid was selected under "Receptor grid". The SP (standard precision) option was chosen for the docking precision. The ligands were docked flexibly for all Glide docking tasks in the dissertation.

8.3. MM-GBSA Task

8.3.1. Parameter settings for energy refinement

The MM-GBSA tasks were performed after the Glide docking exercise and took the output pose files directly from the Glide docking output. All the other parameters were the default values.

8.4. Miscellaneous Schrodinger Tasks

8.4.1. LigPrep – 2D to 3D structure conversion

The structures originating from a commercial library or FRESH initial library are generally in 2D format without explicit hydrogens. The LigPrep program takes these structures as input and generates 3D ligand structures suitable for QikProp, Glide and MM-GBSA. The ionization state is generated by Epik at the physiological pH (7.4). For generating the input file for QikProp, the option “Do not change” is selected. The output format is specified as “SDF”. All other parameters used the defaults.

8.4.2. QikProp – physical/ADMET property estimates

The QikProp program requires neutral 3D input ligand structures. The ligands prepared by LigPrep are used as input for QikProp and the “Fast mode” was selected.

8.4.3. Protein sequence alignment

Sequence comparison for protein receptors (Section 6.2) requires the 3D structure alignment of the receptors. This is completed automatically by the Protein Structure alignment tool included in the Maestro package. The “Reference residues” option is set at “all”. To facilitate the sequence comparison around the binding sites (Section 6.2.3), the selected residues 7 Å away from the bound ligand are colored red while other residues are colored black.

8.4.4. Homology modeling

For the PI3K α case study (Section 3.2.1), a homology model was generated using the Prime – Homology Modeling program. The task followed the established “Structure Prediction Wizard” procedure. The sequence of PI3K α was obtained from uniprot (id: P42336 ⁷⁶) and the structural template employed was the crystal structure of PI3K γ (PDB code :1E8Z ³⁷). The default alignment and the “model built” method were used.

8.4.5. Induced fit docking – Flexible ligand and protein interaction

The SNRI project (Section 4.3) involves induced fit docking. This task was performed by the established “Induced Fit Docking” workflow included in the Maestro package. The pdb-formatted structures of NET and SERT provided by Dr. Spandan Chennamadhavuni were used as the receptor structures. For NET, the grid box was centered at (27.1, 31.6, 21.4). For SERT, the corresponding coordinates were (27.1, 32.2, 21.6). Two ligands, **5** and **6** in **Table 1**, were used for the docking. All other options were taken as default.

8.5. Case-specific Details

8.5.1. PI3K α case study

The FRESH protocol for this case study starts with the construction of QSARs. All the Pipeline Pilot components are in version 7.0. The molecular data set for establishing training and test sets was obtained from the ChEMBL database. The corresponding ChEMBL ID for PI3K α is ChEMBL4005 and the UniProt accession code is P42336. The “Activity Type for Target” was selected as “IC₅₀”. The data set file was downloaded in “xls” format, which included ~2,200 available activity records at the time of the project. The “xls” file format can be directly accessed by the Microsoft Excel program (Version

2010), but it is not suitable for the FRESH workflow as the required format is “sdf”. The component “Molecule from SMILES” was thus employed to convert the “xls” file into the desired “sdf” format by using the property “Canonical_smiles” included in the database. A numeric ID was assigned to each structure with the property “Name” starting from 1.

In the ChEMBL database, the activity value is recorded under the “STANDARD_VALUE” property in nM units. The entire data set was then sorted by the “STANDARD_VALUE” in increasing order to facilitate subsequent processing. As required, the literature compounds should be excluded when generating QSARs except when the compound is either the starting molecule or generated earlier than the publication year. They were removed from the data set by applying the corresponding “PubmedID” property, which in this case equals 21388141. Compounds generated in or after the publication year of the article (2011) were also excluded by the property “YEAR”. Duplicate molecules were then removed using the default “Canonical_smiles”. The activity record sometimes appears as an inequality, for example, $IC_{50} > 100$ nM and is specified in the “RELATION” property (“=” for equality and “>” or “<” for greater or smaller) in the ChEMBL database. For defining active vs. inactives in this case study, molecules with an “=” or “<” relationship and a “STANDARD_VALUE” ≤ 10 nM were defined as actives, while molecules accompanied by an “=” relationship and “STANDARD_VALUE” > 10 nM plus those with a “>” relationship and a “STANDARD_VALUE” no less than 10 nM and were defined as inactives. A new property named “Activity” was added for both the actives (value is 1) and inactives (value is 0) to facilitate further processing.

The data records for actives and inactives were combined and then sorted by “STANDARD_VALUE” in ascending order. All chiral molecules were excluded and the rest of the data records were numbered increasingly starting from 1 (Property name “BayesianGroup”). The following PilotScripts group the records to training and test sets in a 2:1 ratio: “BayesianGroup:= #BayesianGroup_id ; #BayesianGroup_id ++ ; if (#BayesianGroup_id == 4) then #BayesianGroup_id := 1; end if; ” and “BayesianGroup <= 2” for the training set. The hit compound (**3e** in the article) was also added to the training set and defined as “inactive”. The neutral form of each molecule in the training set was then submitted for the “Learn Good From Bad” components for generating Bayesian activity scores. The “TestForGood” option was set at “Activity == 1” and the molecular descriptor was assigned as ECFP_4. After executing this step, a new Pipeline Pilot component was generated, which can be directly incorporated like all other components into the workflow to calculate the Bayesian scores.

All calculations using the Schrodinger Maestro package (Glide, MM-GBSA, QikProp, LigPrep, etc.) were performed in the 2012 released. The Glide and MM-GBSA scores for these ChEMBL molecules were obtained using the procedure described in Sections 8.2. and 8.3. The receptor structure was obtained directly from the PDB database (PDB code: 1LUG). After obtaining the results from Glide or MM-GBSA, only the best score for each molecule was retained and used for further evaluation. This was achieved by first sorting the corresponding score (“r_i_docking_score” for Glide and “r_psp_MMGBSA_dG_Bind” for MM-GBSA) and the “Name” property, then removing the duplicate records by the “Remove Duplicate Molecules” component with the “Name”

property as the criterion for duplication. The QSARs generated in these steps were then employed for potency evaluations.

The next step for FRESH is processing building blocks from the commercial database to obtain the corresponding fragment. The commercial database for this case is the “ZINC bb now”, which is available at <http://zinc.docking.org/subsets/zbb-now>. One major difference in the implementation of FRESH for this case study is the substructure matching for building blocks. As stated in Section 3.3.2, the building blocks are aromatic bromides, in which the bromine atom is attached directly to the aromatic ring (mentioned as Ar-Br below). In this case study, the heterocyclic and fuse-aromatic rings in addition to the phenyl ring were explored. However, the “Substructure Filter from File” component does not support a query structure like “Ar-Br”. At the time of this project, the solution was to use “C-Br” as the query structure with an additional PilotScript filter “Num_aromaticrings > 0”. However, this also allowed structures like “Ar-CH₂Br” to be included. These structures were removed at a later stage of FRESH by another “Substructure Filter from File” component as structures with only “phenyl-CH₂Br” showed up. A more recent solution is to use the case-sensitive regular expression matching provided by the PilotScript, since the atoms in an aromatic system are lower-case letters in the Canonical_smiles string of the structure. The PilotScript code is provided here: Canonical_smiles CASE LIKE '%c[1-9](Br)%' or Canonical_smiles CASE LIKE '%c(Br)%' or Canonical_smiles CASE LIKE 'Brc%' or Canonical_smiles CASE LIKE 'Brc%' or Canonical_smiles CASE LIKE '%c[1-9]Br' or Canonical_smiles CASE LIKE '%c[1-9]Br%' or Canonical_smiles CASE LIKE '%cBr' or Canonical_smiles CASE LIKE '%cBr%'. After obtaining the building blocks, the

conversion from building blocks to the corresponding fragment is straightforward by using the “Perform Reaction on each Molecule” component. Unfavorable fragments were removed by the procedures discussed in Sections 8.1.1 and 8.1.2. For this particular case, additional criteria for selecting a fragment were added by: “I_count + Cl_count + Br_count = 0 and S_count == 0” to reduce the chance of a failed Suzuki coupling. Some silicon-containing compounds appeared on the list during the test stage. They were removed due to potential toxicity concerns by the PilotScript “Si_count == 0”.

The corresponding molecular structures were obtained by covalently attaching the fragments to the core structure shown in **Figure 4** following the procedures in Section 8.1.4. The list of molecules were prepared by Ligprep (Section 8.4.1) then subjected to a series of filters based on physical/ADMET property calculation results from Pipeline Pilot and Qikprop following the procedures described in Sections 8.1.3 and 8.4.2. The Bayesian and Glide scores were obtained using procedure similar to that for the ChEMBL molecules when establishing the QSARs. The final list was sorted by the Bayesian score in ascending order and **19d** was the 3rd on the list.

The iteration step permitting identification of **19k** was performed by first calculating the Tanimoto similarity using the “Fingerprint Similarity” component. All structures were subjected to the same filters as the initial stage with the exception of “N_count + O_count <= 7”, which was designed for this particular iteration.

8.5.2. CA II case study

The FRESH workflow was constructed similar to the PI3K α case study. All the Pipeline Pilot components were in version 7.0, and all calculations using the Schrodinger Maestro package (Glide, MM-GBSA, QikProp, LigPrep, etc.) were performed in the

2012 release. The corresponding ChEMBL ID for CA II is ChEMBL205 and the UniProt accession code is P00918. The “Activity Type for Target” was selected as “Ki”. The downloaded “xls” data set file contained ~6,200 available activity records at the time of the project. The data set was converted to “sdf” format using the same method for the PI3K α case by the “Molecule from SMILES” component. A numeric ID was assigned to each structure with the property “Name” starting from 1. Some data records had an empty canonical_smiles, so they were excluded by the PilotScript “smiles != ””. All literature compounds were excluded from the QSARs by the “Pubmed_ID” property (not equals to 21361354 or 20922253) except one historical compound AAZ (Name == 5906). The publication year cutoff is 2010 (Year < 2010) and all duplicate structures were removed by the “canonical_smiles” property. Treatment of the inequality relationship, defining actives vs. inactives, grouping molecules for training/test set and collection Glide and MM-GBSA scores was identical to the PI3K α case.

After construction of QSARs, the next step for FRESH is to query building blocks in the commercial database to obtain the corresponding fragments. As indicated in Section 3.3.1, the building blocks are phenyl iso-cyanates or acid chlorides. Two separate “Substructure Filter from File” components with two “Perform Reaction on Each Molecule” were employed to obtain the corresponding fragments and duplicates were removed. A temporary fragment file was saved for the 2nd iteration. Unfavorable fragments were removed as discussed in Section 8.1.1.

Target molecular structures were then obtained by covalently attaching the fragments to the core structure. All molecular structures were then subjected to the QSAR score

filters and calculated physical/ADMET properties similar to the PI3K α case. Compound **30** emerged from the final list as one of the top 5 compounds.

During the first iteration step (Section 3.3.3), the logP of the side chain for each input structure was obtained. Unfortunately, the calculation of side chain logP is currently not supported. The present work-around is to extract the side chain from the entire structure, perform the calculation, then merge the results back into the final target structures. The workflow in **Figure 72** demonstrates the process. Each input structure is assigned a unique numeric ID named “MergelogPID”. The workflow then diverges. The bottom part extracts the side chain structure by a “Perform Reaction on each Molecule” component introduced in Section 8.1.3. After the calculation of logP completes, the result is merged to the upper part of the workflow by a “Merge Data” component. The previously created “MergelogPID” property is used to merge data. With the available side chain logP values, all the structures are submitted for further processing.

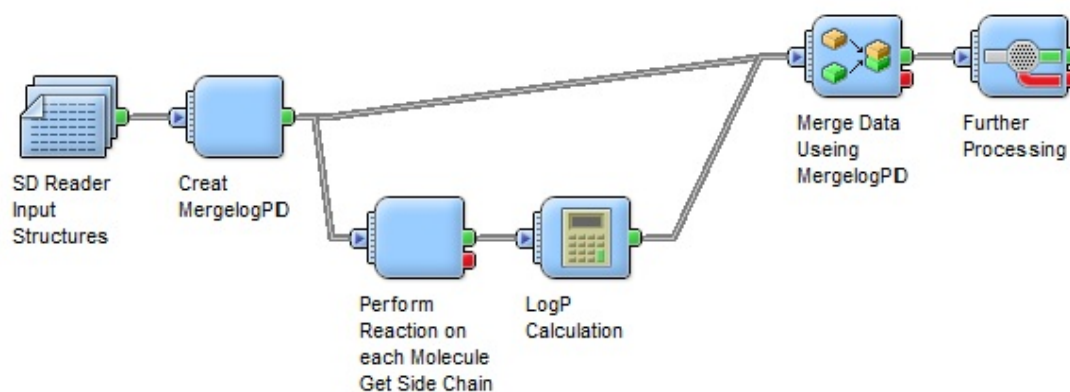


Figure 72. CA II, first iteration demo

The 2nd iteration prioritized all the 2-substitution-only structures with retained hydrophobicity and similarity. The fragments were obtained directly from the saved temporary files and subjected to the same selection criteria for the fragment except the

alogP term was adjusted to “alogp \geq 2.873 (value from the fragment for **3**)” to fit the “retained hydrophobicity” requirement. All 2-substitution-only fragments were covalently attached to the core structure using the “Enumerate using RGroups” components to obtain the target molecular structure. The structures were selected by the same physical/ADMET property requirements (Section 8.1.2) and sorted by Tanimoto similarity with **3**.

8.5.3. HDAC1 case study

The FRESH workflow was constructed similar to with the other two case studies. All the Pipeline Pilot components are in version 7.0. The corresponding ChEMBL ID for HDAC1 II is ChEMBL325 and the UniProt accession code is Q13547. The “Activity Type for Target” was selected as “IC₅₀”. The downloaded “xls” data set file included ~2,300 available activity records at the time of the project. The data set records were converted to “sdf” format using the same method employed for the PI3K α case by the “Molecule from SMILES” component, and numeric ID was assigned to each structure with the property “Name” starting from 1. All literature compounds were excluded from the QSARs by the “Pubmed_ID” property (not equals to 20451378). The publication year cutoff is 2010 (Year < 2010) and all duplicate structures were removed by the “canonical_smiles” property. Treatment of the inequality relationship, definition of actives vs. inactives, grouping of molecules for training/test set and gathering of Bayesian scores was identical to the PI3K α case. This is a ligand-only case study, so Glide and MM-GBSA scores were not involved in building QSARs.

The core structure for this case is shown in **Figure 73**. This case study explored R₁ and R₂ fragments simultaneously. The R₁ fragments were prepared using an approach

identical to the CA II case by querying the phenyl acid chloride building blocks in the “ZINC bb now” commercial database and removing the unfavorable fragments. The R_2 fragment linker chain lengths vary from 1 to 7 (**Figure 74** for an example) and were prepared manually by Chemdraw 2013. The “RG Writer” and “RG Reader” components mentioned in Section 8.1.4 are able to enumerate R_1 and R_2 fragments simultaneously.

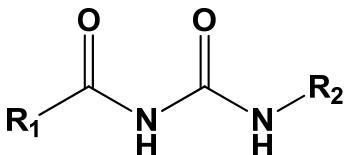


Figure 73. The core structure for the HDAC1 case study

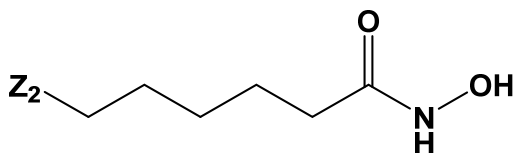


Figure 74. The fragment for the R_2 group with a carbon linker length of 5

The enumerated molecular structures were selected similar to the physical/ADMET criteria discussed in Section 8.1.2, and Bayesian scores were obtained from the constructed QSARs. No Glide or MM-GBSA scores were involved in this case.

The iteration step prioritized structures with 5 carbon length linkers as the R_2 group. The length of the linker is calculated by the number of rotatable bonds. Pipeline Pilot does not support such calculation for a particular side chain. However, an effective solution is similar to the first iteration of the CA II case. As demonstrated in **Figure 75**, the workflow diverges. The bottom part extracts the corresponding R_2 side chain by a “Perform Reaction on Each Molecule” component with a reaction shown in **Figure 76**. The structures with the desired 5 carbon length were selected by the PilotScript “Num_rotatablebonds == 5” and marked with a new property named “isFiveCarbon”.

The results were merged into the upper workflow and structures with the desired carbon linker length were selected by using “isFiveCarbon is defined”.

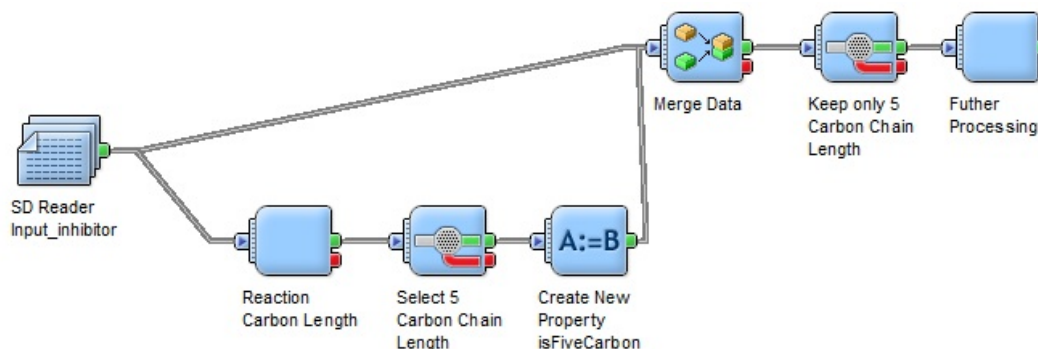


Figure 75. The iteration for HDAC1 case

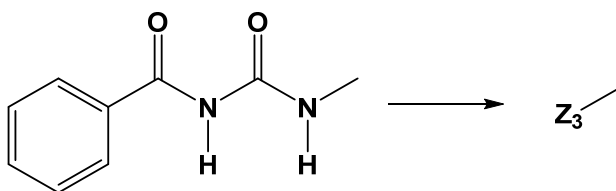


Figure 76. The fragment for the R₂ group with a carbon linker length of 5

8.5.4. SNRI application case

The FRESH workflow for the SNRI project was constructed in a fashion similar to the case studies. The version of Pipeline Pilot used in this project was 6.0 (student version under academic license), and the Schrodinger Maestro Package was the 2009 release.

The QSAR in this application case was constructed from the MM-GBSA scores of the six preliminary compounds listed in **Table 1**. The six ligand structures were manually prepared by Chemdraw 2012 and submitted to LigPrep (see Section 8.4.1) to generate 3D structures. Glide SP docking was performed to generate docking poses for each ligand. The receptor structure used in this Glide docking was generated from the induced-fit docking (See Section 8.4.5). The Glide docking pose file was then submitted to MM-

GBSA program, and the best MM-GBSA score for each ligand was used to establish the QSAR lines depicted in **Figures 29** and **31**.

Construction of the target molecular library started from a query of building block structures in the ChemNavigator building block library (available in our lab, requested by previous lab member Dr. Andrew Prussia from ChemNavigator). The building block structures are listed in **Figure 77**. After converting the building blocks to the fragments, duplicates were removed by the “Canonical_Smiles” property and fragments with unfavorable substructures were removed. The remaining fragments were covalently attached to the core structure depicted in **Figure 78**.

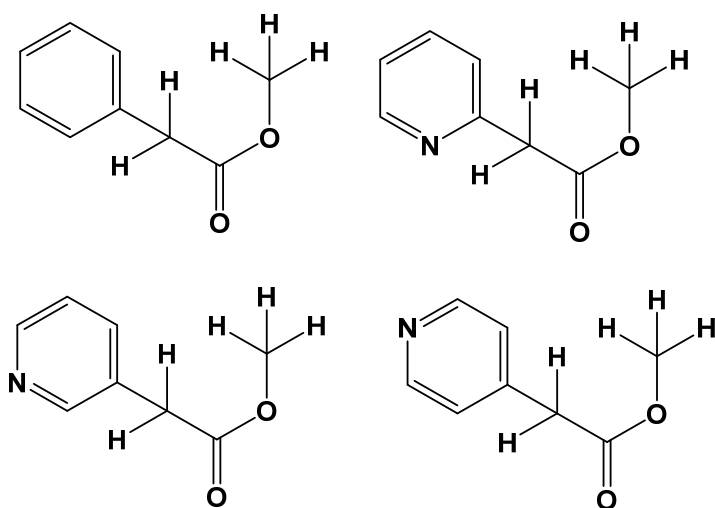


Figure 77. The building block structures for the SNRI project

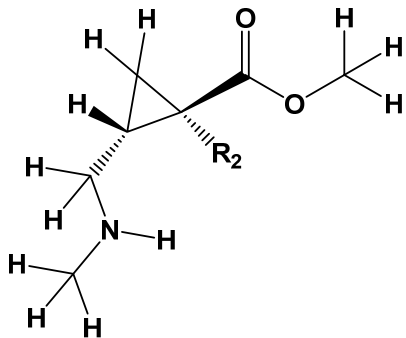


Figure 78. The core structure for SNRI project

The enumerated library was subjected to the following criteria : “(N_Count + O_Count) <= 10 AND Num_H_Donors <= 5 AND AlogP <= 5 AND Molecular_PolarSurfacearea <= 90 And Molecular_solubility >= -5.0 And Molecular_weight < 501” and “

i_qp_nummetab <= 3 and r_qp_qplogbb >= 0 and r_qp_qppmdck >= 500 and r_qp_qppcaco >= 500 and i_qp_CNS > 1 and r_qp_percenthumanoralabsorption >= 80”

One additional filter was employed to prioritize structures with no less than 40% sp³ carbon. It was implemented by the following Perl code (use “Perl Molecule Calculator” component, developed by former lab member Serdar Kurtkaya) and PilotScript:

```
use strict;

use pilot(':constants');

use pilot::chem(':all');

sub onInitialize {

    my $context = shift;

    my $params = $context->getComponentParameters();
```

```

#my @output = $params->getByName("Output")->getValue();

#my %hash;

#@hash{@output} = @output;

#$:do_spiroAtoms = defined $hash{"Num_SpiroAtoms"};

#$:do_freeSpiroAtoms = defined $hash{"Num_FreeSpiroAtoms"};

return READYFORINPUTDATA;

}

sub onProcess {

    my $context = shift;

    my $data = shift;

    # get molecule and props

    my $mol = getMolecule($data);

    my $props = $mol->getProperties()->getHashRef();

    my $sp3 = 0;

    my @atoms = $mol->getAtoms()->getArray();

    foreach my $atom (@atoms) {

        # skip if not Carbon

        if ($atom->getType() != CARBON) {

            next;

        }

        if ($atom->getHybridization() == SP3_HYBRIDIZATION) {

            $sp3++;

        }

    }

}

```

```

    }

    $props->{"sp3"} = $sp3;

    return READYFORINPUTDATA;

}

sub onFinalize {

    my $context = shift;

}

"sp3/c_count >= 0.40"

```

The remaining structures were then subjected to Glide docking and MM-GBSA rescoring. The predicted K_i values were calculated by the QSAR correlations in **Figures 29** and **31**. After discussion with the collaborators, the structures listed in **Table 2** were synthesized and tested.

8.5.5. KCN1 application case – old scaffold

The FRESH workflow for the p300/HIF-1 α antagonist project on the old KCN1 scaffold was constructed in a fashion similar to the case studies. The version of Pipeline Pilot used in this project was 6.0 (student version under academic license), and the Schrodinger Maestro Package was the 2009 release. The QSAR in this case used MM-GBSA scores derived from the two-site binding model discussed in Section 5.3.

As stated in Section 5.4.1, the R_2 group was fully explored (building block is primary amine) while the R_3 group was only chosen from those in **Table 4**. Like the HDAC1 case, the R_2 group was queried in the commercial database (ChemNavigator, identical to the SNRI project) for fragments while the R_3 fragments were manually prepared in

Chewdraw. Unfavorable fragments for the R₂ group were excluded, and the MW cutoff was adjusted to 150 to ensure that the MW of the entire structure stays within 500 amu.

After enumeration of the target molecular structure, the following physical/ADMET property selection criteria were applied: “Molecular_Weight <= 500 AND (N_Count + O_Count) <= 10 AND Num_H_Donors <= 5 AND AlogP <= 5 AND Molecular_PolarSurfacearea <= 90 And Molecular_solubility >= -6.0;” and “i_qp_nummetab <= 6 and r_qp_QPPMDCK >= 500 AND r_qp_qpPCaco >= 500 AND r_qp_qplogbb >= -0.5;”. The ligands were then prepared by LigPrep, docked into both predicted binding sites by Glide and rescored by MM-GBSA. The best MM-GBSA score for each ligand was used for selection by the following PilotScript: “r_esp_Prime_MMGBSA_DG_bind <= -26.5;” for Site 1 and “r_esp_Prime_MMGBSA_DG_bind <= -27.5;” for Site 2. Only six structures (**Table 5**) survived the final list.

8.5.6. KCN1 application case – new scaffold

The FRESH workflow for the p300/HIF-1 α antagonist project on the new diaryl alcohol scaffold was constructed in a fashion similar to the case studies. The version of Pipeline Pilot used in this project was 7.0, and the Schrodinger Maestro Package was the 2012 release. The QSAR was constructed based on ~50 structures. They are listed in Appendix II. The training set structures were submitted to the “learn R PLS model” components for a partial least squares (PLS) analysis by using log IC₅₀ as the property to be learned and ECFP₆ as the molecular descriptor. This produced the QSAR line in **Figure 50**. Like the Bayesian model, a new Pipeline Pilot component based on the QSAR correlation was generated for further calculation.

The building blocks for this case were phenyl bromides and benzaldehydes, which were used to query the ChemNavigator commercial library and obtain fragments. Unfavorable fragments were removed and the rest were covalently attached to form the diphenylalcohol structure. The molecular structures were subjected to physical/ADMET criteria similar to those in Section 8.1.2 with an additional $\log BB \geq 0$ ($r_{qp_qplogBB} \geq 0$;) requirement. The predicted IC_{50} values were calculated using the newly-generated PLS components based on the QSAR line in **Figure 50** with 2 μM as the cutoff. After discussion with GSU collaborators, the structures listed in **Table 6** were further pursued for synthesis and testing.

8.6. Java Programming Language Implementation for NAMFIS

This section of the dissertation reveals some Java implantation details for the NAMFIS software program. The entire length of the Java code for this version is over 2,500 lines. However, most of it (GUI construction, File parsing, data collection etc.) is within the normal scope of any person with basic Java knowledge. Therefore, this section is focused on the newly implemented features of NAMFIS. The user input validation section only provides a description of the algorithm (coding requires only basic level of Java), while the minimization part requires the entire Java code.

8.6.1. Perform user input validation

The validation of NAMFIS user input is initiated by checking the required input parameters. The HashSet which stores all the involved atom indexes cannot be empty. Meanwhile, the ArrayLists for experimental NOE inputs and J couplings cannot be

simultaneously empty. If either is detected, the user is warned. If the input conformation pool file is empty or in an illegal format, the program will also notify the user.

The next round of validation performs a check of all the input experimental parameters. Duplication and invalid inputs are captured at this stage. For atom indices, since the HashSet data structure has already prohibited duplicates, no further validation step is involved. For the NOE pairs, no equal atom index pair or non-positive distance can be present, and all the atom indices should be included in the HashSet. Similar requirements also apply to J couplings and permutations. It should be noted that for comparison of permutation inputs to detect duplicates, overriding the inherited “equals” method is incorrect, as this violates the transitivity (if A equals to B and B equals to C, then A must equal to C) requirement. Thus, a separate method should be adopted by comparing each atom index in the two atom index lists of the two permutation inputs for equal elements.

With available molecular structures from a valid conformation pool file, the first conformer is selected as a reference. The program then performs a check for all the J coupling input. The J coupling constant requires atom input in the order H_1 , C_1 , C_2 and H_2 . Thus, a connectivity check is initiated to verify that H_1 is connected to C_1 , C_1 is connected to C_2 and C_2 is connected to H_2 . An atom type matching validation is also performed to guarantee that the input atom index for the J coupling matches the corresponding input J coupling type. For example, if input atoms are HCCH but the type is selected as HCNH, the input is rejected. After validation of all the experimental input, the entire conformation pool file is screened for conformers with different structures by comparing the calculated canonical_smiles string.

Finally, a procedure is performed to check obvious constraint violations for mathematically infeasible cases. The program examines each input NOE and J coupling to see if the calculated values are all greater/smaller for all conformers. The NAMFIS input must pass all the validation steps before conformer pool minimization.

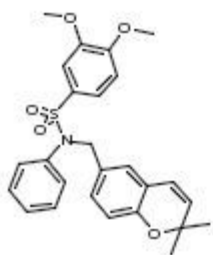
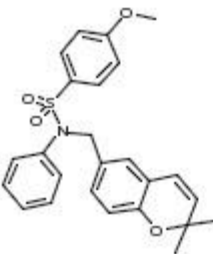
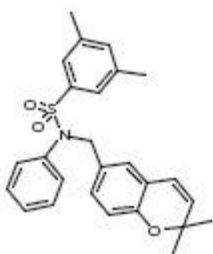
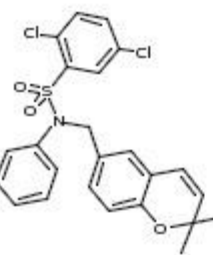
8.6.2. Perform constrained geometry and population minimization

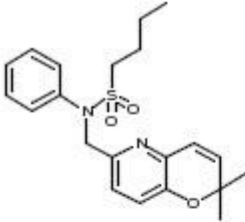
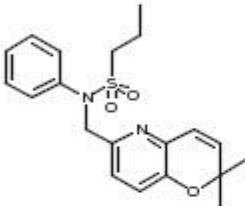
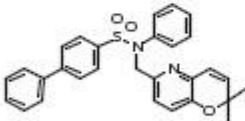
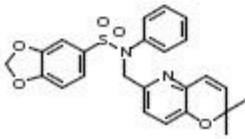
This step is accomplished by the “NAMFISOptimizer” class which extends the “Ipop” class as required. In this step, the input and calculated NOE distances and J couplings are placed in the same array to facilitate processing. The `calculatedResult[i][j]` term represents the j^{th} calculated value for the i^{th} conformer. The `experimentalResult[i]`, `experimentalError[i]` and `weight[i]` stands for the i^{th} input experimental value, experimental error and the weight factor (default value 1.0).

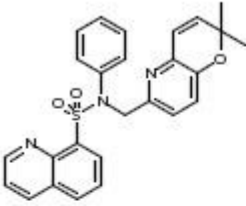
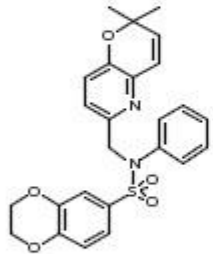
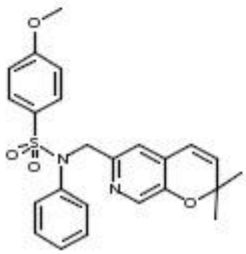
Several integer variables need to be initialized before the minimization. The variable “n” represents the number of conformers in the conformation pull file, while “m” equals the number of constraints (the number of J couplings and NOEs plus 1 for the mole fraction constraint). The variable “nele_jac” stores the number of non-zero terms in the Jacobian matrix of the constraints, which equals $m*n$. The variable “nele_hess” holds the number of non-zero terms in the lower triangular part of the Hessian Matrix of the Lagrangian function, which equals $n * (n+1) / 2$. The detailed Java implementation code for this class is demonstrated in Appendix III.

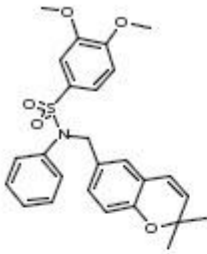
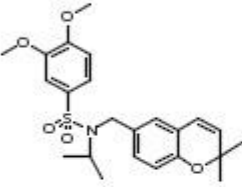
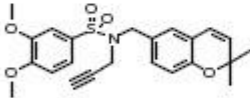
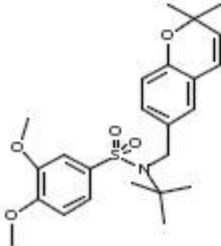
Appendix I: A series of KCN1 analog with experimental IC₅₀ values and predicted MM-GBSA values

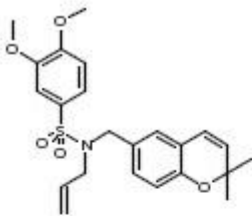
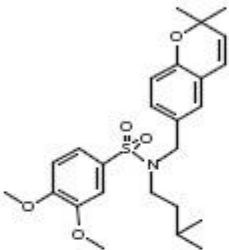
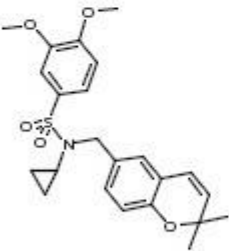
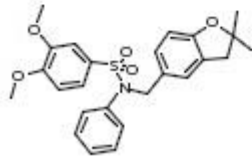
Unit for MMGBSA: kcal/mol, IC₅₀: μ M KCN1 is Compound 1.

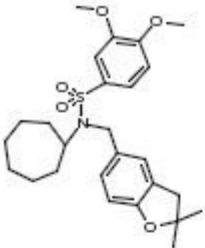
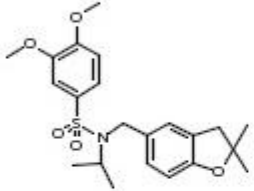
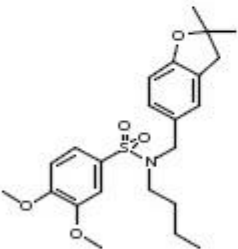
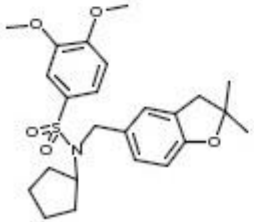
Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	1	-27.5	-28.5	0.7	0.7	I
	510	-23.4	-29.7	0.6	1.2	I
	511	-22.3	-27.3	0.5	1.2	I
	512	-24.9	-28.6	2.1	1.7	I

Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	2614	-24.2	-26.5	5	0.54	I
	2615	-20.6	-24.1	6.4	0.54	I
	2617	-26.1	-24.7	3.4	0.54	I
	2618	-25.8	-25.6	6.5	0.54	I

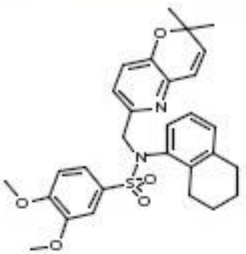
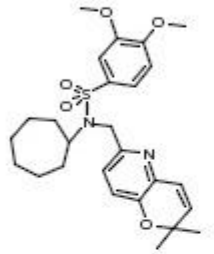
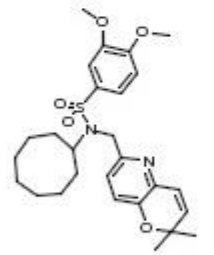
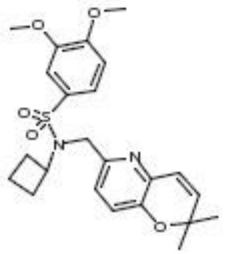
Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	2619	-18.8	-23.7	0.9	0.3	I
	2620	-19.7	-24.1	0.9	0.3	I
	3601	-20.5	-25.9	1.4	0.52	I

Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	1	-27.5	-28.5	0.7	0.7	II
	501	-20.0	-24.5	3.1	0.53	II
	502	-21.4	-24.0	1.3	0.53	II
	504	-20.0	-27.2	3.5	1.9	II

Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	505	-23.3	-28.3	3.4	1.9	II
	506	-20.3	-27.2	1.6	0.53	II
	508	-20.6	-26.6	1.5	0.53	II
	1601	-28.0	-29.1	0.5	1.2	II

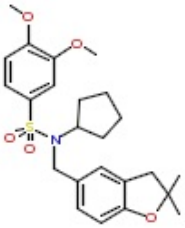
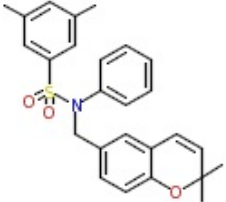
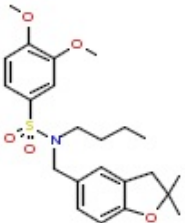
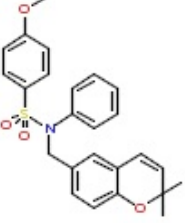
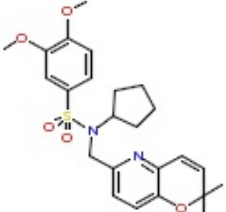
Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	1602	-23.6	-32.0	9.1	0.3	II
	1603	-21.5	-24.9	1.5	0.3	II
	1604	-19.5	-27.3	0.6	0.3	II
	1606	-21.3	-29.7	0.4	0.52	II

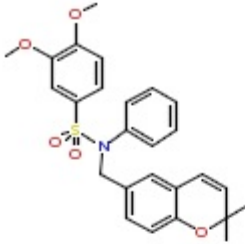
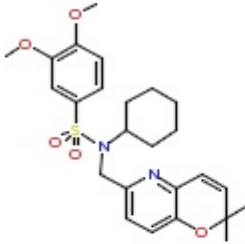
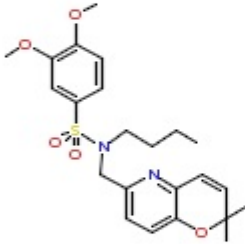
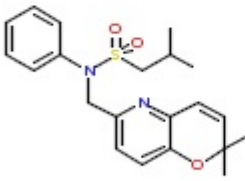
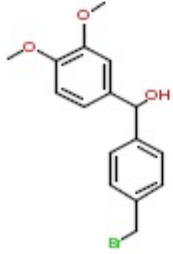
Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	2601	-25.7	-29.2	1.3	1.4	II
	2602	-22.7	-26.6	0.9	1.4	II
	2604	-22.5	-33.2	0.6	1.4	II
	2605	-23.0	-29.8	0.8	1.35	II

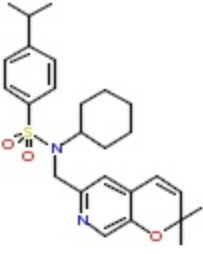
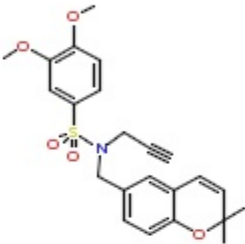
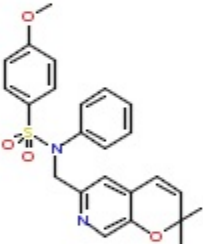
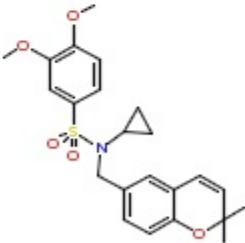
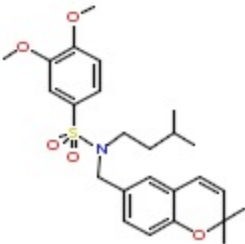
Molecule	Name	MMGBSA_Site1	MMGBSA_Site2	IC50	IC50KCN1	Group
	2606	-23.2	-28.5	6.2	0.7	II
	2607	-22.6	-32.1	6.6	0.7	II
	2608	-21.1	-31.8	0.7	0.4	II
	2609	-26.5	-29.7	0.25	0.59	II

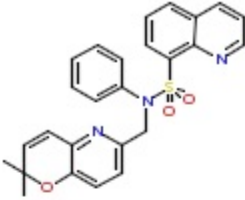
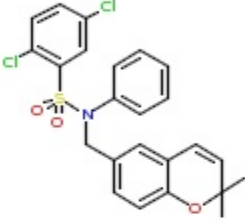
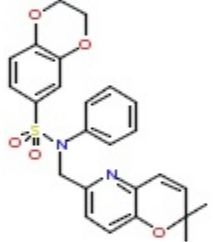
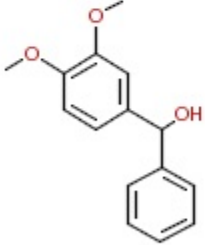
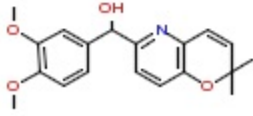
Appendix II: QSAR training and test sets for p300/HIF-1 α antagonists

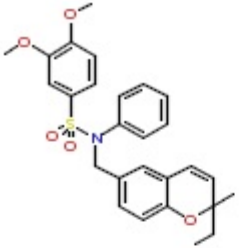
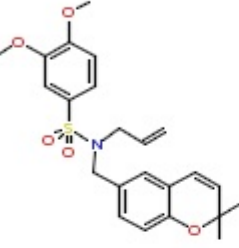
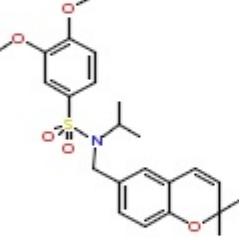
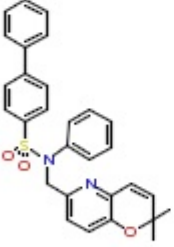
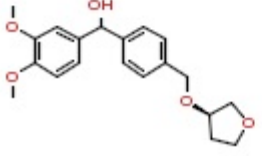
Unit for IC₅₀_p300: μ M.

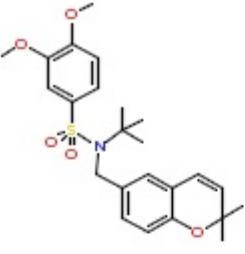
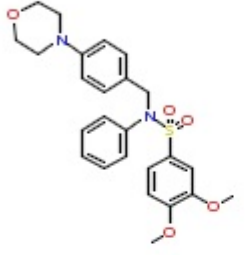
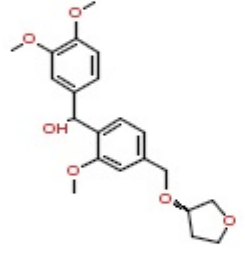
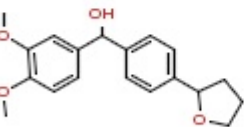
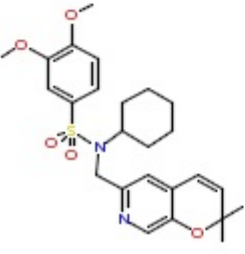
Molecule	Name	IC50_p300	Set
	1606	0.4	Training
	511	0.5	Training
	1604	0.6	Training
	510	0.6	Training
	2604	0.65	Training

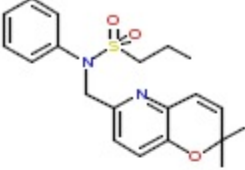
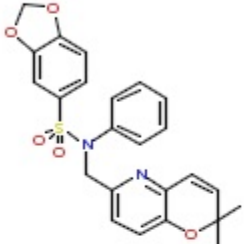
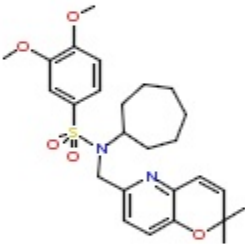
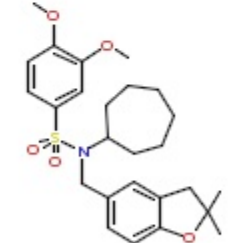
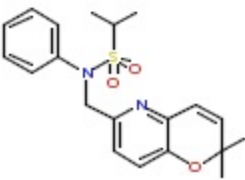
Molecule	Name	IC50_p300	Set
	1	0.7	Training
	2605	0.8	Training
	2602	0.9	Training
	2616	0.9	Training
	97	1	Training

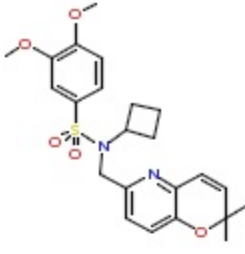
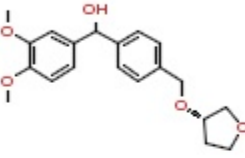
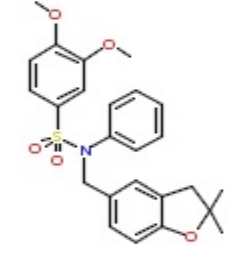
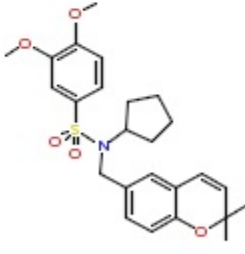
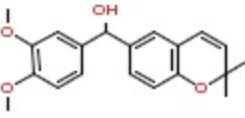
Molecule	Name	IC50_p300	Set
	3603	1.1	Training
	502	1.3	Training
	3601	1.4	Training
	508	1.52	Training
	506	1.6	Training

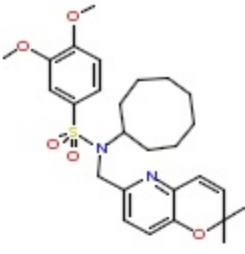
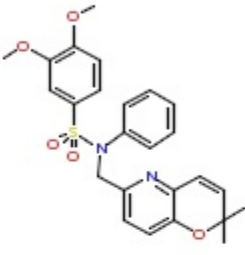
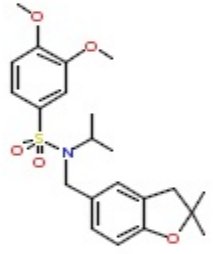
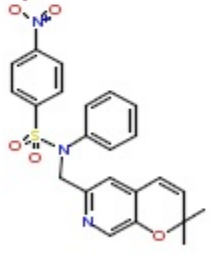
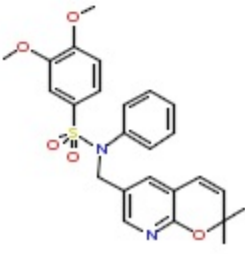
Molecule	Name	IC50_p300	Set
	2619	1.88	Training
	512	2.1	Training
	2620	2.1	Training
	95	2.2	Training
	76	2.25	Training

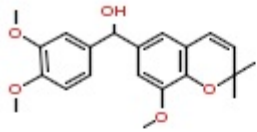
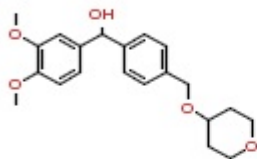
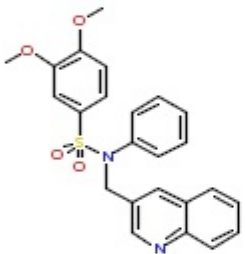
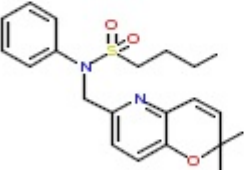
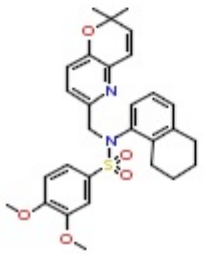
Molecule	Name	IC50_p300	Set
	1001	3.1	Training
	505	3.1	Training
	501	3.12	Training
	2617	3.4	Training
	100	3.4	Training

Molecule	Name	IC50_p300	Set
	504	3.45	Training
	108	3.8	Training
	128	4.2	Training
	127	5	Training
	3604	5.8	Training

Molecule	Name	IC50_p300	Set
	2615	6.4	Training
	2618	6.5	Training
	2607	6.6	Training
	1602	9.05	Training
	2612	13.4	Training

Molecule	Name	IC50_p300	Set
	2609	0.25	Test
	84	0.3	Test
	1601	0.5	Test
	507	0.5	Test
	71	0.6	Test

Molecule	Name	IC50_p300	Set
	2608	0.7	Test
	2601	1.25	Test
	1603	1.52	Test
	3602	1.8	Test
	2001	2.5	Test

Molecule	Name	IC50_p300	Set
	102	3.1	Test
	101	3.4	Test
	1301	3.52	Test
	2614	5	Test
	2606	6.2	Test

Appendix III: The Java implementation of the “NAMFISOptimizer” class

```
package namfis.processing.dataprocess;

import java.util.ArrayList;

import org.coinor.Ipopt;

import namfis.commonclassandmethod.CommonConstant;

import namfis.processing.NAMFISAllData;

public class NAMFISOptimizer extends Ipopt implements CommonConstant {

    /** All the NAMFIS input data */

    private NAMFISAllData namfisAllData;

    /** Number of variables */

    private int n;

    /** Number of Constrains */

    private int m;

    /** Number of Non Zero in the Jacobian Matrix of the constraints */

    private int nele_jac;

    /**
```

```
* Number of Non Zero in the Hessian of the Lagrangian (lower or upper
* triangular part only)
*/
private int nele_hess;

/** Calculated NOE Distances and Coupling Constants for each Conformer */
private double[][] calculatedResult;

/** Experimental NOE Distances and Coupling Constants */
private double[] experimentalResult;

/** Experimental NOE Distance Errors and Coupling Constant Errors */
private double[] experimentalError;

/** Weight Factor */
private double[] weight;

/**
 * Initialize the bounds and create the native Ipopt problem.
 */
public NAMFISOptimizer(NAMFISAllData namfisAllData) {

    this.namfisAllData = namfisAllData;
```



```
boolean enforceBond = false;

boolean isRemove = false;

ArrayList<Integer> outOfBondList = null;

if (enforceBond) {
    isRemove = false;

    outOfBondList = new DetectOutOfErrorBoundValues(
        namfisAllData, isRemove).getOutOfBondValueIndexList();
}

this.n = namfisAllData.getNAMFISConformersList().size();

/*
 * the "+1" for this.m represents one additional constraint :
 *  $x_1+x_2+\dots+x_n = 1.0$  (sum of mole fraction is 1)
 */

this.m = namfisAllData.getInputDataList().size() + 1;

this.nele_jac = this.n * this.m;

this.nele_hess = this.n * (this.n + 1) / 2;

this.calculatedResult = new double[m - 1][n];

this.experimentalResult = new double[m - 1];

this.experimentalError = new double[m - 1];

this.weight = new double[m - 1];
```

```
obtainCalculatedAndExperimentalMatrix();

double x_L[] = new double[n];
double x_U[] = new double[n];
for (int i = 0; i <= n - 1; i++) {
    x_L[i] = 0.0;
    x_U[i] = 1.0;
}

/* set the values of the constraint bounds */
double g_L[] = new double[m];
double g_U[] = new double[m];

/*
 * All the calculated values should fall within a error bar window
 * compared to the experimental ones
 */

for (int i = 0; i <= m - 1 - 1; i++) {
    if (enforceBond) {
        g_L[i] = experimentalResult[i] - experimentalError[i];
        g_U[i] = experimentalResult[i] + experimentalError[i];
    } else {
```

```
g_L[i] = Double.NEGATIVE_INFINITY;
g_U[i] = Double.POSITIVE_INFINITY;
}
}

/*
 * For the calculated values that are out of error bar window, if they
 * are chosen to be kept, their constraint conditions are removed
 */
if (enforceBond && !isRemove && outOfBondList != null
    && !outOfBondList.isEmpty()) {
    for (int outOfBondIndex : outOfBondList) {
        g_L[outOfBondIndex] = Double.NEGATIVE_INFINITY;
        g_U[outOfBondIndex] = Double.POSITIVE_INFINITY;
    }
}

/* All mole fractions should add to 1 */
g_L[m - 1] = 1.0;
g_U[m - 1] = 1.0;

/* Index style for the irow/jcol elements */
int index_style = Ipopt.C_STYLE;
```

```
/* create the IpoptProblem */  
  
create(n, x_L, x_U, m, g_L, g_U, nele_jac, nele_hess, index_style);  
  
}  
  
/**  
  
* Provide the calculated results (NOE distance and Coupling constants,  
* derived from the input conformation file, experimental results and  
* experimental errors. These arrays will be used in the following methods  
* in IPOPT  
*  
* @return void  
*/  
  
private void obtainCalculatedAndExperimentalMatrix() {  
  
for (int i = 0; i <= m - 1 - 1; i++) {  
for (int j = 0; j <= n - 1; j++) {  
calculatedResult[i][j] = namfisAllData  
    .getNAMFISConformersList().get(j)  
    .getCalculatedValuesList().get(i);  
}  
  
experimentalResult[i] = namfisAllData.getInputDataList().get(i)  
    .getExperimentalValue();
```

```
experimentalError[i] = namfisAllData.getInputDataList().get(i)
    .getErrorValue();
weight[i] = namfisAllData.getInputDataList().get(i)
    .getWeightFactor();
}
}

public double[][] getCalculatedResult() {
    return this.calculatedResult;
}

public double[] getExperimentalResult() {
    return this.experimentalResult;
}

public double[] getExperimentalError() {
    return this.experimentalError;
}

/** Initialize the mole fraction array x[] */
public double[] getInitialGuess() {
    double x[] = new double[n];
    x[0] = 1.0;
    for (int i = 1; i <= n - 1; i++) {
```

```

    x[i] = 0.0;
}
return x;
}

```

@Override

```

protected boolean eval_f(int n, double[] x, boolean new_x,
    double[] obj_value) {
    assert n == this.n;
    double sum = 0.0;
    for (int i = 0; i <= m - 1 - 1; i++) {

        double resultForOneSum = 0.0;
        for (int j = 0; j <= n - 1; j++) {
            ;
            resultForOneSum = resultForOneSum + x[j]
                * calculatedResult[i][j];
        }
        resultForOneSum = resultForOneSum - experimentalResult[i];
        resultForOneSum = resultForOneSum / experimentalError[i];
        resultForOneSum = resultForOneSum * weight[i];
        resultForOneSum = resultForOneSum * resultForOneSum;
    }
}

```

```
    sum = sum + resultForOneSum;
}
obj_value[0] = sum;
return true;
}
```

@Override

```
protected boolean eval_grad_f(int n, double[] x, boolean new_x,
    double[] grad_f) {
    assert n == this.n;

    for (int i = 0; i <= n - 1; i++) {

        double sum = 0.0;

        for (int j = 0; j <= m - 1 - 1; j++) {

            double sumK = 0.0;

            for (int k = 0; k <= n - 1; k++) {

                sumK = sumK + x[k] * calculatedResult[j][k];

            }

            sumK = sumK - experimentalResult[j];

            sumK = sumK / experimentalError[j];

            sumK = sumK * weight[j];
```

```
sumK = 2.0 * sumK * calculatedResult[j][i];

sum = sum + sumK;
}

grad_f[i] = sum;

}

return true;
}

@Override

protected boolean eval_g(int n, double[] x, boolean new_x, int m, double[] g) {
    assert n == this.n;
    assert m == this.m;

    /*
     * All the calculated values should fall within a error bar window
     * compared to the experimental ones
     */
    for (int i = 0; i <= m - 1 - 1; i++) {
        double sum = 0.0;
        for (int j = 0; j <= n - 1; j++) {
```



```

    sum = sum + x[j] * calculatedResult[i][j];
}
g[i] = sum;
}
/* All mole fractions should add to 1 */
double sum = 0.0;
for (int i = 0; i <= n - 1; i++) {
    sum = sum + x[i];
}
g[m - 1] = sum;

return true;
}

@Override
protected boolean eval_jac_g(int n, double[] x, boolean new_x, int m,
    int nele_jac, int[] iRow, int[] jCol, double[] values) {
    assert n == this.n;
    assert m == this.m;

    /*
     * iRow[index] and jCol[index] can be regarded as a index converter
     * function for 1D array index to 2D array index. The "index" term in

```

```

* "iRow[index] and jCol[index]" stands for the 1D array index, and
* iRow[index] stands for its corresponding 2D array row index
*/

if (values == null) {
    int index = 0;
    for (int i = 0; i <= m - 1; i++) {
        for (int j = 0; j <= n - 1; j++) {
            iRow[index] = i;
            jCol[index] = j;
            index++;
        }
    }
} else {
    /*
    * return the values of the jacobian of the constraints regarding
    * calculated values
    */
    for (int i = 0; i <= m - 1 - 1; i++) {
        for (int j = 0; j <= n - 1; j++) {
            values[i * n + j] = calculatedResult[i][j];
        }
    }
}

```

```
/*  
 * return the values of the jacobian of the constraint regarding the  
 * sum of mole fraction  
 */  
for (int j = 0; j <= n - 1; j++) {  
    values[(m - 1) * n + j] = 1.0;  
}  
  
}  
  
return true;  
}
```

@Override

```
protected boolean eval_h(int n, double[] x, boolean new_x,  
    double obj_factor, int m, double[] lambda, boolean new_lambda,  
    int nele_hess, int[] iRow, int[] jCol, double[] values) {  
  
    if (values == null) {  
        int index = 0;  
        for (int i = 0; i <= n - 1; i++) {  
            for (int j = 0; j <= i; j++) {
```

```
iRow[index] = i;
jCol[index] = j;
index++;
}
}

} else {
for (int i = 0; i <= n - 1; i++) {
for (int j = 0; j <= i; j++) {

double sum = 0.0;
for (int k = 0; k <= m - 1 - 1; k++) {
sum = sum + 2.0 * calculatedResult[k][i]
* calculatedResult[k][j] / experimentalError[k]
* weight[k];
}
/* Note: should not be values[i * n + j] */
values[(1 + i) * i / 2 + j] = obj_factor * sum;
}
}
}
return true;
}
```

```
}  
  
@Override  
protected boolean eval_grad_f(int n, double[] x, boolean new_x,  
    double[] grad_f) {  
    assert n == this.n;  
  
    for (int i = 0; i <= n - 1; i++) {  
  
        double sum = 0.0;  
  
        for (int j = 0; j <= m - 1 - 1; j++) {  
  
            double sumK = 0.0;  
  
            for (int k = 0; k <= n - 1; k++) {  
  
                sumK = sumK + x[k] * calculatedResult[j][k];  
  
            }  
  
            sumK = sumK - experimentalResult[j];  
  
            sumK = sumK / experimentalError[j];  
  
            sumK = sumK * weight[j];  
  
            sumK = 2.0 * sumK * calculatedResult[j][i];  
  
            sum = sum + sumK;  
  
        }  
    }  
}
```

```
grad_f[i] = sum;

}

return true;
}

@Override

protected boolean eval_g(int n, double[] x, boolean new_x, int m, double[] g) {

    assert n == this.n;

    assert m == this.m;

    /*

    * All the calculated values should fall within a error bar window

    * compared to the experimental ones

    */

    for (int i = 0; i <= m - 1 - 1; i++) {

        double sum = 0.0;

        for (int j = 0; j <= n - 1; j++) {

            sum = sum + x[j] * calculatedResult[i][j];

        }

        g[i] = sum;

    }

}
```

```
/* All mole fractions should add to 1 */  
  
double sum = 0.0;  
  
for (int i = 0; i <= n - 1; i++) {  
  
    sum = sum + x[i];  
  
}  
  
g[m - 1] = sum;  
  
  
return true;  
  
}  
  
@Override  
  
protected boolean eval_jac_g(int n, double[] x, boolean new_x, int m,  
    int nele_jac, int[] iRow, int[] jCol, double[] values) {  
  
    assert n == this.n;  
  
    assert m == this.m;  
  
  
/*  
  
    * iRow[index] and jCol[index] can be regarded as a index converter  
  
    * function for 1D array index to 2D array index. The "index" term in  
  
    * "iRow[index] and jCol[index]" stands for the 1D array index, and  
  
    * iRow[index] stands for its corresponding 2D array row index  
  
*/
```

```
if (values == null) {  
    int index = 0;  
    for (int i = 0; i <= m - 1; i++) {  
        for (int j = 0; j <= n - 1; j++) {  
            iRow[index] = i;  
            jCol[index] = j;  
            index++;  
        }  
    }  
} else {  
    /*  
    * return the values of the jacobian of the constraints regarding  
    * calculated values  
    */  
    for (int i = 0; i <= m - 1 - 1; i++) {  
        for (int j = 0; j <= n - 1; j++) {  
            values[i * n + j] = calculatedResult[i][j];  
        }  
    }  
    /*  
    * return the values of the jacobian of the constraint regarding the  
    * sum of mole fraction
```



```
*/  
  
for (int j = 0; j <= n - 1; j++) {  
    values[(m - 1) * n + j] = 1.0;  
}  
  
}  
  
return true;  
}  
  
@Override  
protected boolean eval_h(int n, double[] x, boolean new_x,  
    double obj_factor, int m, double[] lambda, boolean new_lambda,  
    int nele_hess, int[] iRow, int[] jCol, double[] values) {  
  
    if (values == null) {  
        int index = 0;  
        for (int i = 0; i <= n - 1; i++) {  
            for (int j = 0; j <= i; j++) {  
                iRow[index] = i;  
                jCol[index] = j;  
                index++;  
            }  
        }  
    }  
}
```

```
}

} else {

for (int i = 0; i <= n - 1; i++) {

for (int j = 0; j <= i; j++) { // j<= i or j<= n-1

double sum = 0.0;

for (int k = 0; k <= m - 1 - 1; k++) {

sum = sum + 2.0 * calculatedResult[k][i]

* calculatedResult[k][j] / experimentalError[k]

* weight[k];

}

/* Note: should not be values[i * n + j] */

values[(1 + i) * i / 2 + j] = obj_factor * sum;

}

}

}

return true;

}

}
```

References

The research projects in this dissertation are the continuation of my master thesis in 2012. The background parts and some initial works are either rephrased or directly from that thesis.

- (1) Bohacek R. S.; McMartin C.; Guida W. C. The art and practice of structure-based drug design *Med. Res. Rev.* **1996**, 16, 3–50
- (2) Lipinski C. A.; Lombardo F.; Dominy B. W.; Feeney P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* **1997**, 23, 3–25
- (3) Jorgensen W. L. Efficient drug lead discovery and optimization *Acc. Chem. Res.* **2009**, 42, 724-733
- (4) Morelli X.; Bourgeois R.; Roche P. Chemical and structural lessons from recent successes in protein-protein interaction inhibition (2P2I) *Curr. Opin. Chem. Biol.* **2011**, 15, 475-481
- (5) Leeson P. D.; Springthorpe B. The influence of drug-like concepts on decision-making in medicinal chemistry *Nature Reviews Drug Discovery* **2007**, 6, 881-890
- (6) Morphy R. The influence of target family and functional activity on the physicochemical properties of pre-clinical compounds *J. Med. Chem* **2006**, 49, 2969-2978
- (7) Oprea T. I. Current trends in lead discovery: are we looking for the appropriate properties? *J. Comp Aided Mol. Design* **2002**, 16, 325–334
- (8) Levin V. A. Relationship of octanol/water partition coefficient and molecular weight to rat brain capillary permeability *J. Med. Chem* **1980**, 23, 682–684
- (9) Atkinson F.; Cole S.; Green C.; van de Waterbeemd H. Lipophilicity and other parameters affecting brain penetration *Curr. Med. Chem-Central Nervous System Agents* **2002**, 2, 229-240
- (10) Kelder J.; Grootenhuis P. D. J.; Bayada D. M.; Delbressine L. P. C.; Ploemen J. P. Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs *Pharm. Res.* **1999**, 16, 1514-1519.
- (11) Leeson P. D.; Davis A. M. Time-related differences in the physical property profiles of oral drugs *J. Med. Chem* **2004**, 47, 6338–6348
- (12) Hansch C.; Leo A. J. *Substituent constant for correlation analysis in chemistry and biology*. New York: Wiley, **1979**
- (13) Perola, E.; Walters W. P.; Charifson P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *PROTEINS* **2004**, 56, 235-249
- (14) Friesner R. A.; Banks J. L.; Murphy R. B.; Halgren T. A.; Klicic J. J.; Mainz D. T.; Repasky M. P.; Knoll E. H.; Shelley M.; Perry J. K.; Shaw D. E.; Francis P.; Shenkin P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.* **2004**, 47, 1739-1749
- (15) Halgren T. A.; Murphy R. B.; Friesner R. A.; Beard H. S.; Frye L. L.; Pollard W. T.; Banks J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. enrichment factors in database screening. *J. Med. Chem.* **2004**, 47, 1750-1759
- (16) Zhou Z.; Felts A. K.; Friesner R. A.; Levy R. M. Comparative performance of several flexible docking programs and scoring functions: enrichment studies for a diverse set of pharmaceutically relevant targets. *J. Chem. Inf. Model.* **2007**, 47, 1599-1608
- (17) Guimaraes C. R. W.; Cardozo M. MM-GB/SA Rescoring of Docking Poses in Structure-Based Lead Optimization *J. Chem. Inf. Model.* **2008**, 48, 958–970
- (18) Kubinyi, H. Opinion: Drug Research: Myths, Hype and Reality. *Nat. Rev. Drug Discovery* **2003**, 2, 665–668.
- (19) Kola I.; Landis J. Can the pharmaceutical industry reduce attrition rates? *Nat. Rev. Drug Discovery* **2003**, 3, 711-715.
- (20) QikProp, version 3.2, Schrödinger, LLC, New York, NY, **2009**
- (21) <http://accelrys.com/products/pipeline-pilot/component-collections/adme-tox.html>
- (22) http://www.acdlabs.com/products/pc_admet/admetox.php

- (23) Herrick T. M.; Million R. P. From the analyst's couch: tapping the potential of fixed-dose combinations *Nat. Rev. Drug. Discov.* **2007**, *6*, 513–514
- (24) Sterling J.; Herzig Y.; Goren T.; Finkelstein N.; Lerner D.; Goldenberg W.; Miskolczi I.; Molnar S.; Rantal F.; Tamas T.; Toth G.; Zagyva A.; Zekany A.; Finberg J.; Lavian G.; Gross A.; Friedman R.; Razin M.; Huang W.; Kraus B.; Chorev M.; Youdim M. B.; Weinstock M. Novel dual inhibitors of AChE and MAO derived from hydroxy aminoindan and phenethylamine as potential treatment for Alzheimer's disease *J. Med. Chem.* **2002**, *45*, 5260–5270
- (25) Jorgensen W. L. Challenges for Academic Drug Discovery *Angew. Chem. Int. Ed.* **2012**, *51*, 11680–11684
- (26) Rishton G. M. Reactive compounds and in vitro false positives in HTS *Drug Discov. Today* **1997**, *2*, 382–384
- (27) Congreve M.; Carr R.; Murray C.; Jhoti H. A 'rule of three' for fragment-based lead discovery *Drug Discov. Today* **2003**, *8*, 876–877
- (28) Besnard J.; Ruda G. F.; Setola V.; Abecassis K.; Rodriguiz R. M.; Huang X. P.; Norval S.; Sassano M. F.; Shin A. I.; Webster L. A.; Simeons F. R.; Stojanovski L.; Prat A.; Seidah N. G.; Constam D. B.; Bickerton G. R.; Read K. D.; Wetsel W. C.; Gilbert I. H.; Roth B. L.; Hopkins A. L. Automated design of ligands to polypharmacological profiles *Nature* **2012**, *492*, 215–220
- (29) Rogers D.; Brown R. D.; Hahn M. Using extended-connectivity fingerprints with Laplacian-modified Bayesian analysis in high-throughput screening follow-up *J. Biomol. Screen.* **2005**, *10*, 682–686
- (30) Gaulton A.; Bellis L. J.; Bento A. P.; Chambers J.; Davies M.; Hersey A.; Light Y.; McGlinchey S.; Michalovich D.; Al-Lazikani B.; Overington J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, 1100–1107
- (31) Irwin J. J.; Sterlin T.; Mysinger M. M.; Bolstad E. S.; Coleman R. G. ZINC: A Free Tool to Discover Chemistry for Biology *J. Chem. Inf. Model.*, **2012**, *52*, 1757–1768
- (32) Cain, R. J.; Ridley, A. J. Phosphoinositide 3-kinases in cell migration. *Biol. Cell.* **2009**, *101*, 13–29
- (33) Sawyer, C.; Sturge, J.; Bennett, D. C.; O'Hare, M. J.; Allen, W. E.; Bain, J.; Jones, G. E.; Vanhaesebroeck, B. Regulation of breast cancer cell chemotaxis by the phosphoinositide 3-kinase p110delta. *Cancer Res.* **2003**, *63*, 1667–1675
- (34) Price, J. T.; Tiganis, T.; Agarwal, A.; Djakiew, D.; Thompson, E. W. Epidermal growth factor promotes MDA-MB-231 breast cancer cell migration through a phosphatidylinositol 3'-kinase and phospholipase C-dependent mechanism. *Cancer Res.* **1999**, *59*, 5475–5478.
- (35) Akinleye A.; Avvaru P.; Furqan M.; Song Y.; Liu D. Phosphatidylinositol 3-kinase (PI3K) inhibitors as cancer therapeutics *J. Hematol. Oncol.* **2013**, *6*, 88–104
- (36) Kim, O.; Jeong, Y.; Lee, H.; Hong, S. S.; Hong, S. Design and synthesis of imidazopyridine analogues as inhibitors of phosphoinositide 3-kinase signaling and angiogenesis. *J. Med. Chem.* **2011**, *54*, 2455–2466.
- (37) Walker E. H.; Pacold M. E.; Perisic O.; Stephens L.; Hawkins P. T.; Wymann M. P.; Williams R. L. Structural determinants of phosphoinositide 3-kinase inhibition by wortmannin, LY294002, quercetin, myricetin, and staurosporine *Mol. Cell.* **2000**, *6*, 909–919
- (38) Supuran C. T. Carbonic anhydrases: novel therapeutic applications for inhibitors and activators *Nat. Rev. Drug Discov.* **2008**, *7*, 168–181
- (39) Pacchiano, F.; Aggarwal, M.; Avvaru, B. S.; Robbins, A. H.; Scozzafava, A.; McKenna, R.; Supuran, C. T. Selective hydrophobic pocket binding observed within the carbonic anhydrase II active site accommodate different 4-substituted-ureidobenzenesulfonamides and correlate to inhibitor potency. *Chem. Commun.* **2010**, *46*, 8371–8373.
- (40) Pacchiano, F.; Carta, F.; McDonald, P. C.; Lou, Y.; Vullo, D.; Scozzafava, A.; Dedhar, S.; Supuran, C. T. Ureido-substituted benzenesulfonamides potently inhibit carbonic anhydrase IX and show antimetastatic activity in a model of breast cancer metastasis. *J. Med. Chem.* **2011**, *54*, 1896–1902.
- (41) Behnke C.A.; Le Trong I.; Godden J.W.; Merritt E.A.; Teller D.C.; Bajorath J.; Stenkamp R.E. Atomic resolution studies of carbonic anhydrase II *Acta Crystallogr.* **2010**, *66*, 616–627
- (42) De Ruijter A. J.; Van Gennip A. H.; Caron H. N.; Kemp S.; Van Kuilenburg A. B. Histone deacetylases (HDACs): characterization of the classical HDAC family *Biochem J.* **2003**, *370*, 737–749

- (43) Glaser K. B.; Li, J.; Staver M. J.; Wei R.-Q.; Albert D. H.; Davidsen S. K. *Biochem. Biophys. Res. Commun.* Role of Class I and Class II histone deacetylases in carcinoma cells using siRNA **2003**, 310, 529-536
- (44) Richon V. M. Cancer biology: mechanism of antitumour action of vorinostat (suberoylanilide hydroxamic acid), a novel histone deacetylase inhibitor *Br J Cancer.* **2006**, 95(Suppl 1), S2-S6.
- (45) Wang H.; Lim Z. Y.; Zhou Y.; Ng M.; Lu T.; Lee K.; Sangthongpitag K.; Goh K. C.; Wang X.; Wu X.; Khng H. H.; Goh S. K.; Ong W. C.; Bonday Z.; Sun, E. T. Acylurea connected straight chain hydroxamates as novel histone deacetylase inhibitors: Synthesis, SAR, and in vivo antitumor activity. *Bioorg. Med. Chem. Lett.* **2010**, 20, 3314-3321.
- (46) Johannes C. B.; Le T. K.; Zhou X.; Johnston J. A.; Dworkin R. H. The prevalence of chronic pain in United States adults: results of an Internet-based survey *J. Pain.* **2010**, 11, 1230-1239
- (47) Stewart W. F.; Ricci J. A.; Chee E.; Morganstein D.; Lipton R. Lost productive time and cost due to common pain conditions in the US workforce. *J. Am. Med. Assoc.* **2003**, 290, 2443-2454.
- (48) Ritzwoller D. P.; Ellis J. L.; Korner E. J.; Hartsfield C. L.; Sadosky A. Comorbidities, healthcare service utilization and costs for patients identified with painful DPN in a managed-care setting *Curr. Med. Res. Opin.* **2009**, 25, 1319-1328
- (49) Bruehl S. An update on the pathophysiology of complex regional pain syndrome. *Anesthesiol* **2010**, 113, 713-725
- (50) Dr. Spandan Chennamadhavuni's thesis **2011**, Emory University
- (51) Lee Y. C.; Chen P. P. A review of SSRIs and SNRIs in neuropathic pain *Expert Opin Pharmacother.* **2010**, 11, 2813-2825
- (52) Mochizucki D. Serotonin and noradrenaline reuptake inhibitors in animal models of pain *Hum. Psychopharm. Clin.* **2004**, 19, S15-S19
- (53) Ikeda T.; Ishida Y.; Naono R.; Takeda R.; Abe H.; Nakamura T.; Nishimori T. Effects of intrathecal administration of newer antidepressants on mechanical allodynia in rat models of neuropathic pain *Neurosci. Res.* **2009**, 63, 42-46
- (54) Ravna A. W.; Sylte I.; Dahl S. G. Structure and localisation of drug binding sites on neurotransmitter transporters *J. Mol. Model.* **2009**, 15, 1155-1164
- (55) Brat D. J.; Kaur B.; Van Meir E. G.; Genetic modulation of hypoxia-induced gene expression and vascular proliferation: relevance to brain tumors *Front Biosci* **2002**, 8, d100-116
- (56) Lara, P. C.; Lloret, M.; Clavo, B.; Apolinario, R. M.; Henríquez-Hernández, L. A.; Bordón, E.; Fontes, F.; Rey, A. Severe hypoxia induces chemo-resistance in clinical cervical tumors through MVP over-expression *Radiat Oncol.* **2009**, 4:29
- (57) Semenza G. L. Hypoxia-inducible factors in physiology and medicine *Cell* **2012**, 148, 399-408
- (58) Reid-Mooring, S.; Jin, H.; Devi, N. S.; Jabbar, A. A.; Kaluz, S.; Liu, Y.; Van Meir E. G.; Wang, B. Design and synthesis of novel small-molecule inhibitors of the hypoxia inducible factor pathway *J. Med. Chem.* **2011**, 54, 8471-8489
- (59) Teague S. J.; Davis A. M.; Leeson P. D.; Oprea T. The design of leadlike combinatorial libraries *Angew. Chem. Int. Ed.* **1999**, 38, 3743-3748
- (60) Kuntz I. D.; Chen K.; Sharp K. A.; Kollman P. A. The maximal affinity of ligands *Proc. Natl. Acad. Sci. USA* **1999**, 96, 9997-10002
- (61) De Guzman, R. N.; Wojciak, J. M.; Martinez-Yamout, M. A.; Dyson, H. J.; Wright, P. E. CBP/p300 TAZ1 Domain Forms a Structured Scaffold for Ligand Binding *Biochemistry* **2005**, 44, 490-497
- (62) Shi Q.; Yin S.; Kaluz S.; Ni N.; Devi N. S.; Mun J.; Wang D.; Damera K.; Chen W.; Burroughs S.; Reid-Mooring S.; Goodman M. M.; Van Meir E. G.; Wang B.; Snyder J. P. Binding Model for the Interaction of Anti-cancer Arylsulfonamides with the p300 Transcription co-factor *submitted manuscript*
- (63) Freedman S. J.; Sun Z. Y.; Poy F.; Kung A. L.; Livingston D. M.; Wagner G.; Eck M. J. Structural basis for recruitment of CBP/p300 by hypoxia-inducible factor-1 alpha *Proc. Natl. Acad. Sci.* **2002**, 99, 5367-5372
- (64) Gu J.; Milligan J.; Huang L. E. Molecular Mechanism of Hypoxia-inducible Factor 1 α -p300 Interaction *J. Biol. Chem.* **2001**, 276, 3550-3554
- (65) Glide, version 5.5, Schrödinger, LLC, New York, NY, **2009**
- (66) Prime, version 2.1, Schrödinger, LLC, New York, NY, **2009**

- (67) Thomas S. L.; Zhong D.; Zhou W.; Malik S.; Liotta D.; Snyder J. P.; Hamel E.; Giannakakou P. EF24, a novel curcumin analog, disrupts the microtubule cytoskeleton and inhibits HIF-1 *Cell Cycle*. **2008**, 15, 24409-2417
- (68) Zhu S.; Moore T. W.; Lin X.; Morii N.; Mancini A.; Howard R. B.; Culver D.; Arrendale R. F.; Reddy P.; Evers T. J.; Zhang H.; Sica G.; Chen Z. G.; Sun A.; Fu H.; Khuri F. R.; Shin D. M.; Snyder J. P.; Shoji M. Synthetic curcumin analog EF31 inhibits the growth of head and neck squamous cell carcinoma xenografts. *Integr. Biol. (Camb)*. **2012**, 4, 633-640
- (69) Brown A.; Shi Q.; Moore T. W.; Yoon Y.; Prussia A.; Maddox C.; Liotta D. C.; Shim H.; Snyder J. P. Monocarbonyl curcumin analogues: heterocyclic pleiotropic kinase inhibitors that mediate anticancer properties. *J. Med. Chem.* **2013**, 56, 3456-3466
- (70) Rouse M. B.; Seefeld M. A.; Leber J. D.; McNulty K. C.; Sun L.; Miller W. H.; Zhang S.; Minthorn E. A.; Concha N. O.; Choudhry A. E.; Schaber M. D.; Heerding D. A. Aminofurazans as potent inhibitors of AKT kinase. *Bioorg. Med. Chem. Lett.* **2009**, 19, 1508-1511
- (71) Cicero D. O.; Barbato G.; Bazzo R. NMR Analysis of Molecular Flexibility in Solution: A New Method for the Study of Complex Distributions of Rapidly Exchanging Conformations. Application to a 13-Residue Peptide with an 8-Residue Loop *J. Am. Chem. Soc.* **1995**, 117, 1027-1033
- (72) Lakdawala A.; Wang M.; Nevins N.; Liotta D. C.; Rusinska-Roszak D.; Lozynski M.; Snyder J.P. Calculated conformer energies for organic molecules with multiple polar functionalities are method dependent: Taxol (case study) *BMC Chem Biol* **2001**, 1:2
- (73) Coxon B. Developments in the Karplus equation as they relate to the NMR coupling constants of carbohydrates. *Adv. Carbohydr. Chem. Biochem.* **2009**, 62, 17-82
- (74) JChem Base was used for structure searching and chemical database access and management, JChem 6.2.2, 2014, ChemAxon (<http://www.chemaxon.com>)
- (75) Wächter A.; Biegler L. T. On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming, *Math. Program.* **2006**, 106, 25-57
- (76) <http://www.uniprot.org/uniprot/P42336>