VIRTUAL DRUG SCREENING OF THE HIV-1 CAPSID PROTEIN (P24)

By

JAY C. HOUSTON

(Under the Guidance of Dr. Cory Momany)

## Abstract

Despite the worldwide attention garnered by AIDS, the HIV therapies of today do not extend to a wide variety of viral targets. The HIV-1 capsid structure bears significance as a drug target via its protective properties for the overall maturation of the virion. Individual capsid dimer units combine to form an electron dense, protective core around the RNA strands via proteolytic cleavage of the capsid protein (p24) from the Gag polyprotein complex. A hydrophobic "pocket" is created prior to final beta-hairpin/helix formation of the initial 13 N-terminal residues of the capsid protein. Amino acid residue sequence alignment studies of the protein (p24) show the N-terminal Proline and residue 51 (Aspartate) to be highly conserved. These two residues form a salt-bridge that stabilizes the final formation of the protein. The UNIX-based program DOCK 4.0 was used to screen more than 400,000 compounds in search of ligands and pharmacophores capable of interacting with key N-terminal pocket residues. Numerous lead compounds have been analyzed which have notable potential for inhibiting capsid formation and subsequent retardation of HIV maturation.

INDEX WORDS: Virtual Drug Screening; HIV-1; Capsid assembly; anti-viral inhibitors; UNIX-based Docking; DOCK 4.0; Molecular Genetics.

# VIRTUAL DRUG SCREENING OF THE HIV-1 CAPSID PROTEIN (P24)

By

JAY C. HOUSTON

B.S. Biology, Cleveland State University, 1996

A Dissertation submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

Doctor of Philosophy

ATHENS, GEORGIA

2004

# VIRTUAL DRUG SCREENING OF THE HIV-1 CAPSID PROTEIN (P24)

By

JAY C. HOUSTON

Approved by:

Major Professor:

Dr. Cory Momany

Committee:

Dr. Will Taylor
Dr. Russell Malmberg
Dr. Stuart Feldman
Dr. Anthony Capomacchia

Electronic Version Approved by:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2004

## DEDICATION

To my Parents: Ann & Jerry, with all my love

I told you *I'd be right back* mom.

# ACKNOWLEDGEMENTS

First and foremost, I want to thank my major professor, Dr. Cory Momany for his knowledge, supreme scientific intellect and availability. I wish you and your family (Dale & Dr. Michelle Momany in Botany) continued success both on and off of the bench-top. I also want give thanks to my committee members. Dr. Russell Malmberg, your advice to me, and the Bioinformatics emphasis of your Genomics classes were very much appreciated. Dr. E. Will Taylor, your support of me personally as well as your white-board, CCMSD break-room RNA lectures involving all of us in Computational Chemistry were excellent. Dr. Stuart Feldman, active involvement and support of me personally and SNPhA overall, including the many gatherings at your house were all critical, given the known history of this College and this University. Dr. Anthony Capomacchia is mostly responsible for my presence in the department of PBS. From recruiting me here, to listening to my issues, to always being extremely supportive, Dr. "C" is one of the great *people* to come into my life. I thank you.

With the nature of my research, I also need to thank the esteemed post-doc Sandra Haddad from Microbiology. You not only sat with me in your lab and broke down some of the subtle minutiae of restriction enzyme molecular genetics, you are a *real person* as well as a scientist, imagine that.

Inside the Momany lab, I am humbled and extremely grateful to have worked with Nandita Bose and Ashwini Nadkarni. Professionally and personally, you each have meant much more to me than the typed words on this page can ever express. Nandita: you taught me so much, from monomers and heterodimers to how to catch the UGA bus (it's *free*?). And you listened to all my "Oh-My" antics/stories at the Tate Center. Ashwini: Our discussions from the rain in Bombay to the Nickel columns and buffer solutions of the HPLC were all worth their weight in

Gold. I know you feel left behind, but you'll be fine, because from that train station in India to Room 302 Pharmacy, you were meant to be here. I wish you nothing but the best in your future with your loving husband Yash.

I also have to give a 'Bang Head Here' thanks to lab-mates: Laura-Lee Kelley, Betty Ngo, Avis Scott (we still have to race), Michael Pedreira ('techno' Mike), and Philip Mar and Cassie Inman. Cassie, Avis, Phillip and Ashwini, when one of you guys hit the ear signal, the rest of us will *always* do whatever it takes to try and rescue you.

Outside the lab, but inside the department: "The PBS Family": Mervin Williams, Tiffany Adams, Carey Hines, Summer Lewis, Solomon Garner, Babatunde Olubajo and Kim Hill. Can any of you imagine doing this without each other? I can't, so we'll always be in touch forever. Special thanks to old head family members Chandra Brown and Monica Grandison. New-school family status was endowed upon Johnetta Farrar. Honorary PBS Family status was bestowed upon my good friend, Marc A. Grimmett, a recent Ph.D. recipient from the Counseling Psychology department. Guoping Su? We miss you, and uh, thanks for crushing me in badminton.

Judi-Lee Nelson, also a Counseling Psychology graduate, was the key that opened my Ph.D. door. I am humbled by that fact and to have her presence in my life - again. The delicate beauty of Sheriase Sanders, inside and out, has been a blessing in my graduate school career even before *she knew* she was in my life. Thank you Shindari, I am hugging you *again* right now.

Along the way at UGA, I met and must thank individuals who supported, be-friended and encouraged me. Curtis Byrd, Monique "GRE" Harris, Uneisha Minor, Chastity Cyprian, Tameika Chambliss, Angela Black, Jeff Bond, Marilyn Wilson, Helen Riley and Sean Chaplin - thank you. Also thanks to the epic personal and professional support of Michele P. Godwin, "L".

Penultimate, this is the part where I stand up and give a thunderous standing ovation to my boys from Wash, DC and Cleveland, Ohio as *they* come across the stage one by one. Sometimes, I have no idea what I did to have such outstanding individuals come into my life as my friends. Kelley “Schnake” Warren, Daniel Mushala, George “Morales” Brownlee, Keith Smiley, Mike Carter, Harold Oliver, Jason Henderson, Alicia “Nici” Montgomery, Evette Cordell and James Shepard, I’ll have to thank you all personally – and I’ll try not to break down – individually at some later time. I even have to thank my Main Mormon Man, Steve Bryar. Douglas “Sage” Hoston, Ed “Staph Aureus” Saxon and Chauncey “C-money” Williamson, thank you for your whatever-it-takes-I-got-your-back support of my grad school career and the trips you all made to Athens.

Lastly, let me thank my wonderful parents: Two people who spent most of their lives as educators in our native Washington, DC. They set the bar high for their son. I love you both dearly, and this degree is dedicated to you and each brick of 3127 NE.

Everyone, listen, I will spend the balance of my entire life paying you all back for the life-support you have given me – and that may not be enough.

Houston out

# TABLE OF CONTENTS

# Chapter I

## HIV, Drug Design & Capsid Discussion

### *AIDS & HIV Statistics and History*

Human Immunodeficiency Virus (HIV); the causative agent of Acquired Immune Deficiency Syndrome (AIDS) has caused 3.2 million cumulative human deaths worldwide in the year 2003. With an additional 5 million people being diagnosed with the virus each year, the worldwide total stands at more than 45 million infected and living with HIV (Centers for Disease Control). Almost 30 million of the planet's total infected population resides in sub-saharan Africa with another 6 million coming from South and Southeast Asia. Both of these areas of the globe struggle with low literacy rates, political turmoil, poverty and insufficient access to health care that is further complicated by cultural misunderstandings of viruses themselves and the nature of how diseases are spread (Folch *et al.*, 2003).

In some nations, the more staggering statistics come from the percentage of the population living with HIV or full-blown AIDS. Avert, a subdivision of the World Health Organization, estimates 38.8% of the overall population of Botswana as being infected with HIV/AIDS. High percentages of infected populations also exist in Zimbabwe (33.7%), Namibia (22.5%) and South Africa (20.1%). In the United States, as of December 2001, 774,467 AIDS

cases had been reported to the Centers for Disease Control and Prevention (CDC). The state of Georgia, USA, ranks 9th in the United States in total number of AIDS cases.

With the staggering current and projected future numbers of human sickness and death, the amount of research attention garnered is equally enormous. 41,834 documents can be found with a single search of HIV/United States within the CDC web files. As of late December 2003, the National Library of Medicine's PubMed returned 132,811 journal publications in response to a simple word search/query for HIV. By late March 2004, the number had grown to 135,920, which clearly shows the plethora of worldwide focus via the more than 1,000 article-per-month rate of addition to the body of literature.

AIDS was first discovered in homosexual men in the United States, and was soon observed in other groups, including users of intravenous drugs (IV), hemophiliacs, recipients of blood transfusions, sexual partners of HIV/AIDS patients and eventually, infants born to mothers with the disease. Such findings drew the attention of the CDC, which published a newsletter in 1981 discussing 5 cases. The CDC began asking in 1982 that all AIDS cases be reported.

***HIV as a retrovirus***

The tremendous international attention paid to HIV has overshadowed other members of the retrovirus family. Retroviruses can be divided into two groups: transforming and cytopathic. The transforming retroviruses induce changes, often via oncogenes, in cell growth that lead to cancer. In this group are bovine leukemia virus, avian type C virus, mammalian type C virus and the widely studied human T-cell lymphotrophic virus type 1 and 2 (HTLV-1 and HTLV-2). HTLV-1 causes T-cells to overexpress the high-affinity receptor for

interleukin 2 (IL-2). As the cell secretes IL-2, it thereby auto-stimulates its own division in an unregulated fashion causing T-cell leukemia. Cytopathic retroviruses are members of the lentivirus family. One branch of the group includes visna virus, equine infectious anemia virus and feline immunodeficiency virus (FIV). The other branch of this group includes human immunodeficiency viruses 1&2 (HIV-1, HIV-2) as well as simian immunodeficiency virus (SIV).

***HIV Infection of target cells***

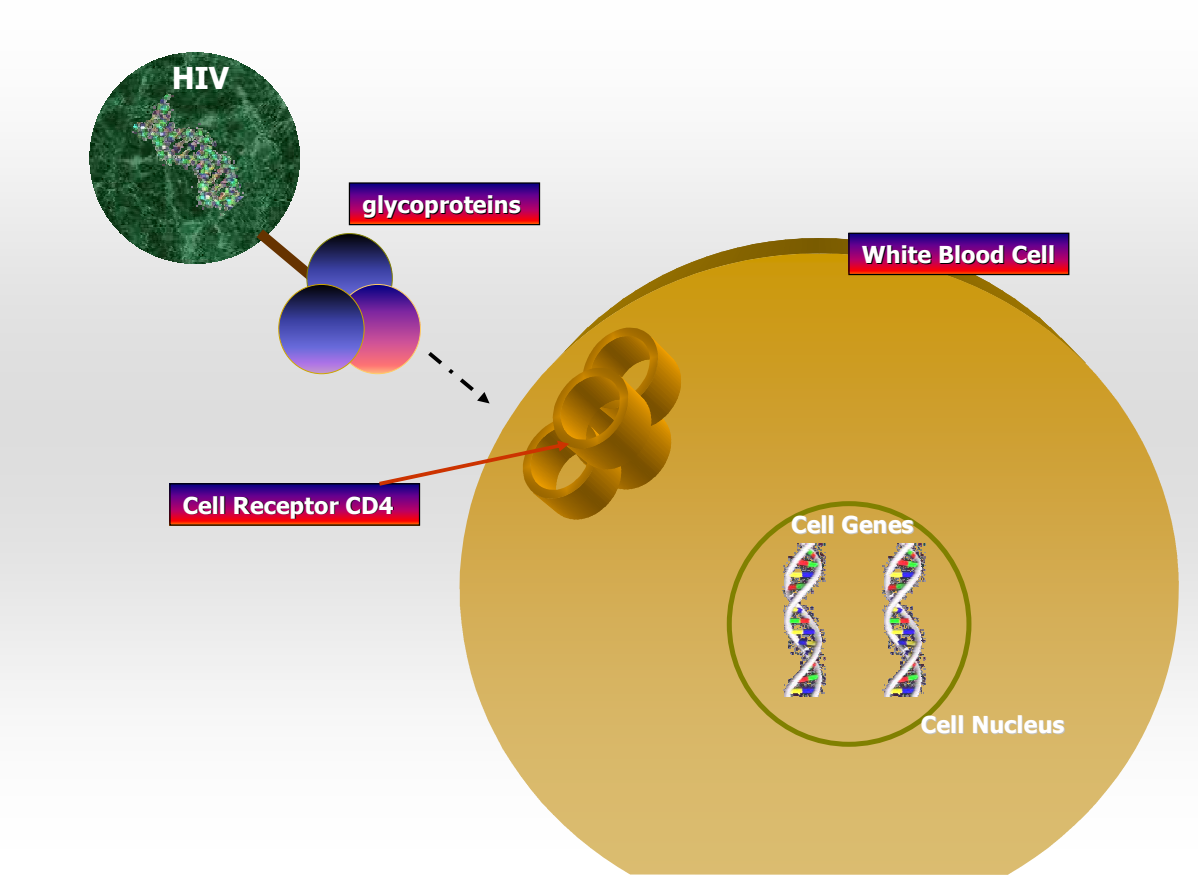



**Figure 1.1 - HIV Attachment to Leukocyte. As the virus approaches the leukocyte, it locates the CD4 complex along the cell surface. HIV glycoproteins gp41 and gp120 bind with CD4 facilitating entry into the host white blood cell.**

Entry of HIV into target cells involves two steps: binding of virions to receptors on target cells followed by fusion of the viral envelope with the plasma membrane of the target cells. The two envelope glycoproteins, gp120 and gp41, make up the surface projections of HIV and play vital roles in these initial steps in HIV infection. These viral glycoproteins bind with the cellular receptor CD4 that is predominately found on $T_{4H}$ cells (see Figure 1.1).

HIV-1 has a 25 fold higher affinity for CD4 receptors than HIV-2 (Kuby 1994) that may explain its greater pathogenicity. CD4 alone is not sufficient for HIV entry. Additional proteins, such as CD26 (as well CCR5 and CXCR4) bind to another location on gp120 called the V-3 loop (Callebut et al., 1993).

Once the HIV RNA genome has been introduced into a target cell and uncoated, a DNA copy is synthesized by the reverse transcriptase enzyme. The viral DNA then integrates into the host-cell genome forming a provirus, which can remain in a latent stage for years. Integration of the HIV DNA into the host-cell chromosomal DNA is mediated by the viral enzyme integrase, which is packaged together with the reverse transcriptase enzyme in the virion. Once integrated, the viral DNA is permanently associated with the host-cell DNA and is passed on from daughter cell to daughter cell during generational mitosis (Kuby 1994). In the latent state, viral genes are not expressed and HIV is able to remain hidden from the host immune system.

Proviral activation initiates transcription of the structural genes into mRNA, which is then translated into viral proteins. As the viral proteins begin to assemble within the host cell, the host-cell plasma membrane is modified by insertion of gp41 and gp120. The viral RNA and core proteins then assemble beneath the modified membrane, acquiring the modified host plasma membrane as its envelope during budding.

Lysis of the host cell versus survival of the host cell depends on the level of CD4 expressed on the membrane (Kuby 1994). As viral gp120 is expressed on the cell membrane, it binds to CD4. If the CD4 level is high, this membrane auto-fusion destroys the membrane's integrity as the virus exits (budding from) the cell, resulting in lysis and cell death (see figure1.2).

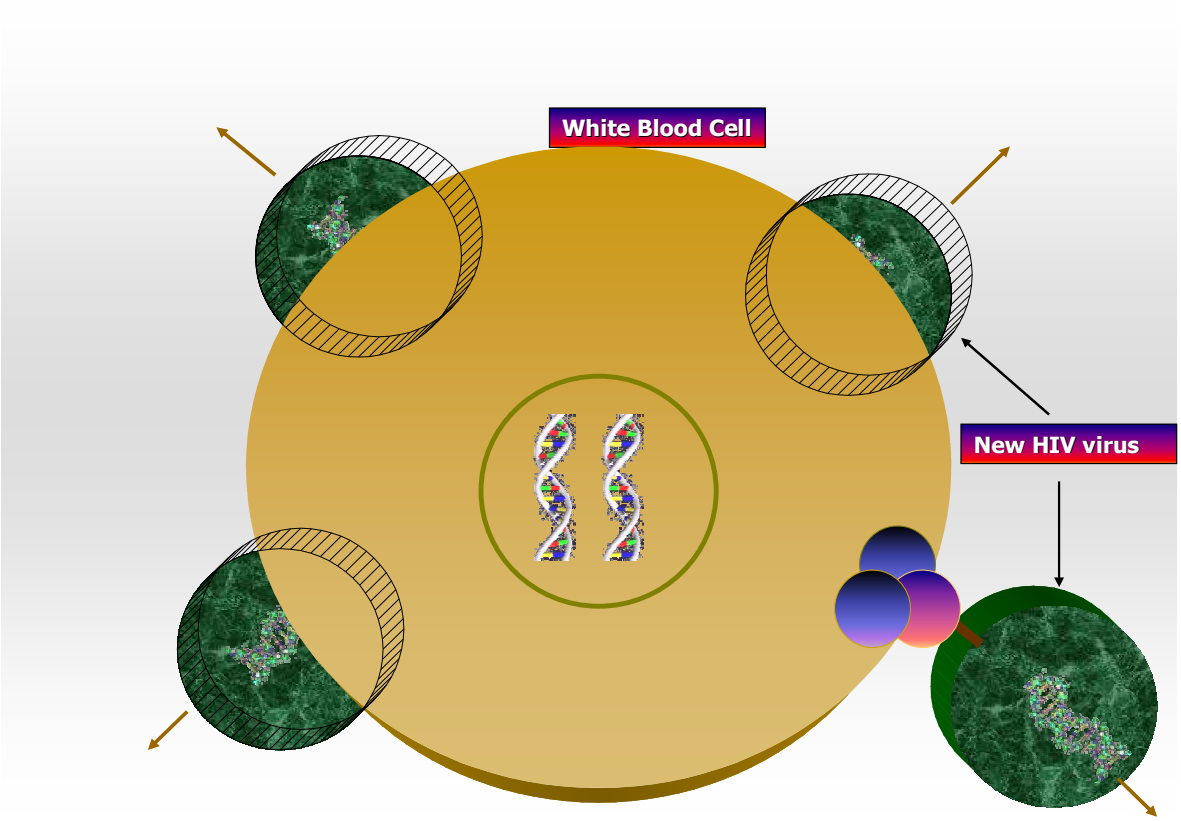



**Figure 1.2 Viral Budding & Cell lysis. As mature HIV virions depart the leukocyte they rupture the cell surface, causing lysis of the cell.**

However, if CD4 expression is low (as in macrophages, dendritic cells and monocytes) the budding of HIV does not lead to extensive membrane damage and the cell continues to live, producing low levels of HIV. With their high levels of CD4 expression, $T_4$ helper cells are at the greatest risk for lytic annihilation and their decreasing levels can be an indication of the progression of AIDS.

***HIV Genome***

The formation of the mature HIV-1 virion involves at least two major assembly processes, one for the viral envelope and the other for the viral core (Cann, *et al.,* 1989, Dickson *et al.,* 1984, Wills *et al.,* 1991). Synthesis, processing and glycosylation of the envelope precursor occur in the endoplasmic reticulum. The provirus contains the *gag, env* and *pol* genes, which encode respectively the viral core proteins, the surface envelope glycoproteins and the non-structural proteins required for replication. The entire HIV genome contains six additional genes: Virion Infectivity Factor (*vif*), Viral Protein R (*vpr*), Transactivator (*tat*), Regulator of Expression of Virion proteins (*rev*), Negative Regulatory Factor (*nef*), and HIV-1 viral protein *U* (viral protein *X* is found in HIV-*2*). Three of these genes, *tat*, *rev* and *nef* encode regulatory proteins that control expression of the structural and enzymatic genes *gag, pol* and *env*. In common with all other retroviruses, HIV-1 encodes all internal structures of the virus within one single gene: *gag*, located in the 5' part of the genome (Figure 1.3) (Jones *et al.,* 1997).

Figure 1.3 HIV-1 Genome

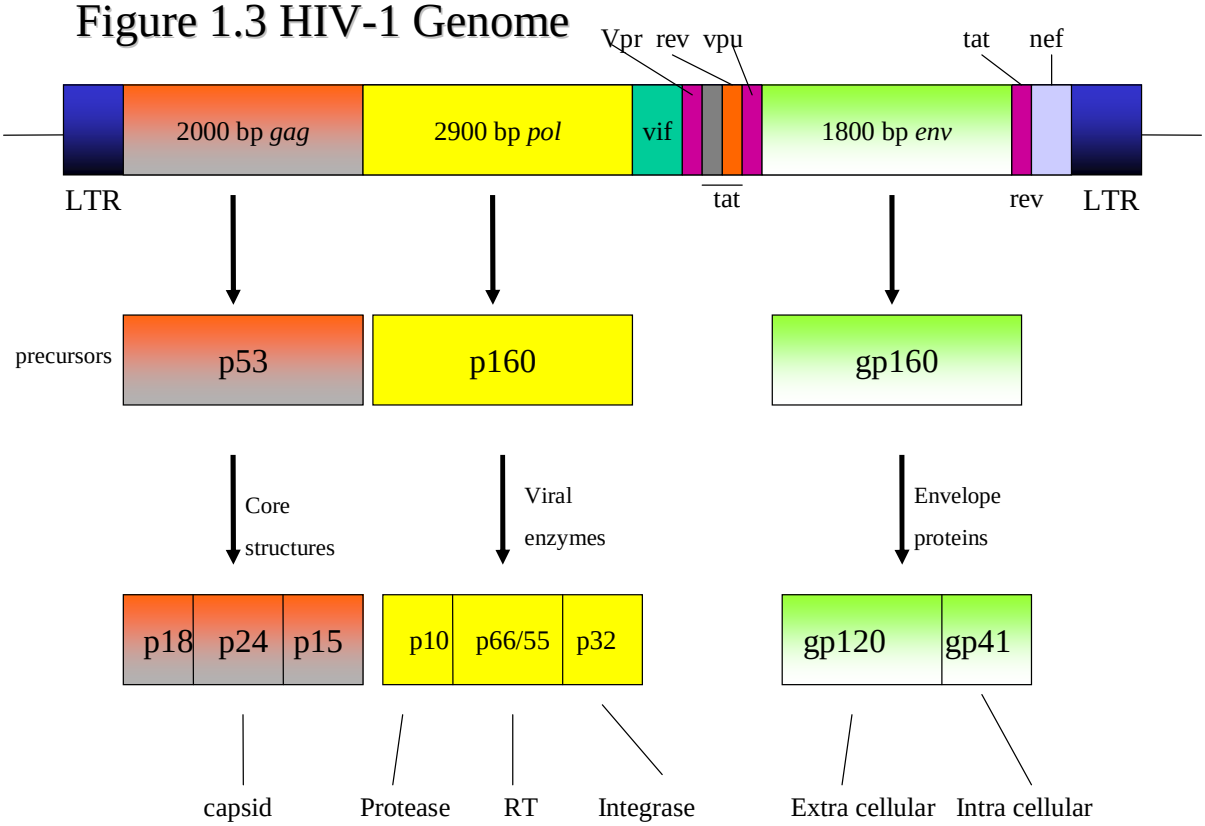



**Fig 1.3 Three main domains lead to 3 main precursor proteins and eight subsequent protein products.**

The *gag* gene products are then translated from an unspliced RNA as a 55-kDa polyprotein precursor ($Pr55_{gag}$). Translation and co-translational myristylation of this polyprotein is followed by an assembly stage, which consists of several temporally ordered events (Cann, *et al.*, 1989, Dickson *et al.*, 1984). During or shortly subsequent to the particle release, cleavage by the viral proteinase yields the following mature Gag products (listed in order within the gene from amino to carboxyl terminus): p17 (Matrix), p24 (capsid), p7 (nucleocapsid) and a small proline-rich protein (p6), as well as the *pol*-encoded domains for the viral enzymes protease, reverse transcriptase and integrase (Mervis *et al.*, 1988, Henderson *et al.*, 1992). Assembly takes place on the cytoplasmic

side of the cell membrane *simultaneously* with the budding and release of immature viral particles (See Figure 1.2). In developing a 'pack and go' strategy, the virus simplifies the mechanism of assembly in that it forfeits the need to evolve a packaging signal in every protein required to form an infectious particle. Instead, all necessary functions are bundled together in a single polyprotein which is assembled and then unpacked (by cleavage) and re-arranged once the virus is formed (Jones *et al.,* 1998). These assembly events represent a key stage in the viral life cycle. Indeed, it is generally true that the specific mutations that allow assembly, but produce aberrant particles, are non-infectious (Barklis *et al.,* 1998).

### *Highly Active Anti-Retroviral Therapy (HAART): Current Multi Drug Therapy for HIV*

Individuals infected with Human Immunodeficiency virus type-1 (HIV-1) are treated with combinations of drugs available from the 16 currently approved antiretroviral agents (Jordan *et al.,* 2003). The target for each one of the 16 is one of the two viral enzymes: protease (PR) or reverse transcriptase (RT). Three classifications exist for each enzyme. Protease inhibitors bind to the active site of protease, non-nucleoside RT inhibitors (NNRTI) bind directly to the RT molecule, and nucleoside RT inhibitors act as chain terminating substrates during reverse transcription.

Despite the development of HAART, a therapy that combines three to six different inhibitors from at least two different drug classifications, it is still impossible to fully eradicate all traces of the virus from patients. Aspirations for current treatment strategies involve

suppressing the viral load levels (the number of free virus particles in the blood plasma) for as long as possible.

Long-term limitations to HAART therapy arise from strong side effects (ranging from nausea, fatigue and diarrhea to anemia, elevated cholesterol and glucose all leading to a decrease in patient compliance) and the evolution of drug-resistant variants (Beerenwinkel *et al.,* 2003). Viral replication can be found in a variety of tissues and cell types even in patients with viral loads suppressed to below detectable limits (50 copies/ml). Persistent virus production is further facilitated by sub-inhibitory drug levels in infected cells or by outright host immune failure. Thus, preexisting or newly produced drug resistant mutants can emerge that have a selective advantage under drug pressure. These escape mutants become dominant in the virus population and lead to viral rebound and therapy failure (Jetz *et al.,* 2000).

The genetic basis of drug resistance is HIV's high mutation rate, approximately $3x10^{-5}$ per nucleotide per round of replication due to the faulty mechanisms of the DNA polymerase. This stems from HIV not having a proof reading mechanism to check and/or balance its high replication rate. Polymorphisms in the viral genome have been linked to drug resistance. In protease alone, 49 of the 99 residues that could alone or in combination render the enzyme resistant have been observed to have amino acid changes (Weber *et al.,* 2002). Thus, alternative targets and therapies must be created to support, if not replace the existing strategies for individuals diagnosed with HIV.

### *Drug Design Targeting HIV-1 Protease*

While enzymatic inhibition generally focuses on competitive or non-competitive binding at the substrate active site, research done on HIV-1 protease inhibition has not been just limited to small molecule inhibitors. Rozzelle *et al.,* (2000) discussed creating defective heterodimers of protease by altering from one to three residues critical around the homodimeric active site. Substitutions D25K, G49W and I50W allow not only for substrate binding inefficiency but also for formations of defective heterodimers over wild-type homodimers. The replacement of lysine for catalytic aspartate-25 may stabilize the heterodimer by both (1) hydrophobic interactions between the methylene groups of the lysine side chain and the hydrophobic substrate-binding pocket and (2) favorable charge-charge interactions with the aspartate of the wild-type monomer (McPhee, 1996). Replacement of Gly-49 and Ile-50 with tryptophan induces favorable interactions with the Sp1 and Sp2 recognition sequences and destabilizes homodimers formation via steric and electrostatic repulsion.

A compound has been developed (QF34) that binds to the protease in quite an unusual binding manner. Its side-chains do not fit tightly into their respective pockets, but are positioned between them (Weber, 2002). The characteristic mutations, leading to weaker binding of inhibitors such as saquinavir or indinavir to the mutated protease species, thus do not influence the binding of QF34, possibly resulting in the development of a new class of protease inhibitors.

With innovative studies on dimer-interfaces and atypical compound classes, the HIV protease enzyme has been fertile soil for copious investigations regarding structure based drug design. The crossroads of ligand-based (shape-similarity) and receptor-based (shape-complimentarity) drug design strategies has unfolded a new field: Shape Signatures. This style of

computer aided design has been applied by Zauher *et al.,* (2003) to design new inhibitors predicted to be active against HIV protease. Their results, which focused more on shape than on chemical properties of candidates, was done versus the National Cancer Institute (NCI) database and yielded a list of 50 best inhibitors within 11.6 CPU hours.

Other computer aided, structure-based studies assist in the therapeutic targeting of highly conserved residues in an attempt to avoid the non-essential residue mutations that HIV protease undergoes. Using energy minimizations and molecular modeling of the substrate inhibitor MVT101, a new synthetic tetra-peptide has been developed which interacts with several conserved residues (Siddiqui *et al.,* 2001).

The HIV protease active site motion during the actual binding process has been studied via the program F-DycoBlock, which takes receptor flexibility into account. The known inhibitor L700417 was examined in regards to accuracy of recovery, binding energy, solvent accessible surface area (SASA) and positional root-mean-square (RMS) deviation. Protein flexibility was accurately associated with each stage of drug design: search for the binding sites, dynamic assembly and optimization of candidate compounds.

### *Drug Design Targeting HIV-1 Reverse Transcriptase*

Perhaps no other single mechanism in the HIV life cycle is more problematic for researchers than the error-prone, mutation-causing enzyme, reverse transcriptase (RT), which copies the RNA genome into double-stranded DNA. An example of the practical dilemma in the fight against AIDS can be seen by comparison to the flu. As vaccines for the influenza (common flu) virus have always been hampered by the virus' mutation rate, HIV has a mutation rate that is

65 times greater than influenza (Kuby 1994). Reverse transcriptase introduces an estimated 5-10 errors during each round of replication. Sequencing reveals that not only do no 2 individuals have the same exact virus, but that isolates taken from the same individual at different times can vary substantially in sequence (Kuby 1994).

There are several non-nucleoside inhibitor trials relying on structure-based and/or docking procedures. New pyrrolyl aryl sulfone (PAS) compounds have been designed and synthesized via studies with SYBYL (Silvestri 2003). The newly designed PAS derivative is characterized by a *p*-chloroaniline pharmacophore that enables the $NH_2$ to hydrogen bond with the carbonyl oxygen of Lys101 within the NNRTI binding site. One PAS has an $IC_{50}$= 0.05 μm. Other locations of interactions within the site, such as at Tyr181, Tyr188 and Trp229 are intended to assist this compound in binding even in the face of an ever-altering binding site.

DOCK 4.0 has been utilized to predict binding energies between the commercially available efavirenz (SUSTIVA™) within the NNRTI binding site with favorable results (Wang 2001). This group utilized MM-PBSA (Molecular Mechanics Poisson-Boltzmann/surface Area) to identify the correct binding mode, which has a binding free energy about 7kcal/mol more favorable than the next best binding energy. Moreover, the calculated binding free energy (-13.2 kcal/mol) is in reasonable agreement with experimental (-11.6 kcal/mol). These results, which included modeling the complex within the structure, were achieved without prior knowledge of the structure of the efavirenz/RT complex. These findings illustrate that molecular docking in combination with modeling analysis is an attractive approach for understanding energy dynamics of ligand-receptor interactions.

The Molecular Operating Environment (MOE) assisted the group led by Zhou *et al.,* (2002) to dock known and proposed non-nucleoside RT inhibitors in the NNRTI binding site. Three separate charge schemes (PEOEKLMN, PEOEMF and MF) were used to compile lists of the five lowest energy configurations of Nevirapine$^{R}$ in comparison to two compounds proposed to have inhibition within the same site.

***Drug Design Targeting the HIV-1 Enzyme Integrase***

HIV-1 integrase, the third enzyme originating from the p160 precursor, is essential for retroviral replication. It is involved in the integration of HIV DNA into host chromosomal DNA and appears to have no functional equivalent in human cells (Maurin *et al.,* 2003). These qualities make integrase a rational and attractive target for selective therapy. However, there is no current therapy available today that targets the Integrase enzyme. While several integrase inhibitors have been shown to have activities in the micromolar ranges (Rao *et al.,* 2002) *in vitro*, they have yet to be proven *in vivo* and move forward clinically.

The lack of compounds involving this crucial step in the life cycle of the virus has in part led towards efforts involving structure-based design and computer modeling. Maurin *et al*., (2003) and Chen *et al*., (1998) reported the structure-activity relationships of HIV-1 integrase inhibitors expected to interact within or near the active site. Specifically discussed is the emergence of diketo-acids (DKAs) and dicaffeoyltartaric acids as a result of the recent report detailing the crystal structure of the integrase core domain. Inhibitory compounds were categorized by their different proposed mechanisms of action as well as their proximity of

binding to the active site. Results of the study were both ligand-based (pharmacophore) and target-based (docking).

Perola *et al.,* (2000) detailed the feasibility of using virtual screening as an approach to drug discovery with their work on the metalloprotein farnesyltransferase (FT). Virtual screening had come under scrutiny concerning validity due to the lack of studies to show that metalloproteins such as HIV-1 integrase, matrix metalloproteins and farnesyltransferase were viable drug targets via this method. Utilizing the program EUDOC (an upgrade from their own program SYSDOC) and the 219,390 compounds from the Available Chemicals Directory (ACD) they identified 21 compounds having an $IC_{50}$ range from 25 to 100 μm. This stands in stark contrast to the $IC_{50}$ results stemming from 21 *randomly* selected compounds, none of which have an $IC_{50}$ lower than 100μm.

***Drug Target Design for the HIV-1 protein Gp120***

There is also considerable attention being paid to glycoproteins 120 and 41, coded by the *env* as part of the next wave of hopeful therapeutics for HIV. These surface proteins are essential for viral entry as they bind to CD4 expressing cells although CD4 independent attachment has been reported in conjunction with CCR5 and CXCR4 cell surface proteins (Geijenbeek *et al.,* 2000).

Co-receptor binding involving CXCR4 has been targeted via AMD-3100, a non-peptidic, low molecular weight bicyclam compound. AMD-3100, which has also moved on to clinical trials, prevents the electrostatic interactions between CXCR4 and gp120 and completely blocks

signal transduction from CXCR4, inhibiting replication of T-trophic HIV-1 (DeClerq *et al.*, 1994).

Glycoprotein 41 is specifically targeted by the phase II clinical trial-level synthetic peptide T-20. Homologous to 36 conserved residues within the C-terminal heptad repeat region of gp41 (Rimsky *et al.,* 1998), T-20 binds to the highly conserved hydrophobic groove of gp41 C-peptide that normally mediates the conformational change from a pre-hairpin intermediate to a fusion-active hairpin, thereby inhibiting HR1-HR2 complex formation, preventing membrane apposition, fusion and entry (Chen *et al.,* 1998).

Zollner *et al*., (2001) discussed compounds under development, including fusion inhibitors that block viral cell entry by targeting the HIV envelope protein gp41. Cooley *et al*., (2003) detailed two compounds that are currently in the clinical trial phase. Each agent targets gp120-CD4 binding, as HIV-1 almost always infects CD4 expressing leukocytes. Attachment inhibitor PRO452 is CD4-immonoglobulin $G_2$ ($IgG_2$), a recombinant antibody like fusion protein designed to bind and neutralize HIV-1 prior to cellular attachment. Cyanovirin-N (CV-N) has four gp120 binding sites thereby surpassing the number of sites available on monomeric CD4 or dimeric CD4-heavy chain constructs.

***HIV-1 Capsid p24 – Structure, function and physiological properties***

The distinct conical structure of mature HIV-1, as opposed to the typical spherical form seen in other retroviruses, is directly due to the capsid shell (Hoglund *et al.,* 1990). The protective shells of retroviruses like HIV-1 are more complex than icosahedral capsids, however

similar principles of conformational control also apply. Mature retroviral particles contain their genome as a condensed ribonucleoprotein (RNP) core encased in a proteinaceous cone-shaped capsid shell and surrounded by a lipid bilayer derived from the host cell (Nermut & Hockley 1996). The transformation from immature particles to mature structures of the mature virion is accompanied by proteolytic cleavage.

This cleavage is not only essential for infectivity while leading to the condensation of the inner core/capsid shell but it serves to convert the stable immature capsid shell into a metastable mature core (Gross *et al.,* 2000). Maturation also involves the translocation of the capsid protein from a peripheral position to a more internal position relative to the viral membrane as seen in figure 1.4 (Gelderblom *et al.,* 1987). A virion with capsid still in the peripheral location is labeled immature and is not infectious.

A complete understanding of how the capsid complex disassembles upon viral entry into the host cell has not been realized. HIV-1 (but not HIV-2 or Simian Immunodeficiency Virus (SIV)) requires the cytosolic protein cyclophilinA (CypA) from the host cell for infectivity (Braaton, 1996). CypA is specifically incorporated into HIV-1 virions via its interaction with the capsid domain of Gag (Franke, 2002; Thali, 2002; Luban, 1993). The CypA binding site is located in the amino terminal domain localized to a proline-rich flexible exposed loop.

# Figure 1.4 Virus Assembly

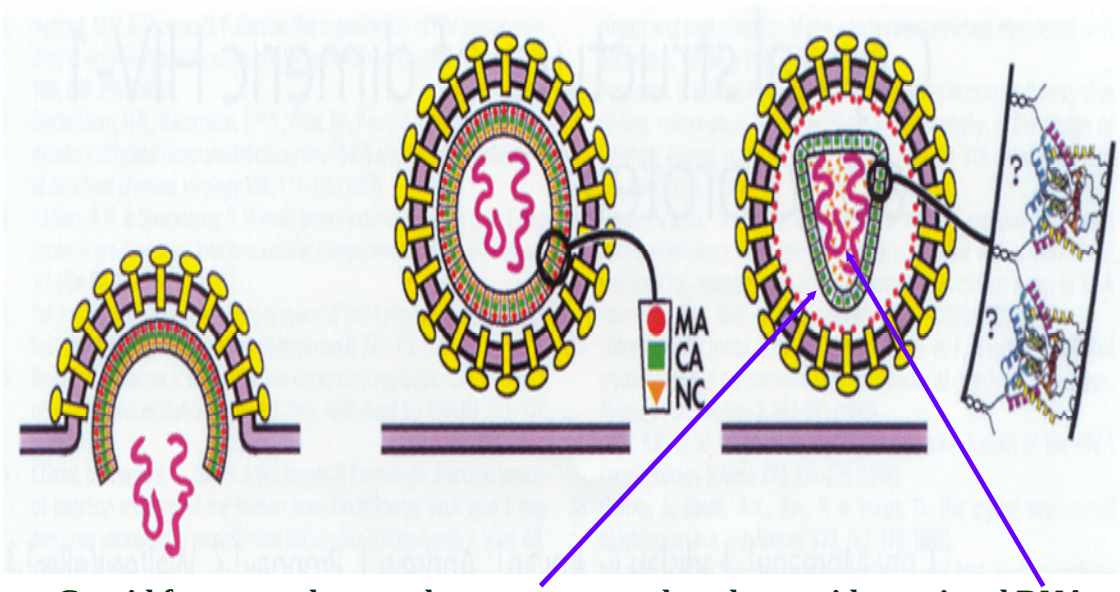



**Figure 1.4 – The role of the GAG polyprotein and capsid in the assembly of the HIV virion (used with permission from Nature Structural Biology - permissions@natureny.com). The capsid begins as part of the *gag* polyprotein while the mature virion has dimer units collectively forming a shell around the RNA.**

The domain encompasses residues 85-93 which includes the important residue Pro90. Disruption of the cyclophilinA-capsid interaction either by binding cyclosporin and its analogs by mutagenesis of residues in the binding site, or by deletion of the CypA gene blocks CypA incorporation into virions and greatly reduces HIV-1 viral replication and infectivity (BonHomme *et al.*, 2003). For every 10 Gag molecules, approximately one single CypA molecule is incorporated into the virion (Ott *et al.*, 1997).

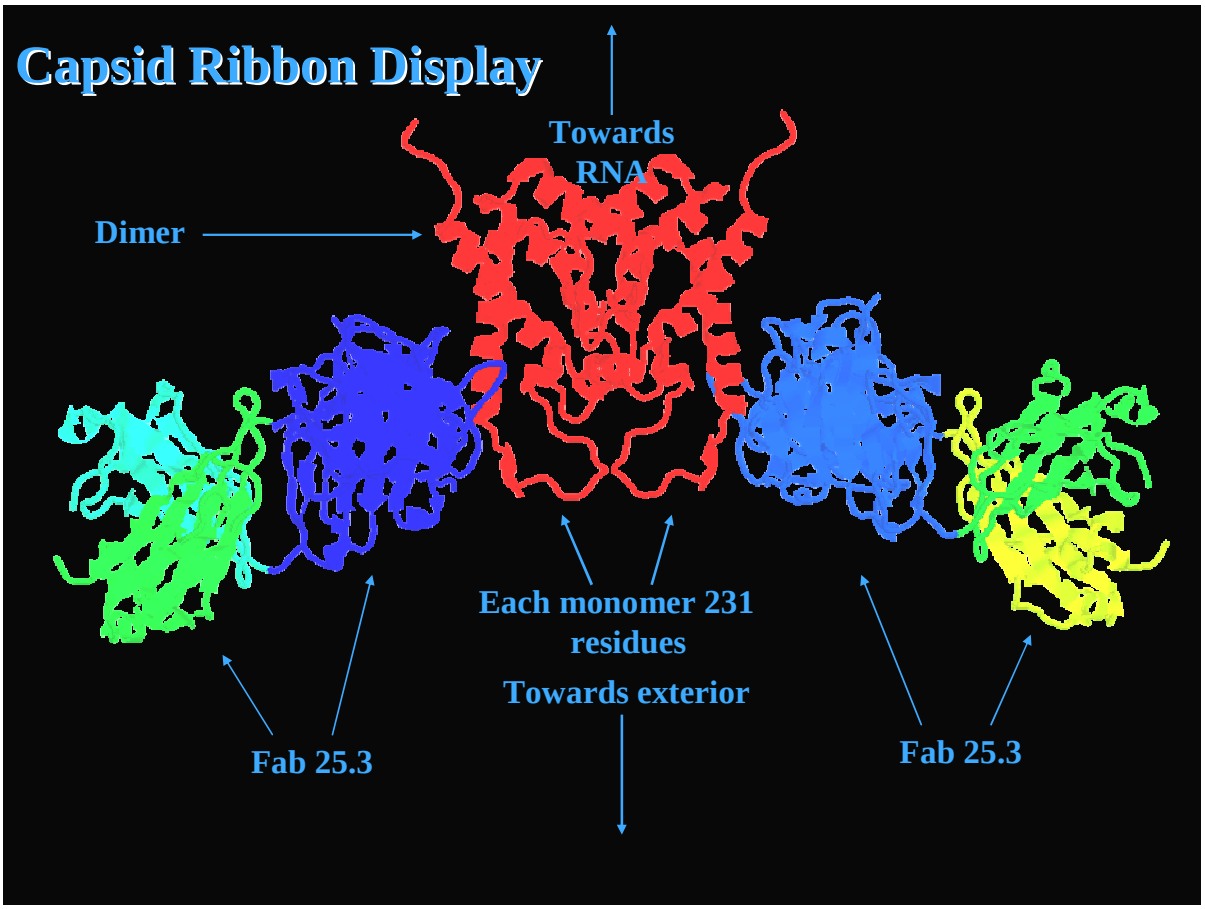



**Figure 1.5 – Dimeric Capsid Ribbon Display representation with associated Fragment Antibodies. (Figure prepared using Rasmol)**

Momany *et al.*, (1996) showed that the wild-type is a dimeric capsid protein (see Figure 1.5) with each monomer consisting of 231 residues, and forms an electron dense, elongated core (Figure 1.4) within the virion (Momany *et al.*, 1996). Towards the amino terminus of the capsid is small spacer peptide (Sp1), and the carboxyl terminus of the protein is spacer peptide 2 (Sp2). Both are targets for viral protease in the separation of the capsid from the upstream matrix and downstream nucleocapsid gene products (BonHomme *et al.*, 2003).

The spherical shape of the immature HIV particles is due to the presence of an N-terminal matrix protein sequence extension on the capsid domain (Gross *et al.*, 1998). The

liberation of this sequence during viral maturation leads to the cylindrical particles capable of dimer and oligo formations (Erlich *et al.,* 1992). It has also been determined that cylindrical particles were also observed when the initial 13 residues were eliminated from the Capsid while up to but no more than five residues upstream of the Pro1 (into the protease cleavage recognition sequence) were also allowable for cylindrical formation (Gross *et al.,* 2000). Sequences after the C-terminus of the capsid are not required for spherical particles as were seen with extension N-terminal beyond the aforementioned five residues.

Alterations in ionic strength and pH affect the assembly and disassembly of the capsid (Ehrlich *et al.,* 1992). Assemblies of oligomeric structures are stable under various conditions *in vitro*. High ionic strength salt (1M NaCl) or non-ionic detergent (0.1% octyl-beta glucopyranoside) led to limited disassociation only after several weeks of incubation, indicating that the oligomers were actually quite stable under these conditions. In addition, oligomerization occurred more readily at pH 7.0-8.0 than pH 6.0 as shown by dynamic light scattering techniques (BonHomme *et al.,* 2003).

The significance of the capsid protein towards the overall fortitude of HIV and hence, its value as an attractive drug target can be illustrated with two main points: 1) mutations which inhibit assembly are lethal (Schwedler, 1998; Gross, 1998; Tang, 1998; Reicin, 1996; Forshey, 2002) and 2) mutations which alter capsid stability hamper replication (Forshey *et al.,* 2002). Therapies have been developed targeting capsid proteins in picornoviruses such as polioviruses (Smith 1986). Deres *et al.,* (2003) recently identified inhibitors for capsid assembly in hepatitis-B (HBV).

The capsid shell of HIV-1 remains an elusive target for drug design. Gitti *et al.*, (1996) and Gamble *et al.,* (1996) each theorized that significant structural differences lie between the predominantly beta-sheet configuration of non-retroviral capsid monomers and the majority helical arrangement of retroviral capsid proteins. This difference in structure may well entail how the downfall of strategies anticipated to inhibit capsid formation up to a decade ago have yet to come to fruition (Rossman 1998).

### *Amino Terminus Discussion of Capsid Protein*

The N-terminal amino acid of the capsid is a proline that forms a salt bridge with Asp51 (their proximity to one another can be seen in Figures 1.6 through1.8). This interaction is the paramount molecular interaction of the drug target pocket. The folding back of the initial 13 residues within this beta-hairpin loop of the capsid is the final step towards completion of the monomer (Tang *et al.,* 2003). Locating compounds that emulate the nitrogen group of Pro1 and interact with the carboxyl group of Asp51 is the central objective of this dissertation. Further re-enforcing the significance of Pro1 and Asp51 in the stability of the final capsid protein are the results of homology studies in comparable retroviruses. Sequence alignment analysis with Rous Sarcoma virus (RSV), Mouse Mammary Tumor Virus (MMTV), Bovine leukemia virus (BLV), Feline Immunodeficiency Virus (FIV), residues 51 and the amino terminus proline were seen to be highly conserved throughout (Momany *et al.,* 1996). Each of these homologs contains the N-terminal proline and the only substitution for aspartic acid as residue 51 is to glutamic acid (E) that harbors similar properties and is equally capable of interacting with the amino group on the

cyclic N-terminal proline (carboxyl group made readily available, thrusting the dual oxygens into the pocket).

BonHomme *et al.,* 2003 studied the N-terminal extension of the capsid with a hexa-histidine tag to prevent the Pro1-Asp51 salt bridge and examine the resulting immature capsid formations. Not only was this study key in understanding that cleaving of the tag was similar in task to the wild-type clipping done by protease at the Matrix-Capsid (sp1) site changing the capsid from hollow spheres to cylindrical particles, but that specific physiological conditions must exists as well. A pH value of 6.0 or more and the capsid monomers undergo condensation to an *oligomerization-ready* form. This form not only has a high potential to self-associate into the overall capsid shell, but it also has an increased capability of contacting CypA.

To better understand how the capsid assembles *in-vitro*, Gross *et al.*, (1998) created various N-terminal extensions of the capsid. The number of amino acids required to convert the phenotype from wild-type spherical particles to tubular (cylindrical) formations were evaluated. Their discovery that the spherical shape of immature HIV capsid particles is determined by the presence of an N-terminal extension on the capsid domain and that core condensation during virion maturation requires the liberation of some N-terminus of the capsid illustrates the importance of the amino terminus.

However, there are a few more interactions occurring within the N-terminal “pocket”:

- The side chain of Isoleucine 15 packs into a hydrophobic binding site (Figure 1.7).
- A second H-bond between the amino on Pro1 and Gln13 oxygen (Figures 1.6, 1.7).
- van der Waals contacts between the invariant Pro1 ring and the C-alpha atoms of Ile15 and Gly46 (Figures 1.6, 1.7).

- H-bonding interactions between the two strands of the B-hairpin (Figure 1.8).

Significant evidence exists in the literature (Tang 2002, Schwedler 1998) showing that a salt bridge between Pro1 and Asp51 stabilizes the beta-hairpin structure at the amino-terminus and finalizes the folding of the protein. The carboxyl group of the highly conserved negatively charged Asp51 points directly into the pocket, available for potential binding with the N-terminus. Hence, the search for compounds that emulate the ring amino group of N-terminal Proline is the central ambition for retardation of the capsid protein's final formation.

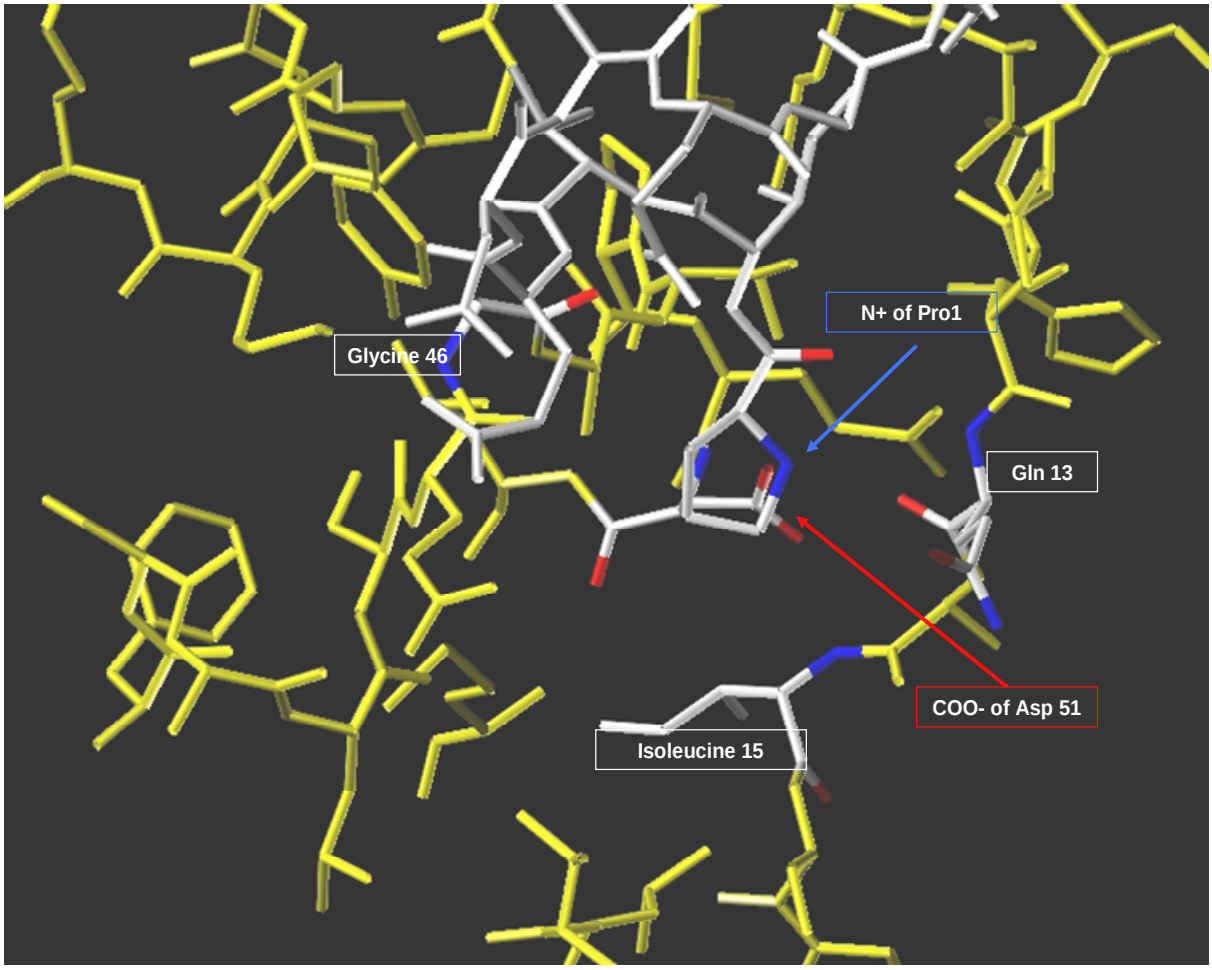



**Figure 1.6 – Key residues in the N-terminal pocket of the HIV-1 Capsid. The proximity of the Pro1 and the Asp51 can be seen, as well as the location of residues GLN13, GLY46 and ILE15. The oxygen of Gln13 is seen to project into the pocket, providing an H-bond opportunity for Pro1. The C-alpha atoms of Ile15 and Gly46 enable van der Waals' interactions to occur with the Pro1 ring.**

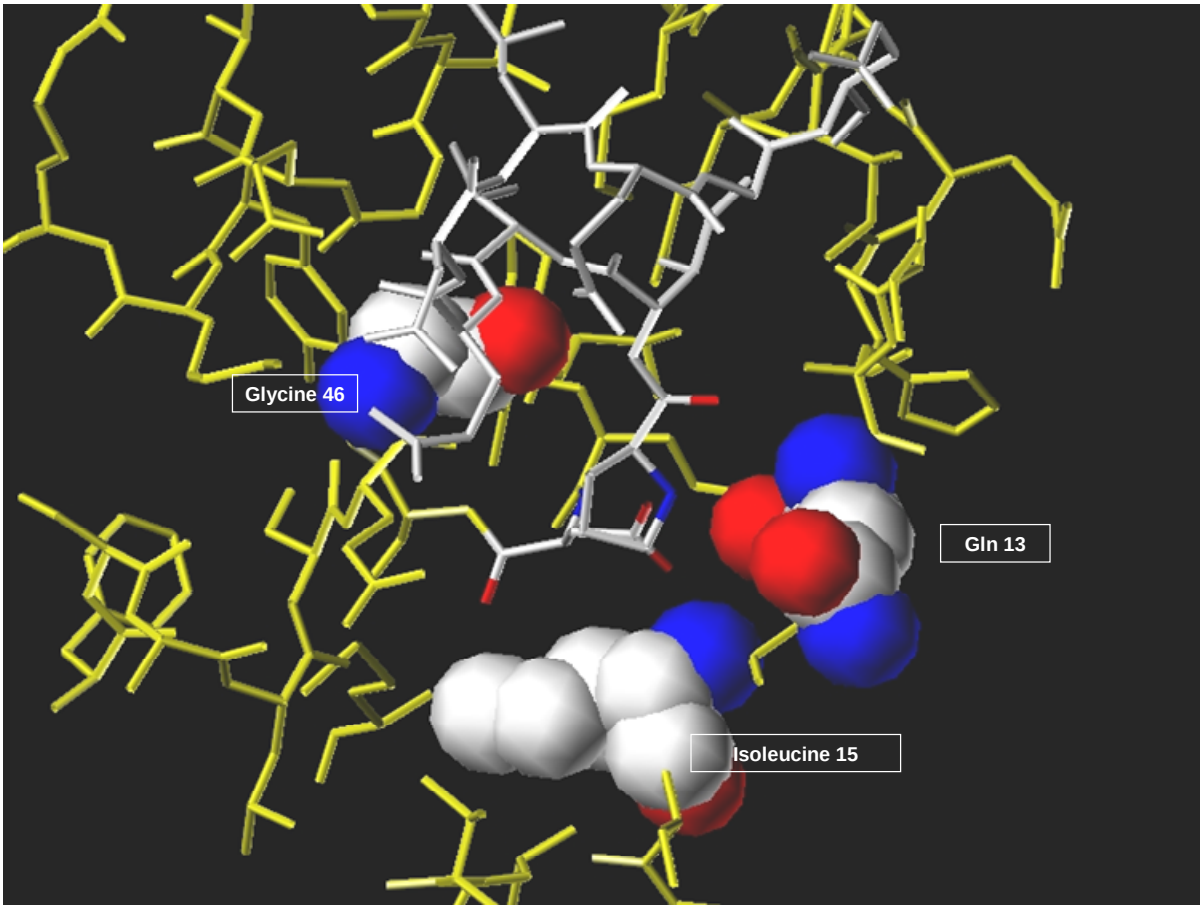



**Figure 1.7 Residues G46, I15, Q13 shown in space-fill within the HIV-1 capsid N-terminal pocket. The projections of the oxygen atoms (in red) are more evident when displayed as how they truly occupy room within the pocket.**

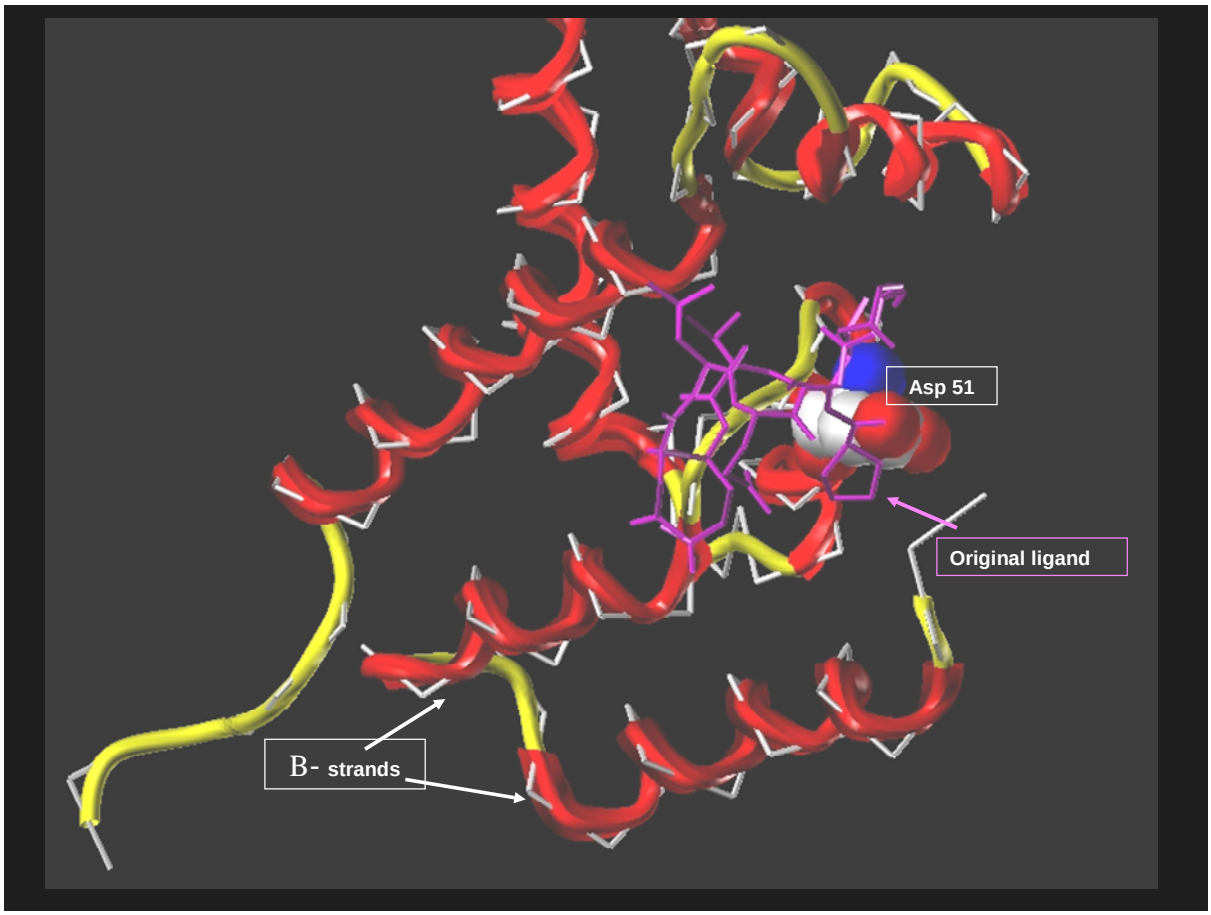



**Figure 1.8 – C-Alpha backbone showing secondary structure as a tube, Asp51 as space-fill. The label beta-strands points towards the beta-hairpin turn that occurs at a secondary structure level.**

# Chapter II

## Virtual Drug Screening

The essence of medicinal chemistry is discovering new therapeutic molecular agents to bind, inhibit or activate a bio-molecular target. Historically, discovery was a serial process, screening and optimizing compounds at a singular rate (Lamb 2001). While the development of high-throughput robotic methods (Houston 1997) hastened the process of screening large corporate databases and combinatorial libraries, it is still not feasible to screen all available compounds experimentally. While recent laboratory advances in combinatorial chemistry have dramatically increased the number of available compounds (Gordon, 1999; Suto, 1999), the hit rates when screening a large unbiased library are frequently very low. In addition, many libraries have not yet been associated with interesting biological activity (Lamb 2001).

Consequently, a critical need has developed for fast, reliable computational methods for virtual screening of large three-dimensional (3D) libraries and databases to predict the putative geometry of protein-ligand complexes. Receptor-based, virtual screening uses knowledge of the target protein structure to select candidate compounds with which it is likely to favorably interact (Koh 2003). This type of drug design has been shown to work well for excising potential ligands from the plethora of compounds that exist in contemporary databases (Amzel *et al.,* 1998; Marrone *et al.,* 1997).

Several algorithms for the structure-based drug design and evaluation of combinatorial libraries have been developed (Sun *et al.,* 1998; Kick *et al.,*1997; Haque *et al.,* 1999; Murray *et*

*al.,* 1999;) and the advantage of structure-based, directed library design in favor of designs based only on molecular diversity has been shown (Kick 1997). It has been estimated (Balkenhohl 1996) that the number of organic compounds with molecular weights < 750 approaches $10^{200}$, whereas $10^{12}$-$10^{15}$ is a practical limit on the size of a library that may be screened via computer (Walters *et al.,* 1998). The virtual library must then be reduced to a smaller library, containing the compounds most likely to be active against a given target.

Although the fundamental goals of screening methods are to identify those molecules with the shape, hydrogen bonding, and electrostatic and hydrophobic interactions for the target receptor, the complexity of the problem is in reality far greater. As an illustration, the ligand and the receptor may exist in a different set of conformations when in free solution than when bound. The entropy of the unassociated ligand and receptor is generally higher than that of the formed complexes. Favorable interactions with water are lost on binding. These energetic costs of association must be offset by the gain of favorable intermolecular protein-ligand interactions (Koh 2003).

The magnitude of the energetic costs and gains is typically much larger than their difference and thus, potency becomes extremely difficult to predict even when relative errors are small. Although methods have indeed been developed to account for the strength of molecular association events from entropic and solvation effects (Reynolds 1992, Zhang 2001), these methods are costly in terms of computational time and are inefficient for the virtual screening of large compound databases.

### *Surface Representations*

The basic description of a protein or ligand surface is the atomic representation of exposed residues. However, such a representation is usually used only when based on real potential energy functions (Halperin 2002). In the grid approach, the atomic details of the ligand and of the receptor-binding site are simulated explicitly while the other bulk portions of the system are represented as grids.

Grids aside, the more frequent characterization of the protein surface is via its geometric features (Luty 1995). Some of the origins of protein surface analysis begin with Connolly (1983). The Connolly surface consists of the part of the van der Waals surface of the atoms that is accessible to the probe sphere (contact surface) connected by a network of convex, concave, and saddle shape surfaces that smoothes over the crevices and pits between the atoms (Connolly 1983). Based on the Connolly analysis the surface is described by sparse critical points (Lin 1994, Lin 1996) defined as the projection of the gravity center of a Connolly face.

The program MS-DOT calculates discrete points along with three types of surface faces representing the molecular shape (Connolly 1983). For each face an interest point and a normal are computed. The interest point is a cap, belt, or pit for convex, toroidal, and concave faces, respectively (Lin 1994).

With the origins of surface representation relating to simple conformational search and ranking of potential solutions; other features have been added as enhancement. Physiochemical features of the protein surface are added into the purely geometrical description. The protein surface may also be fitted with spherical harmonic functions to include electrostatics (Ritchie 2000).

***Algorithms: The impetus of docking programs***

The dilemma in developing practical virtual screening methods is to utilize a search algorithm that is rapid enough to evaluate potentially millions of compounds while maintaining sufficient accuracy to correctly identify compounds (or subsets of compounds) with significant hits. Those are two contradictory requirements and compromises have to be made. Recent advances include higher clock speeds of modern computers, representation of the surface of a protein by molecular surface, scoring functions (which will be discussed later) and parallel computing (Gabb *et al.*, 1997; Jiang *et al.*, 1991; Kuntz *et al.*, 1982; Norel *et al.*, 1994).

The underlying foundation of the docking methodology is the force field or *docking function* defining the energetics of the system. It is this target function that all docking algorithms attempt to optimize (Wu *et al.*, 2003). The next major development in the molecular docking came from the application of grids to store physio-chemical properties of the receptor. The grid only needs to be calculated *once*, assuming the receptor is rigid. Programs map the extended van der Waals radii onto a grid, which could be used to explore the unoccupied surface/volume with a ligand (Busetta *et al.*, 1983). In the program GRID, for each of the grid points, the potential energy of the interaction between the receptor and a probe atom/group is calculated (Goodford *et al.*, 1985). Some probe groups are amino, methyl, and hydroxyl groups. The grids thus display the kind of interactions (polar and non-polar) that are preferred at each position in the binding site.

*Disadvantages* for GRID implementations that use interaction energies between receptor and ligand are not easily pinpointed (Broomijmans, 2003). Two different implementations (Pattabiraman 1985, Tomoika 1987) allow for direct energy evaluation while manipulating the

system. Pattabiraman *et al.*, (1985) mapped van der Waals and Coulombic contributions of the receptor onto the grid. Another implementation builds on Goodford's GRID and maps the interactions of different probes with the receptor on the grid, which is stored in a look-up table. The interaction energy of the ligand with the receptor can be estimated by summing up the interaction energies of each probe atom that corresponds to a ligand atom. In addition to the van der Waals and Coulombic interactions, a grid is also made to map the hydrogen bonding properties (Tomoika 1987). The first docking programs with algorithms involving grids in an automated fashion were DOCK (Meng 1992) and AutoDOCK (Goodsell, 1990). The time saved with implementation of grids permitted itself in the development of more sophisticated scoring functions.

With flexible docking, the search algorithm explores different positions of the ligand in the receptor active site translational, torsional (hence, the flexibility) and rotational degrees of freedom. Ligand flexibility algorithms fall into three categories: stochastic, systematic and deterministic.

Stochastic search algorithms make random changes, involving one degree of freedom of the system at a time. A major unknown variable of stochastic searches is the uncertainty surrounding convergence (Broomijams 2003). Multiple, independent runs can be done to improve convergence. Monte Carlo (MC) methods and evolutionary algorithms are examples of this type of searches.

Systematic search algorithms are based on a grid of values for each formal degree of freedom, and each of those grid values is explored in a combinatorial fashion during the search. As the number of degrees of freedom increase, the number of evaluations *needed* increases

rapidly. Termination criteria are inserted to prevent the algorithm from sampling space that is known to lead to the wrong solution. An example of systematic search is the anchor-and-grow (incremental) construction algorithm (to be discussed in further detail later as this is the mechanism utilized by DOCK 4.0).

In deterministic searches, the initial state determines the move that can be made to generate the next state, which generally has to be equal to or lower in energy than the initial state. Deterministic searches performed on exactly the same starting system (including each degree of freedom) with the same parameters will generate exactly the same final state (Golke 2000). Problems with deterministic algorithms arise when they get trapped in local minima because they are unable to traverse barriers (Broomijmans 2003).

The first incremental construction algorithm for docking was described by DesJarlis *et al.,* (1986) and was incorporated into the widely used program DOCK. Incremental (anchor and grow) construction algorithms divide a ligand into rigid and flexible regions. One (or more) rigid "anchors" with flexible parts are defined by perception of rotatable bonds. Most implementations dock the anchor first, with the flexible parts added sequentially, with systematic scanning of the torsion angles. Several procedures developed since the initial incremental construction algorithms have been able to search the degrees of freedom of the flexible part more explicitly during the docking. Leach and Kuntz (1992) docked the anchor rigidly first, with the flexible portions added later. Each of the dihedral angles was sampled systematically, but the number of dihedral angles allowed *per angle* was kept low – to prevent combinatorial explosions.

The DOCK 4.0 algorithm (discussed at length anon) docks the anchor based on steric complementarity. The flexible parts are grown incrementally; the dihedral degrees of freedom

are explored and minimized. Pruning occurs at each step of the growth to ensure diversity. When the molecule is complete, it is re-minimized and the final score is calculated (Ewing 2001). Minimization is possible because the scoring function is analytic in form and differentiable.

Compare this with FlexX (Rarey *et al.*, 1996) which docks the anchor based on chemical interactions instead of steric complementarity (using an algorithm called "pose clustering"). With fewer matches accepted with chemical complementarity, a clustering algorithm is implemented to merge similar transformations of the ligand into the active site (Rarey *et al.*, 1996).

The program ICM implements the principles of Monte Carlo algorithms, using pseudo-Brownian motions (Totrov 1998). However, AutoDock was the first docking program to implement simulated annealing MC methods using tens of thousands of steps performed during each cycle, reducing temperature at the beginning of each cycle (Goodsell 1990). Other methods that use MC methods are MCDOCK (Liu 1999), Prodock (Trosset 1999) and PRO_LEADS (Murray 1999).

The most widely acclaimed example of an evolutionary algorithm is the program GOLD. Jones *et al.*, (1997) were actually the first group (via GOLD) to use a docking algorithm tested on a large set (> 100) of protein-ligand complexes. GOLD uses multiple subpopulations of the ligand, rather than a single large population, and manipulates these simultaneously.

FLOG and EUDOC are docking programs under the category of Pre-generated Conformational Libraries. These libraries are an efficient way of taking ligand flexibility into account because the cost of generating multiple conformers per molecule only has to be incurred once and the internal energy of the conformers can be assessed prior to docking. Each of the generated conformers will be docked rigidly into the receptor, and its fit with the binding site

will be determined. The program FLOG generates and docks conformational libraries called Flexibases (Kearsley *et al.,* 1994; Miller *et al.,* 1994) and its algorithm is similar to that of DOCK. EUDOC uses conformational searches of the ligand to generate different ligand structures. Each of the structures is then placed in the active site of the receptor via a systematic search (Pang *et al.,* 2001). An energy evaluation is subsequently performed for each receptor-ligand complex.

Another major goal of a good docking function is to discriminate between the manifold of true solutions, usually defined as poses within 2.0 Angstroms root mean square deviation (RMSD) from the X-ray geometry, and false solutions or incorrectly docked structures (Vieth, *et al.,* 1998; Roche *et al.,* 2001). The importance of RMSD can also be illustrated by comparing the docked binding mode with the experimental binding mode. A successful prediction of a binding mode is when that calculated difference is below the aforementioned 2.0 angstrom value (Verdonk *et al.,* 2003).

***Scoring Functions: The limiting utility***

While the soundness and vulnerabilities of docking algorithms are well documented in the literature, a strong case can be made that the major weakness of docking programs currently lies not in the algorithms, but in the inaccuracy of the techniques used to *estimate the affinity* between receptor and ligand - the scoring functions (Ajay *et al.,* 1995, Bohm *et al.,* 1999, Tame *et al.,*1999). These scoring functions are required to serve two purposes: during the ongoing docking process, they optimize orientation and conformation for comparison with *other* ligand molecules and they are used to estimate binding affinity for the fully docked molecule. Although

in principle different functions can be used for these two purposes, in most applications the same function is used (Stahl *et al.,* 2003).

A weak scoring function equates to a substandard docking program. There are various criteria for evaluating the quality of scoring function: its ability to identify the correct binding mode of a ligand out of alternative docking solutions (Taylor *et al.,* 2002); its ability to rank related ligands with respect to their binding affinity and its ability to select a number of (however weak) inhibitors out of a large database of inactive compounds (Stahl *et al.,* 2003).

Of primary importance, scoring functions employed in library docking must be extremely swift. Many terms and variables associated with the full thermodynamic cycle defining free energy in solution may very well be neglected in achievement of this primary objective (Gilson 1997). In addition, scoring functions must be error-tolerant, since fast, flexible ligand docking approaches crystallographic accuracy only for relatively rigid ligands and in the absence of induced fit phenomena. These numerous limitations keep expectations low for giving accurate affinity predictions.

Nevertheless, scoring functions should recognize solutions displaying good steric and electrostatic complementarity between receptor and ligand and give lower ranks to other solutions with unlikely relative orientations of ligand and receptor groups (Blaney *et al.,* 1993). Functions and the programs they serve should return a full description of the two major driving forces of complex formation: hydrogen bonds and hydrophobic interactions.

Considerable efforts have been devoted to the development of better scoring functions to alleviate the real or perceived weakness that exists in today's docking programs. During the past

few years, four categories of computational methods have been investigated: Force field (or First-Principle), knowledge-based, empirical and semi-empirical.

Force field methods are scoring functions that use the Coulomb and van der Waals terms of force field functions, as seen in DOCK (Meng *et al.*, 1992) and Auto Dock (Goodsell *et al.*, 1990). To account for the screening effect of the solvent on electrostatic interactions, a distance-dependent dielectric constant is used. Internal ligand energies and entropic terms are completely ignored. EUDOC even uses this brand of force-field function *without* a grid (Pang 2001). Stoichet *et al.*, (1999) later added the effects of the solvent on protein-ligand interactions using implicit solvent methods. The van der Waals interactions were calculated using the Leonard-Jones potential; the electrostatic interaction between ligand and receptor was estimated using a pre-calculated receptor potential, which solves the Poisson-Boltzmann equation. These solvent-corrected scores were closer to experimental binding free energies than the regular DOCK scores, but still they were too favorable (Stoichet *et al.*, 1999).

Semi-empirical methods deal directly with the issue of the CPU time required when sampling physically irrelevant states of the ligand. When only small mutations can be made at a time to any given ligand, MC simulations can spend vast amounts of precious CPU time reviewing these intermediate conformations. Aqvist (1996) proposed a method that only sampled the initial and final states of the ligand free in solution and bound to the receptor. GOLD scoring functions fall under this umbrella and the function itself consists of three terms, a hydrogen bonding term, a van der Waals term and an internal energy term (Jones *et al.*, 1997). The total energy is a weighted sum of the three, making it semi-empirical.

Empirical scoring functions (Bohm *et al.*, 1999) try to capture those elements of binding free energy that are intuitively deemed important by a sum of terms – mainly hydrogen bond, contact surface and entropic contributions – whose relative weights are either derived from experimental data or by physical reasoning. Empirical scoring functions in docking are based on receptor-ligand structure properties rather than on ligand properties alone (Broomijmans *et al.*, 2003). However, there are disadvantages, the first being the difficulty of knowing what each term exactly accounts for and to assess where errors stem from. In addition, binding free energy predictions can only be successful if the molecules make similar interactions to the ones in the training set complexes (transferability issues). Finally, pH, salt concentration and temperature can influence the measured binding constants greatly and these conditions are ignored when calculating free energies from experimental binding constants. This limits training of empirical functions and accuracy of predicted binding free energies. The first program to use empirical scoring functions to predict binding free energies was LUDI (Wang *et al.*, 2003).

Conversely, knowledge based scoring functions are derived from statistical analysis of protein-ligand atom pair distances in x-ray structures of protein-ligand complexes (Muegge *et al.*, 1999; Mitchell *et al.*, 1999; Gohlke *et al.*, 2000). Converting the frequency of the atom-atom interactions using Boltzmann distribution into potentials derives the term Potentials of Mean Force (PMF). The developers of FlexX and DrugScore both utilized PMF in creating those programs (Sadowski *et al.*, 1990). The main difference between knowledge-based and empirical potentials is that no binding data are needed. This is a big advantage to devise relatively large training sets.

Any docking program of high standards will have the ability to reproduce the experimental binding modes of ligands. This is usually tested by taking a ligand out of the X-ray structure of its protein-ligand complex and docking it once again into its original binding site.

The suite of programs called DREAM+2 has also been used for docking computationally generated ligands into macromolecular binding sites. It is composed of three separate entities: ORIENT++, REACT++ and SEARCH++. The program ORIENT++ positions molecules in a binding site with the DOCK algorithm. Its output can be used as input to REACT++ and SEARCH++. REACT++ performs user-specific chemical reactions on a docked molecule, so that reaction products can be evaluated for three-dimensional complementarity with the macromolecular site. SEARCH++ performs an efficient conformation search on the reaction products using a hybrid backtrack and incremental construction algorithm (Makino *et al.*, 1999).

TreeDock addresses the issue of enumerating possible high resolution docking orientations in a rigid-body search. By representing molecules as multidimensional binary search trees and by exploring a sufficient number of docking orientations such that two chosen atoms, one from each molecule, are always in contact, TreeDock is able to explore all clash-free orientations at very fine orientations in a reasonable amount of time (Fahmy *et al.*, 2002). Due to the speed of the program, many contact pairs can be examined to search partial or complete surface areas. The deterministic systematic search of TreeDock is in contrast to most other docking programs that use *stochastic* searches such as Monte Carlo or simulated annealing methods (Fahmy *et al.*, 2002).

### *Protein-Protein versus Protein-Ligand Docking*

Although the physical principles that govern protein-protein association are similar to those responsible for other ligands, docking algorithms designed for protein-protein association differ somewhat from those for small ligands used in drug design (Janin *et al.,* 1993). In protein-protein docking the rigid body approximation is still the standard because of the large number of degrees of flexibility and because it is much harder to predict where the protein interaction site is (Halperin 2002). The scoring function should be soft because some atom clashes are likely to occur even at near-native configurations. When both protein surfaces are sampled fully, the number of generated complexes can be extremely large requiring *efficient* sampling and scoring functions. In comparison with protein-protein interactions, in which electrostatics plays an important role, that role is even more emphasized when docking small molecules into protein targets (Hou *et al.,* 1999). In protein-ligand docking, the complementary contact surfaces between the ligand and the receptor are substantially smaller and less discriminating than in the case of protein-protein docking. Single water molecules in the interface may be particularly important in small ligand docking, mediating hydrogen bonds (Lengauer *et al.,* 1996).

### *Pharmacophore identification of lead compounds*

The starting point of recognition of a pharmacophore is the collection of small molecule ligands that were experimentally observed to interact with the given receptor. The underlying logic is that such an interaction is obtained either via a set of geometric features common to the data set of ligands, or alternatively, they may be chemical attributes, translated into geometrical features (e.g., hydrogen bonds, coordinates of hydrophobic atoms, points representing charged groups, etc.). These features combine to outline a pharmacophore, which is recognized by the

receptor. The ideal situation would be that once a pharmacophore is identified, other ligands with potential for similar functionality can be found by screening for molecules containing a similar constellation. A multiple structural alignment algorithm is the natural method for identifying pharmacophores (Leibowitz *et al.*, 2001).

Finn *et al.*, (1997) developed an algorithm for pharmacophore identification, Randomized Pharmacophore Identification for Drug Design (RAPID). The algorithm is designed to find the structural alignment between a pair of molecules. To extend the algorithm, the group iteratively took all solutions of a certain pairwise problem. For each of these, they generated a new molecule composed of the core found by that solution. RAPID compared the next molecule from the original ensemble against each of these new molecules.

Rigoutsos *et al.*, (1996) developed an algorithm for flexible 3-D structure matching against massive databases of small molecules. For any given database of 3-D structures and a single query molecule as an input, the method determines those molecules from the database, which contain substructures in common with substructures in the query molecule, allowing for torsional flexibility around rotatable bonds.

Miller *et al.*, (1999) have recently developed SQ, an atom-based clique matching, followed by an alignment scoring function that recognizes pharmacologically relevant atomic properties. Pharmacophore searches have also been used in conjunction with DOCK in flexible ligand docking (Carlson 2000). Studies have been done with dynamic pharmacophore construction algorithm and tested versus the HIV-1 integrase. The study tackled the problem of inherent flexibility of the integrase active site and attempted to reduce the entropic penalty that is associated with binding a ligand. Lead compounds were indeed identified using this method. Yet

while constructing pharmacophores and docking compounds containing them has obvious benefits, there are limitations. The obvious trade off is the limit to diversity of lead compounds. Since the volume and shape of targets can change, compounds with various attributes will always be fundamental to therapeutic discoveries.

### *DOCK 4.0*

The analysis and comparison of various docking algorithms, scoring functions and the programs that utilize them, transition to the topic of DOCK 4.0 itself, the central instrument of this individual endeavor. The original DOCK (University of California at San Francisco) algorithm addressed rigid body docking using a geometric matching algorithm to superimpose the ligand onto a negative image of the binding pocket (Eucharis *et al.*, 1988; Meng *et al.,* 1992). Earlier versions saw advancements in database processing, force field based scoring (Gschwend *et al.,* 1996) and on-the-fly optimization (Ewing *et al.,* 1997). The particular version, 4.0 used in this dissertation, utilizes an improved matching algorithm for rigid body docking and a new algorithm for flexible ligand docking.

The anchor and grow algorithm is shown in Figure 2.1. Step one involves the rigid part of the ligand (the anchor) being docked using a geometrical matching procedure on the receptor/target. The resulting anchor positions are then used to start a pruned conformation search in steps 2 & 3. The conformation search is performed breadth-first on each anchor position simultaneously, with the most promising partially built conformations retained during each stage of the search. When finalized, each conformation is locally optimized (step 4) to

relieve any strain incorporated during the construction process. If *additional* portions of the ligand are suitable as anchors, this can all be repeated in step five.

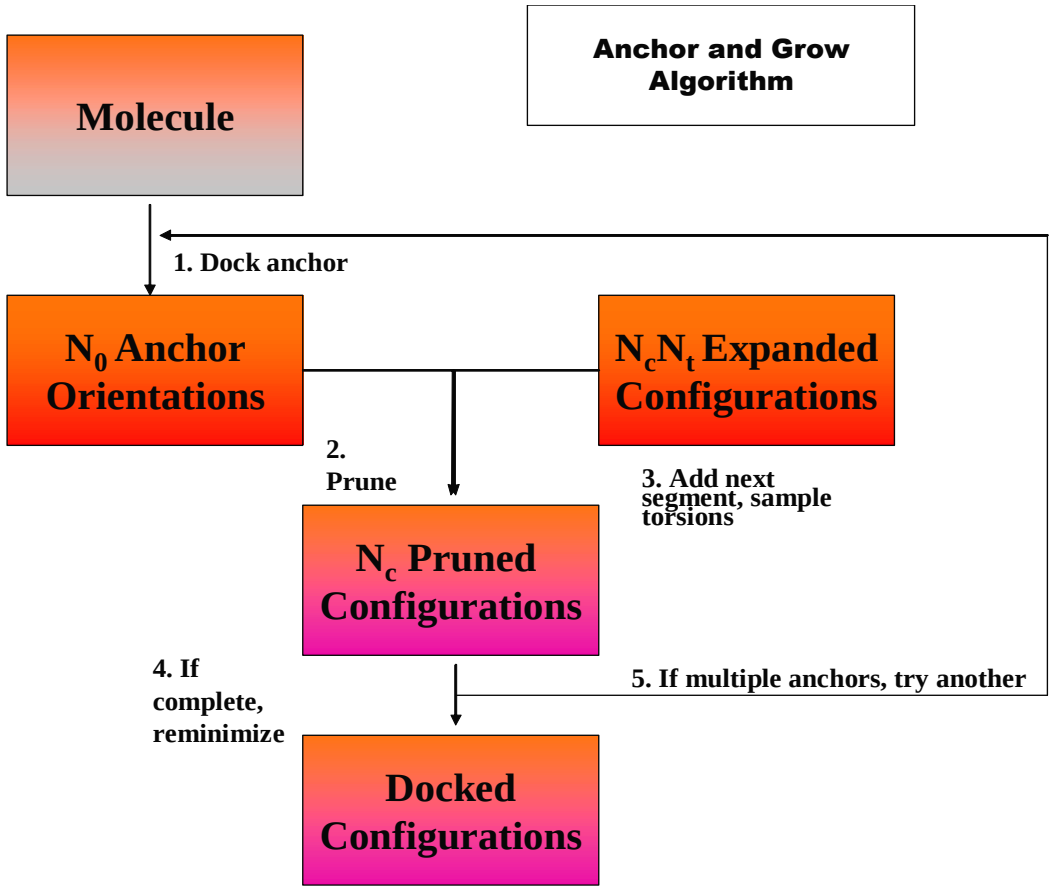



**Figure 2.1 - Anchor and Grow Algorithm**

With DOCK 4.0, the matching based orientation search not only can be performed using a variety of protocols, it can be run manually or automatically. In the manual mode, the user can specify geometric parameters such as the distance tolerance, and the program builds all matches that fit those parameters. Manual controls of matching work better with docking a database of ligands simply because it biases the sampling (Eucharis *et al.*, 1994) towards molecules that contain more internal distance similarity to the site points.

DOCK 4.0 also implements a search technique in which multiple random conformations are docked independently. It is similar to the ‘flexibase’ approach (Leach 1992) in which the molecule database is seeded with multiple conformations of each molecule, and each conformation is docked independently as a rigid molecule. DOCK can reproduce this technique as a default if the flexible ligand option is requested and Anchor-First docking is not requested.

With anchor and grow being implemented, the scoring function of DOCK is simply used to guide intermediate stages of the search (Ewing 2001). Accuracy is paramount since entire modes of binding could be missed because of mistakes in the calculated interactions of a portion of the complex. In this work, the existing molecular mechanical scoring function is kept, but the importance of continued development of more accurate scoring functions cannot be emphasized enough. The anchor and grow strategy constrains the type of scoring function to be atom pairwise decomposable since the interactions of a partially built molecule are evaluated.

# Chapter III

## Materials and Methods

In this chapter the methods of obtaining small molecule leads via DOCK 4.0 (within the N-terminal pocket created by the deletion of the initial 13 residues at the N-terminus of the HIV-1 capsid protein) will be discussed. Databases of molecules were derived from two separate locations. The Cambridge Structural Database is the principal product maintained by the Cambridge Crystallographic Data Centre. The drug discovery and developmental arm of the National Cancer Institute (NCI) is the Developmental Therapeutics Program, (DTP) which houses the NCI DIS 3D Database. The databases combined contain more than 400,000 small-molecule compounds.

The Protein DataBank (PDB) accession number 1AFV, the atomic structure of the capsid protein, was used in docking studies as the receptor (see appendix A). The receptor pocket was defined by removing the first 13 residues from the atomic structure. A separate ligand file was created using the initial 13 residues of the capsid and named “ligand.pdb’ (see appendix A). All other molecules from all other databases and sources analyzed within this experiment were then directly and/or indirectly compared to this original ligand file.

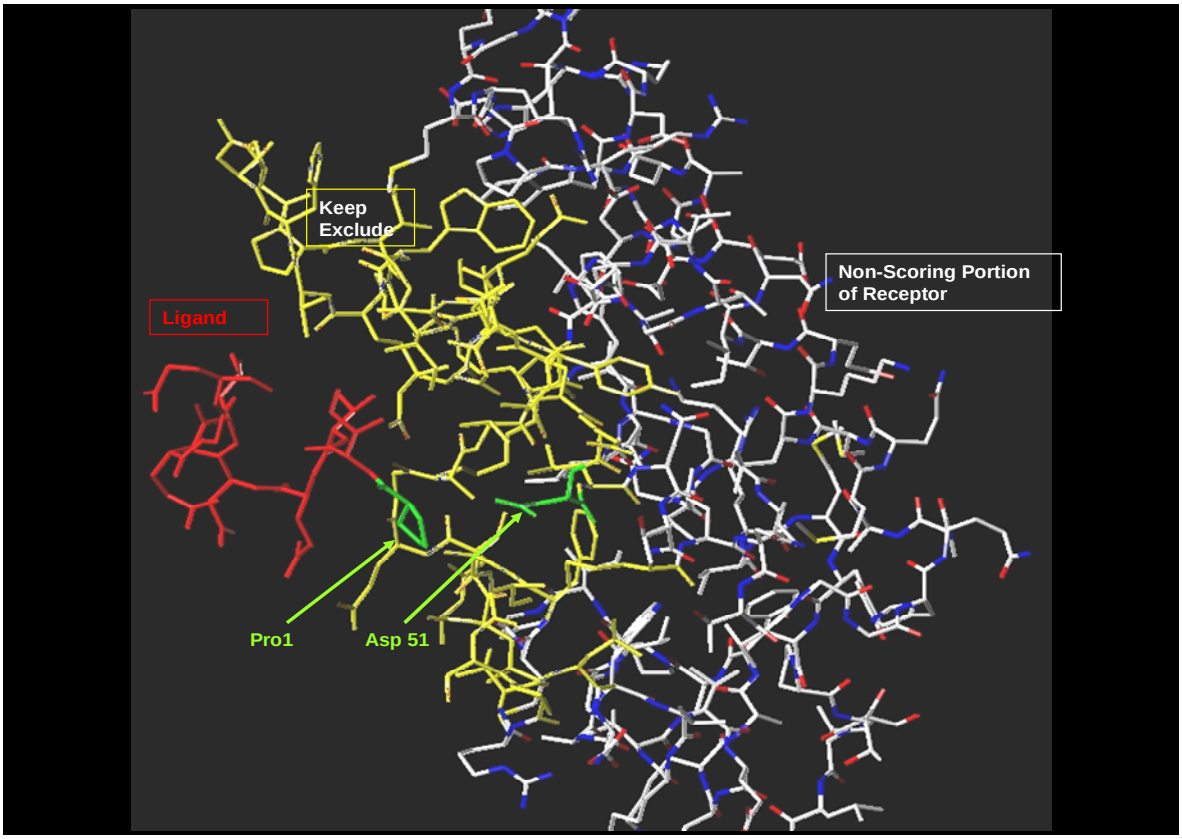
DOCK 4.0 is actually a suite of sub-programs, each of which is briefly albeit awkwardly described in the accompanying loose-leaf manual. Progressing though these programs in an organized manner channels the user through the four basic stages: Ligand Preparation, Site

Characterization, Scoring Grid Calculation and finally, Docking. The overall Program Sequence is displayed in Appendix A.

### *Ligand Preparation*

All residues in the receptor within 10 Å of the ligand were identified by the program get_near_res and place in the file keep_exclude.pdb (see appendix A). Verification of proper structure of both files should be done using visualization with Sybyl. In Figure 3.1, the ligand file can be seen in red within the specified pocket. The molecule in yellow represents the portion of the receptor designated as the keep_exclude region. The excluded portion of the capsid receptor is colored by atom type. The principle interaction of the N-terminus is featured, with both Pro1 and Asp51 shown in green.

**Figure 3.1 - The molecule in yellow represents the portion of the receptor designated as the keep_exclude region. The excluded portion of the capsid receptor is colored by atom type. The principle interaction of the N-terminus is featured, with both Pro1 and Asp51 shown in green. The original ligand is red.**



***Site Characterization***

The sub program 'invertPDB' calculates the surface of the receptor. The script AutoMS prepares an INSPH file for running SPHGEN (DOCK manual). SPHGEN output has two key components: OUTSPH and "receptor.sph". The receptor.sph file was converted into a .pdb file and in order to be viewed by Sybyl by the sub-program showsphere. Showsphere requested the name of the sphere cluster file: receptor.sph. It then requested which of the available subsets (from zero to n) the user wished to turn into a pdb file. Cluster zero was chosen. Upon SYBYL visual inspection, none of the further subsets of clusters gave an accurate description of the ligand in the appropriate pocket, despite the manipulation of variables within subprograms leading to this point. Therefore, the sphereset to be used in further grid and dock procedures will be done with manual sphere deletions within Sybyl.

Figure 3.2 details the shows the spheres used in subsequent programs. The yellow molecule is the keep_exclude file, with the original ligand in orange. Pro1 and Asp51 are represented in green. The spheres accurately follow the shape of the ligand that fills up the volume of the N-terminal pocket. Figure 3.3 has these same spheres within the pocket, surrounded by a dot surfaces file representing a van der Walls' radius (magenta) around each sphere (generation of radii detailed in appendix A). Molecular details of the sphere's vdw radius within the pocket are visualized more concretely with the magnified view shown in Figure 3.4. The proximity and electron donating potential of Asp51, as well as the aggregate number of spheres in the pocket space vacated by Pro1 are unmistakably discernable. The illustration of spheres representing *only* the selected portion of the ligand that fits tightly into the pocket is also demonstrated in figures 3.3-3.6. The stick figure of the ligand in figure 3.6, against the backdrop

of the space-fill keep_exclude region, clearly contains segments that are not affiliated with red-colored spheres.

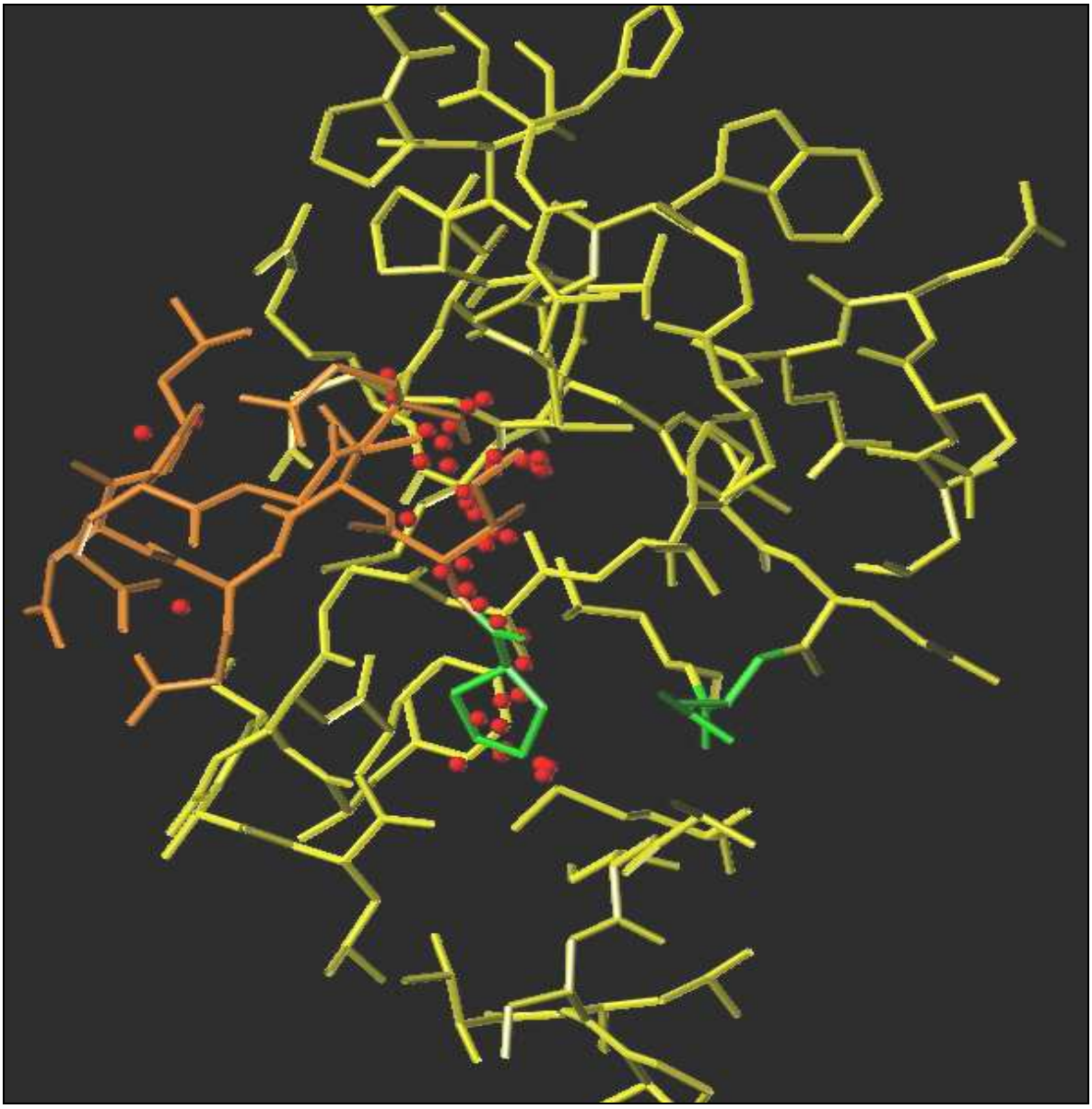


**Figure 3.2 – Generated Spheres (red), original ligand (orange), Keep_exclude region (yellow), Asp51 and Pro1 are shown in Green.**

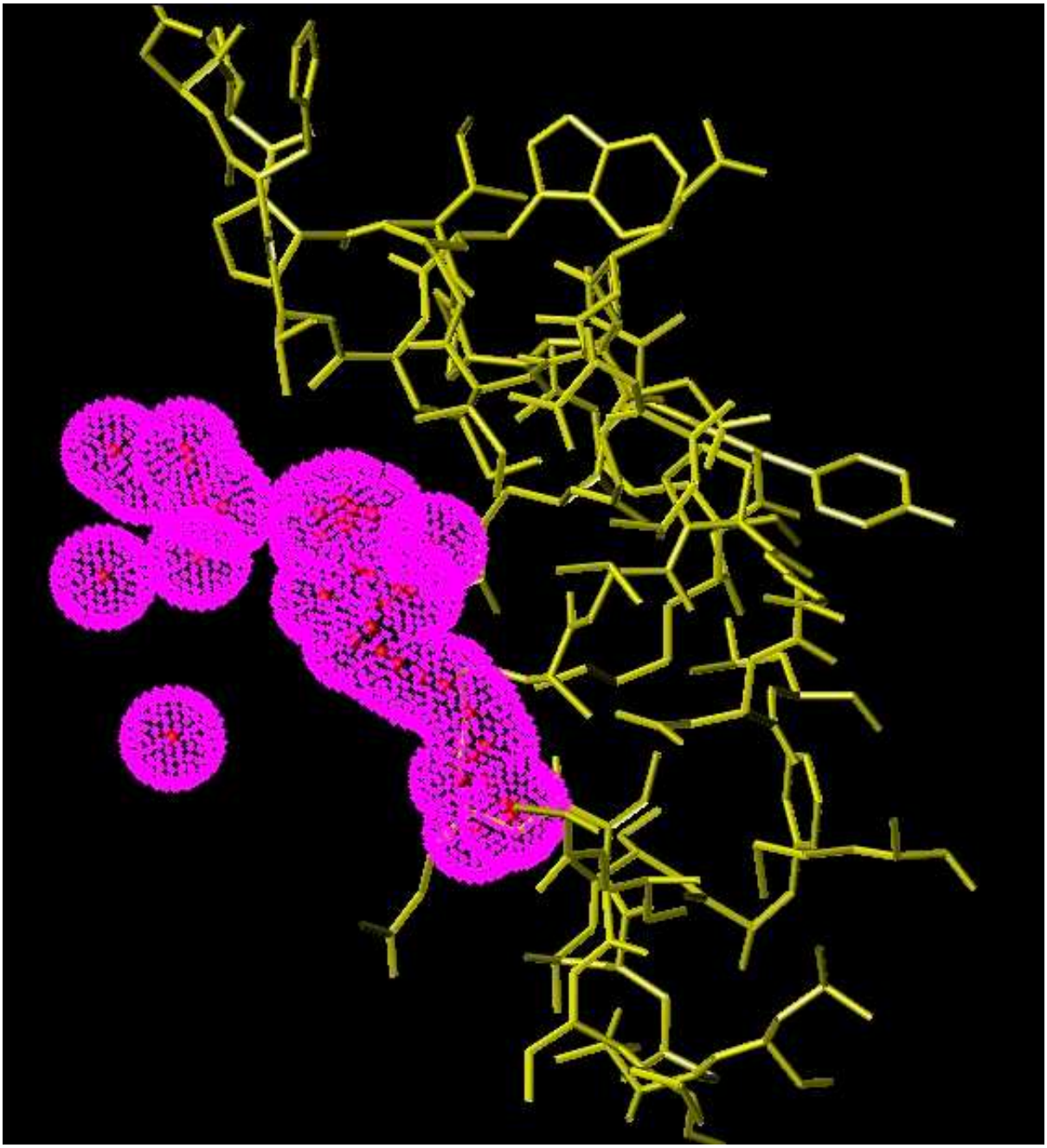


**Figure 3.3 - van der Waals radii encircling generated spheres within keep_exclude pocket. A single dot surfaces file represents the magenta spheres surrounding each red/ligand sphere.**

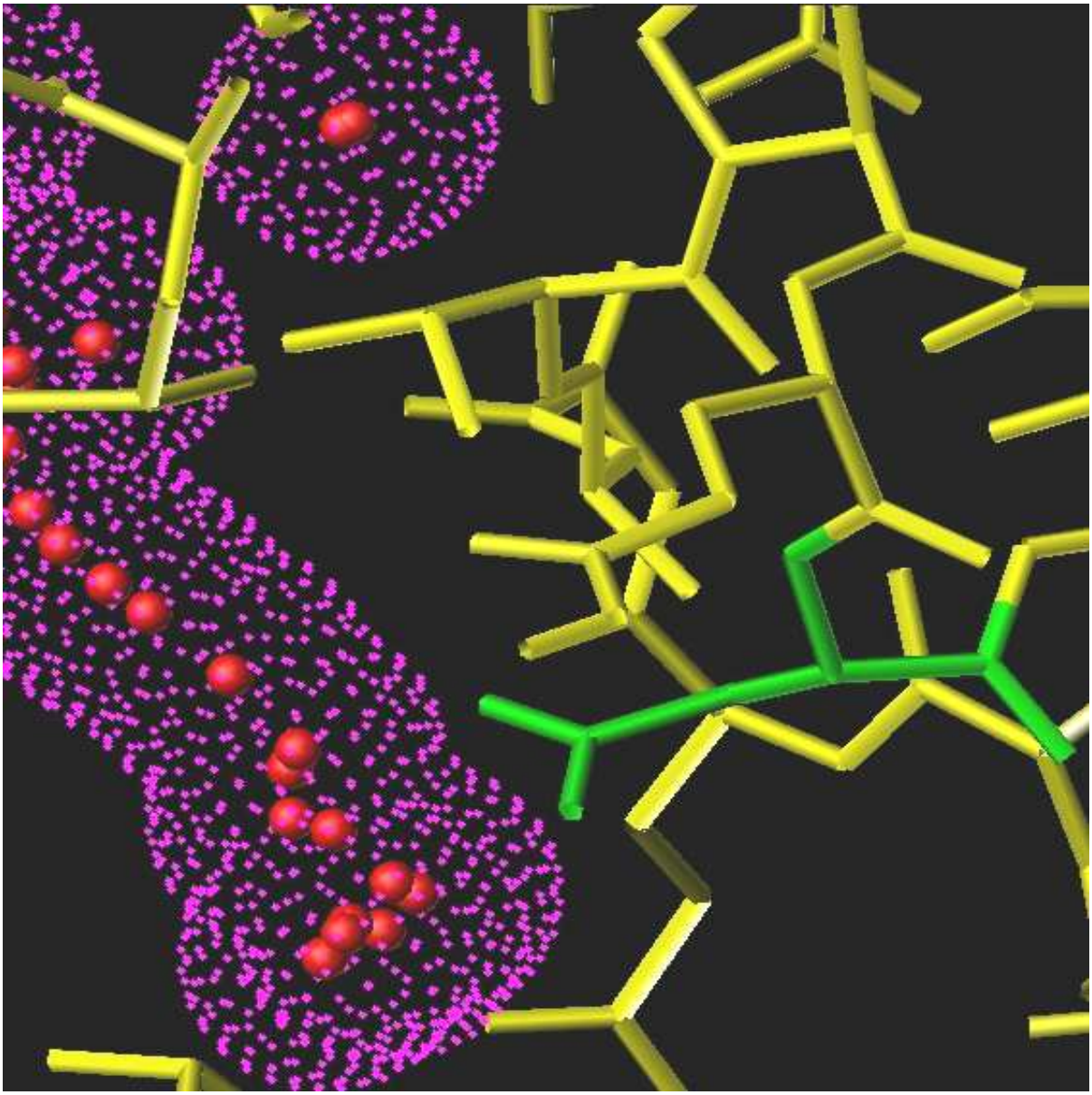


**Figure 3.4 – Key pocket components can be seen interacting as Asp51 (green) is in close proximity toVDW radii (magenta) of generated spheres (red).**

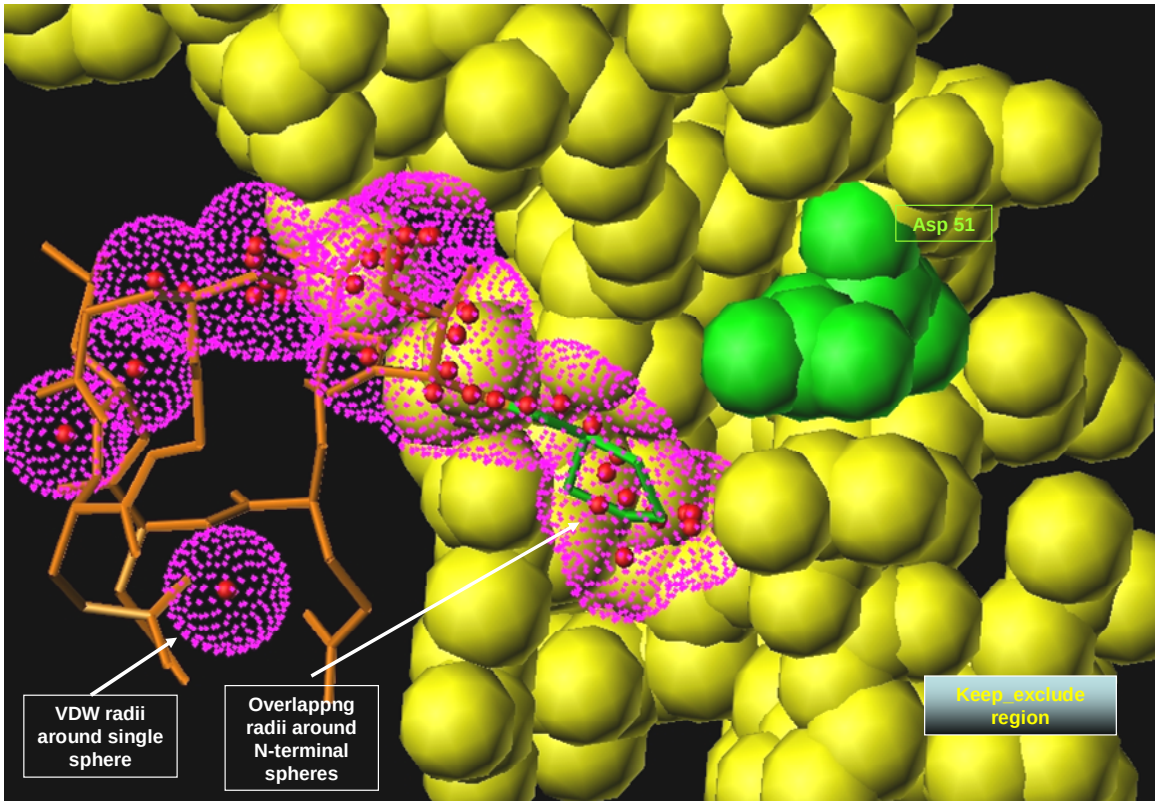



**Figure 3.5- Ligand, VDW radii (magenta), within Space-fill representation of Keep_exclude region**

### *Scoring grid calculation*

Showbox is an interactive sub-program that gives the user the capability to visualize the location and size of the grids that will be calculated by the program '*grid*'. The final program prompt is for the name of the output file: 'site_box.pdb'. The sub-program GRID creates the files that are required for high-speed evaluation/scoring in DOCK. The only type of scoring that was utilized in this project was the energy-based scoring (contact and chemical scoring are the others). Docking multiple ligands in multiple orientations over parallel servers, with each server distributing compounds to multiple clients requires supplemental programming. Details of Showbox, GRID and DOCK are available in appendix A.

# Chapter IV

## Results & Discussion

Upon execution of the docking protocol, listings of energy values were returned to the user. Figure 4.1 lists the energy readings for the top 5 compounds found per client for the Cambridge database on the left and the National Cancer Institute's (NCI) on the right. It is immediately noticeable that most of the superior scores are listed as analyzed by Clients 1 through 4. The relative speed at which the first 4 clients sort through the orientations of their server-assigned ligands become clearer as clients 6,7,8 do not receive as many molecules overall to score, and hence do not have many, if any of the top scoring compounds. The first four clients do not process molecules at any faster rate than clients 5 through 8 (all clients process the same CPU power as each SGI mainframe has two equal sub-processors). However, the server always begins at client number 1 and works sequentially through the clients, searching for the next idle client on which to dispatch the next database compound. The clients actually list the top 100 compounds found, and a grouping which exceeds the top 5 and displays the top 10 compounds found by servers 1 through 3 are seen in figure 4.2. An illustration of this client/server relationship is seen in Figure 4.3.

More frequently than not, one of the first 4 clients is once again available and open to analyze a ligand *before* the server reaches to clients higher than five. Hence, with the parameters set in these particular docking runs, the number of clients could have been set to six, or possibly five.

Client 1
1: -25.07 KOXBAA
2: -24.86 TMTFTC
3: -23.32 SETJAC
4: -21.67 DBNTHR02
5: -21.50 BUFNEV
Client 2
1: -24.81 SOCVIP
2: -22.40 BALNAD01
3: -22.36 TETTRI01
4: -21.58 VOWDOA
5: -21.45 PNPTCC01
Client 3
1: -25.18 GAYHUJ
2: -24.96 MTAZNI
3: -21.47 CUQUIN05
4: -20.27 VEXPIX
5: -20.20 MTHFPC10
Client 4
1: -22.98 GADMIH
2: -22.02 FAVPAT
3: -21.30 ANDREO
4: -19.52 TEMYUF
5: -19.28 CASRPP03
Client 5
1: -18.21 TAGUDN01
2: -17.94 YOVYUD
3: -17.79 GAZFAO
4: -17.51 BEBRIJ
5: -17.34 YABVIG
Client 6
1: -17.93 VUXTIR
2: -16.35 ZOZRAH
3: -16.20 YEGJOU
4: -16.12 TCHPOC
5: -15.93 PHENAZ02
Client 7
1: -16.79 RONDON
2: -15.15 HIBZTH
3: -15.01 DMANTR
4: -13.26 SAZGIJ
5: -12.23 WEMDEX
Client 8
1: -12.37 POTGIO
2: -12.04 TZCXHY
3: -11.37 SRFORM10
4: -10.89 TETBEZ
5: -10.87 TMETTS01

Client 1
1: -51.06 Str69897
2: -31.86 Str87814
3: -27.70 Str9620
4: -24.88 Str118402
5: -24.75 Str104122
Client 2
1: -28.84 Str29785
2: -24.50 Str34901
3: -24.19 Str34901
4: -23.72 Str89088
5: -23.56 Str72220
Client 3
1: -22.82 Str45269
2: -22.60 Str92429
3: -22.06 Str143780
4: -21.85 Str75476
5: -21.74 Str46068
Client 4
1: -23.86 Str54971
2: -21.70 Str114333
3: -21.63 Str114333
4: -20.27 Str63082
5: -20.13 Str112277
Client 5
1: -20.46 Str141315
2: -18.23 Str87650
3: -17.77 Str65488
4: -17.59 Str54771
5: -16.94 Str125850
Client 6
1: -16.75 Str138374
2: -15.77 Str89325
3: -14.86 Str126053
4: -14.56 Str126053
5: -14.49 Str81588
Client 7
1: -12.82 Str42860
2: -12.60 Str65345
3: -12.02 Str105993
4: -11.97 Str92668
5: -11.47 Str116163
Client 8
1: -12.13 Str68246
2: -11.37 Str130116
3: -11.26 Str143128
4: -11.08 Str68246
5: -11.02 Str114576

**Figure 4.1 Database results. Cambridge on left, NCI on right, Clients 1 through 8**

**Camb_Client_1**

```
Compounds read                    :      18263
Compounds docked                  :      18263
Compounds skipped                 :          0
Elapsed CPU time (sec)            :  106428.42
Time per docked compound (sec)    :       5.83

Current best energy scorers:
 1:  -25.07 KOXBAA
 2:  -24.86 TMTFTC
 3:  -23.32 SETJAC
 4:  -21.67 DBNTHR02
 5:  -21.50 BUFNEV
 6:  -20.89 VEXPIX
 7:  -20.70 GADMIH
 8:  -20.37 PAGQIX
 9:  -20.35 DBANQU
```

**Camb_Client_2**

```
Compounds read                    :      24144
Compounds docked                  :      24144
Compounds skipped                 :          0
Elapsed CPU time (sec)            :   97588.84
Time per docked compound (sec)    :       4.04

Current best energy scorers:
 1:  -24.81 SOCVIP
 2:  -22.40 BALNAD01
 3:  -22.36 TETTRI01
 4:  -21.58 VOWDOA
 5:  -21.45 PNPTCC01
 6:  -21.29 FAVPAT
 7:  -21.19 ANDREO
 8:  -20.21 SAOXPD
 9:  -19.70 OXFVOX
10:  -19.67 CALPCE
```

**Camb_Client_3**

```
Compounds read                    :      10934
Compounds docked                  :      10934
Compounds skipped                 :          0
Elapsed CPU time (sec)            :   41810.89
Time per docked compound (sec)    :       3.82
Current best energy scorers:
 1:  -25.18 GAYHUJ
 2:  -24.96 MTAZNI
 3:  -21.47 CUQUIN05
 4:  -20.27 BIXBIT03
 5:  -20.20 MTHFPC10
 6:  -19.21 FURALA10
 7:  -19.06 FEYZUE
 8:  -19.04 BIXBIT03
 9:  -18.91 KASJAP
10:  -18.77 BAVFEJ
```

**Figure 4.2- The top 10 compounds of Clients 1-3**

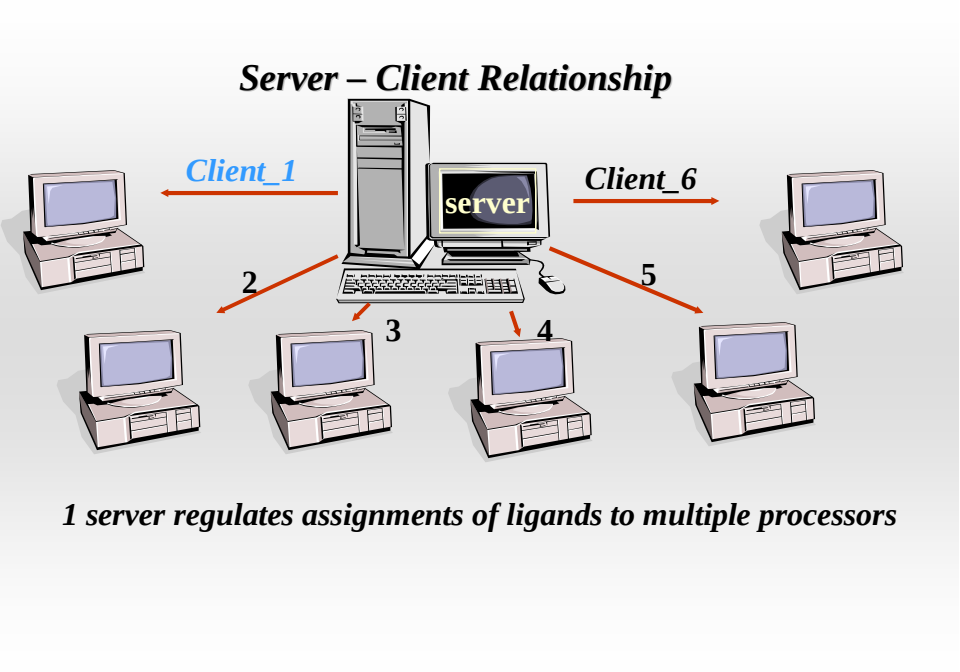



**Figure 4.3 – The server / client relationship**

Output files for single ligand runs (Figure 4.4) or database runs (Figure 4.5) are specified in input decks and contain the suffix '.info'. Each database run utilized a master input deck which specifies that 100 of the top ligand configurations be returned to this '.info' file. Once these molecules were obtained, the top 40 compounds (from each database respectively) were re-evaluated and re-docked individually with more CPU time being allotted for exploring a more generous (500) number of orientations per compound. It was found that re-docked compounds returned better energy scores once given more range of flexibility within the pocket. Example: NCI compound Str34901 (shown in Figure 4.4 as having a score of -27.12) originally achieved energy readings of -24.50 and -24.19 even with two conformations being scored in the top five of client 2. The initial database run was then a pharmacophore screening process to achieve the necessary chemical components of a high-scoring ligand. Subsequent analysis of the top 40 compounds fit better into the category of shape complimentarity as the torsional flexibility of

each compounds was explored. The top 15 compounds are illustrated in this chapter within the space-fill, N-terminal pocket.

```
_________________________Docking_Results_____________________________
Name         : Str34901
Description  : ****
Orientations tried                                  :       4763
Orientations scored                                 :        500

Best intermolecular energy score                    :     -27.12
RMSD of best energy scorer (A)                      :     119.14

Elapsed cpu time (sec)                              :       0.97


Writing restart information to disk.
Writing top scoring molecules to disk.

Finished processing molecule in 1.01 seconds.
```

**Figure 4.4 – '.info' file showing the single docking output of NCI molecule Str34901**

```
Compounds read                      :     181343
Compounds docked                    :     181343
Compounds skipped                   :          0
Elapsed CPU time (sec)              :   86372.23
Time per docked compound (sec)      :       0.48

Current best energy scorers:

 1:  -26.93 Str1732
 2:  -24.40 Str576
 3:  -23.35 Str1373
 4:  -23.31 Str1372
 5:  -23.15 Str726
 6:  -23.10 Str726
 7:  -22.97 Str1788
 8:  -22.87 Str1788
 9:  -22.44 Str626
10:  -22.36 Str3083
11:  -22.08 Str555
12:  -22.07 Str2691
13:  -21.86 Str1262
14:  -21.85 Str2076
15:  -21.62 Str2532
```

**Figure 4.5 - '.info' file listing the top 15 molecules from a database run**

Compounds emulating the size, torsional flexibility and chemical properties of the original 10-residue ligand within the capsid N-terminal pocket are examined in the next several pages. Again, the primary interactions are not just an amino group interacting with the carboxyl of Asp51 but: 1) isoleucine side chain packing into a hydrophobic binding site, 2) hydrogen bonding between the Pro1 & Gln13 main chain oxygen, 3) van der Waals contacts between Pro1 ring and alpha carbon atoms of Ile15 and Gly46 and 4) hydrogen bonding between the two strands of the B-hairpin itself and were collectively illustrated in chapter one (figures 1.6-1.8).

NCI Screening: Analysis by virtual screening is obviously not solely sufficient as the best overall score of -51.06 by Str69897 (Figure 4.6B) may not have the desired size to fully deter the N-terminal from forming. One of its amino groups is readily available and located in a para position. Each of these amino groups is capable of interacting with the carboxyl of Asp51 and the oxygen of Gln13 and the ring group is capable of interacting with the side-chains of Ile15 and Gly46. Dock placed this molecule high up in the pocket and configured it so that the ring faces *away* from the D51 complex (see figure 4.6A). Despite the size and configuration drawbacks of this molecule, its top score is the result of excellent energy interactions through its chemical makeup.

Other top scores include the -31.06 of Str87814 (Figure 4.7A) and the -28.84 of Str29785 (Figure 4.8A). Both of these compounds include two rings and the requisite amount of size and torsional flexibility to fit into the pocket in a number of configurations. Differences lie in the groups on the flexible regions. The lower (and more favorable) energy reading assigned to Str87814(Figure 4.7B) may be largely due to its two amino groups, while Str29785 (Figure 4.8B) has one amino, one oxygen, which DOCK 4.0 appropriately placed away from D51. The

lack of amino groups directly on either ring of either compound is detrimental for interactions with D51 and Q13. Side-chain reactions, while less important, are still maintained with each ring group. Figure 4.9A shows Str9620 with a sulfur atom being placed in front of D51 with an amino group in the nearby vicinity. The molecule (Figure 4.9B) has the requisite size, placement within the pocket by DOCK and ring groups for side chain as well as van der Waals' interactions hence, its good score of -27.70.

High scoring compounds which have the appropriate chemical groupings, yet appear somewhat bulky and non-flexible included Str104122 (Figure 4.10B) and Str34901 (Figure 4.11B). The five-member ring compound Str104122 has torsional flexibility, with its ring groups on either end, neither containing an amino group for electron acceptance from D51. The amino groups which are present on Str104122 were placed by DOCK in such a manner than neither ring group could sufficiently interact with Ile15 and Gln46 simultaneously, although DOCK did an excellent job of pointing the oxygen away from the D51 carboxyl group. Figure 4.11A shows the orientation of Str34901 in the pocket. The five rings of the compound give a wide variety of interactions throughout the pocket, including Ile15 and Gly46, however the location of the two amino groups leaves neither with a direct association with D51 or Q13.

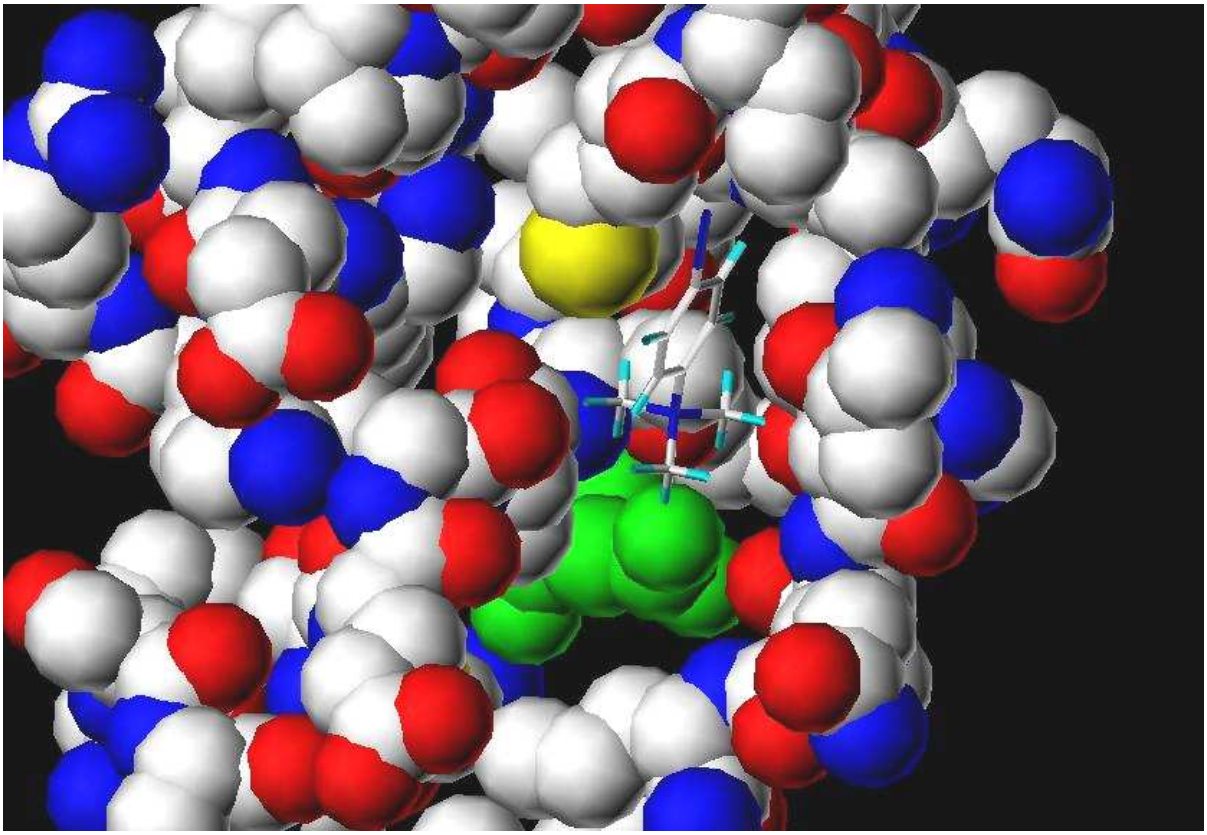


**Figure 4.6A Str69897, score -51.06, despite the size drawback of this molecule, its top score is the result of excellent energy interactions through its chemical makeup**

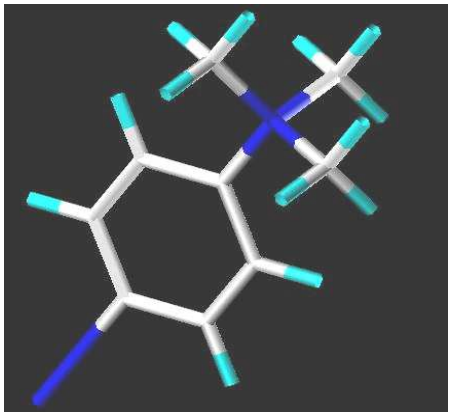


**Figure 4.6B Str69897 from NCI**

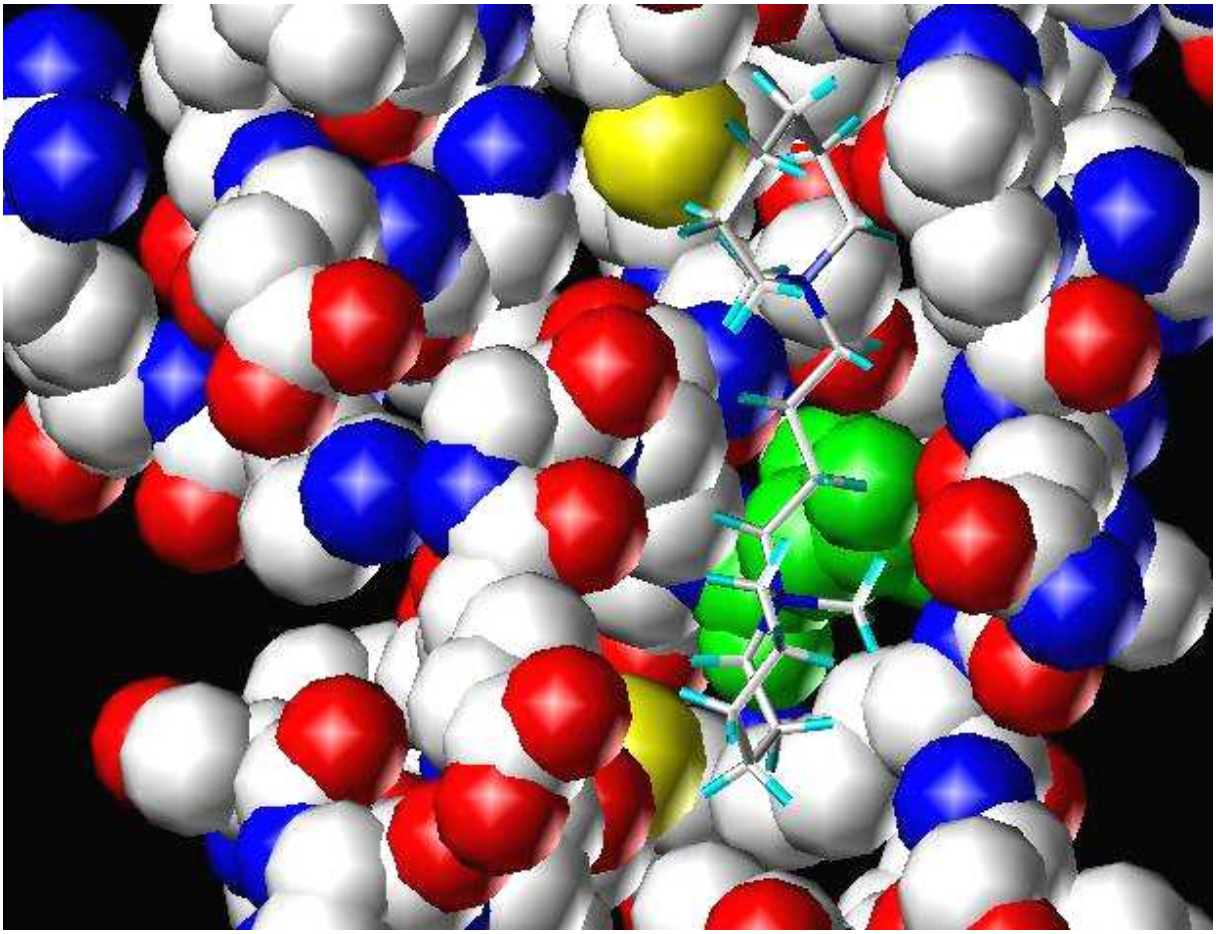


**Figure 4.7A Str87814, score -31.06**, t**he favorable energy reading assigned to Str87814 may be largely due to its two amino groups.**

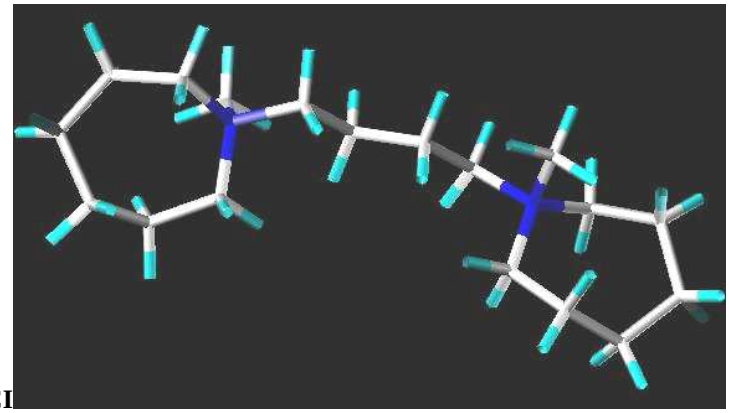


**Figure 4.7B Str87814 from NCI**

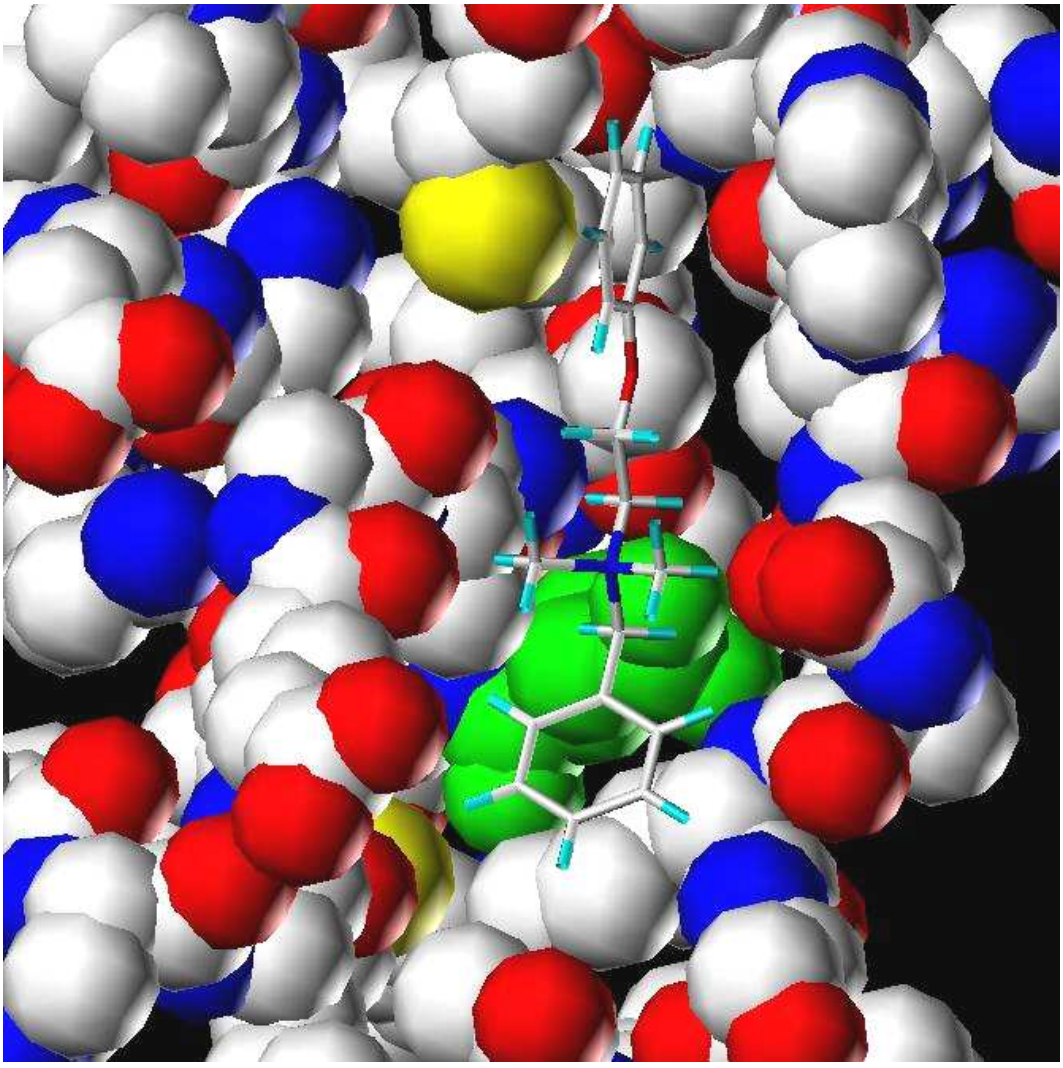


**Figure 4.8A Str29785 - score -28.84, has one amino group and an oxygen which DOCK 4.0 appropriately placed away from D51 (in green)**

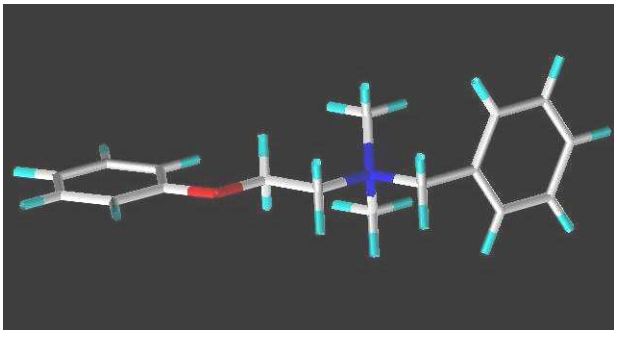


**Figure 4.8B Str29785 from NCI**

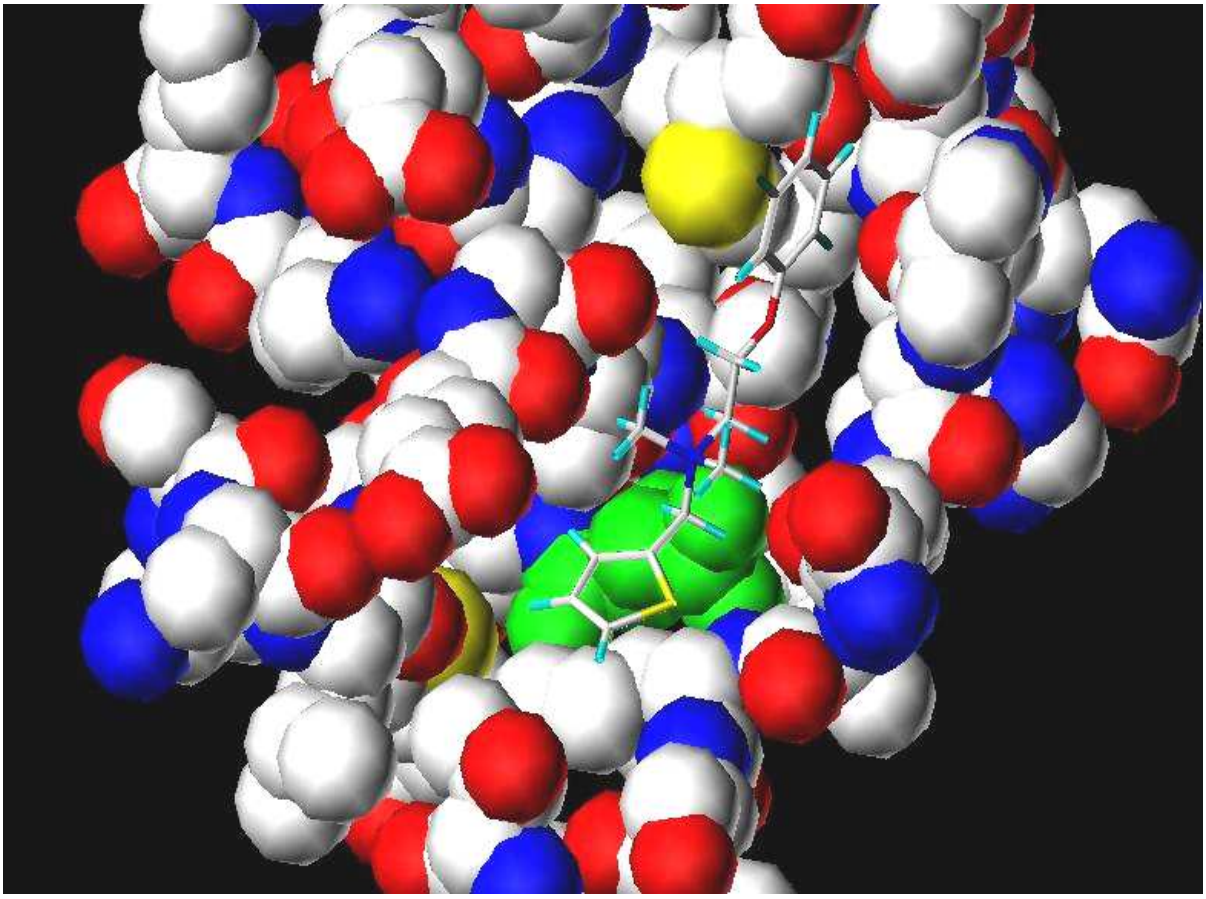


**Figure 4.9A Str9620, score -27.70, with a sulfur atom being placed in front of D51 with an amino group in the nearby vicinity. The molecule has the requisite size, placement within the pocket by DOCK and benzyl groups for side chain as well as van der Waals' interactions.**

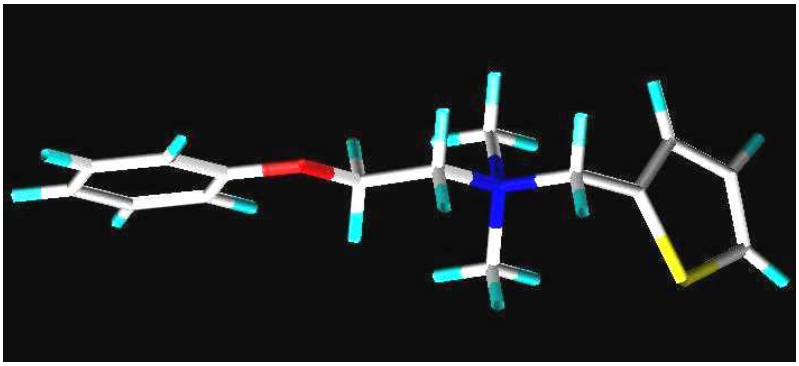


**Figure 4.9B Str9620 from NCI database**

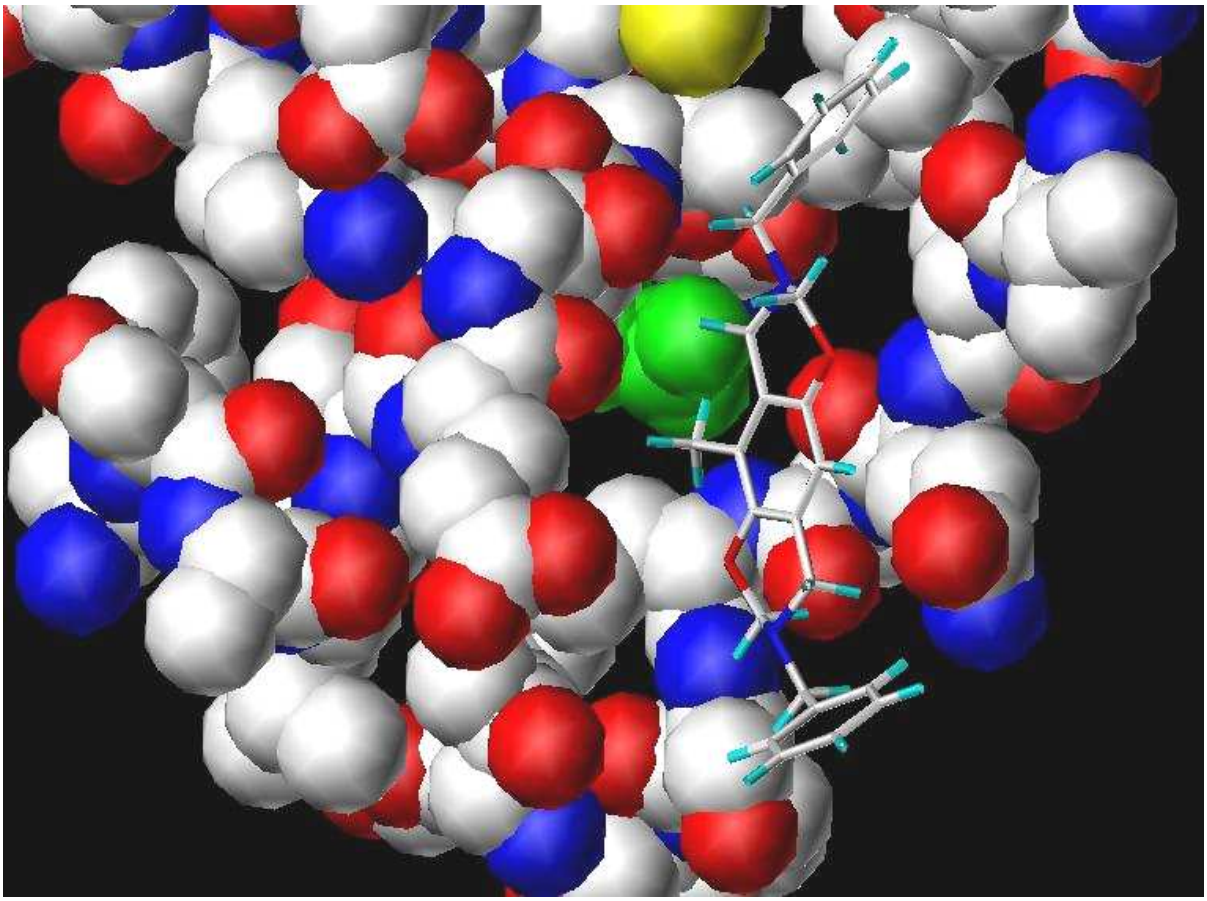


**Figure 4.10A Str104122, score -24.75, has torsional flexibility, with its ring groups on either end, neither containing an amino group for electron acceptance from D51 (in green).**

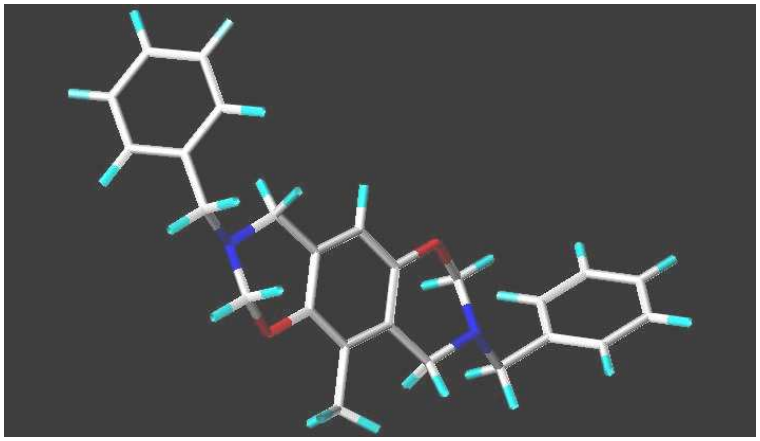


**Figure 4.10B Str104122 from NCI**

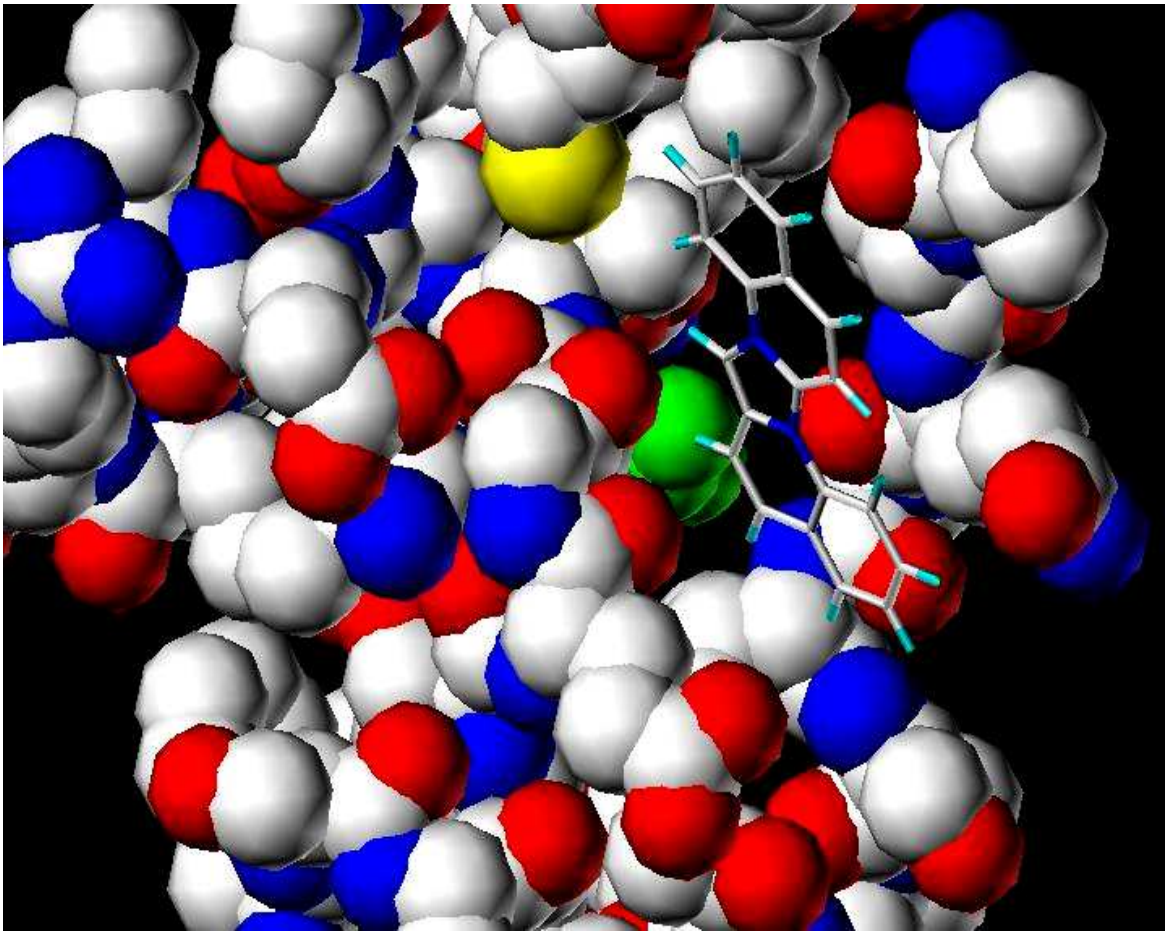


**Figure 4.11A Str34901 score -24.50 the compound has a wide variety of interactions throughout the pocket, however the location of the two amino groups leaves neither with a direct association with D51 (in green) or Q13.**

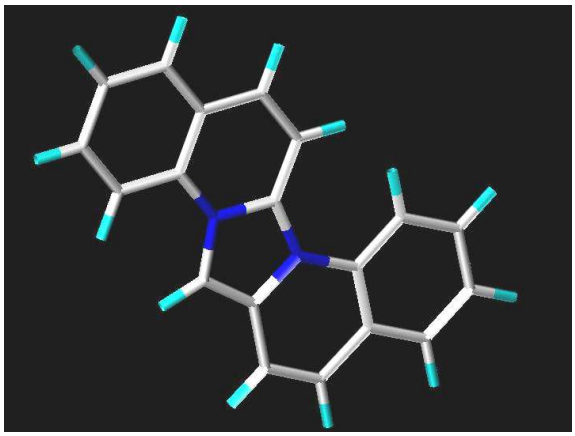


**Figure 4.11B - Str34901 from NCI**

Other compounds in the National Cancer Institute top 15 include compounds that have obvious and immediately detectable drawbacks in either shape, chemical composition or both and their features should not be included in forming pharmacophore assessments of lead compounds. The size and configuration of Str118402 (figure 4.12A scoring -24.88) made it difficult for DOCK to place a ring group and amino group in proximity to D51. While DOCK did accomplish this task, the multiple electronegative chlorine groups on the molecule add unwanted chemical interactions to an already awkward binding inside this pocket. Figure 4.13B shows Str54971, the only compound in the top 30 (-23.86) with absolutely no ring groups. Nitrogen, sulfur and oxygen groups are available and complementary to many wanted pocket interactions; however the lack of ring group dispels any notion of van der Waals' interactions with Ile15 or Gly46. Conversely, Str89088 and Str72220 have the size shape and ring groups to be ideal candidates but lack any complementary chemical features (Figures 4.14A & 4.15A). Str143780, Str75476 (Figures 4.17A, 4.18A)and Str92429 all have excellent shape complementarity and amino groups; however their scores (while still being top 15) are restricted by the side oxygen groups. Str92429 (Figure 4.16B) has the distinct disadvantage of having an oxygen in close proximity to D51.

Str45269 rounds out the Top 15 of the NCI database scores. This compound's bulky awkwardness is only supplemented by having a multiple array of ring groups for interactions with Gly46 and Ile15. DOCK has a difficult time placing and amino group near the D51 carboxyl and avoiding the electronegative chlorine from wreaking havoc deep within the pocket. Figure 4.19A shows how DOCK places the chorine away from the pocket and attempts to adhere to other pocket requirements.

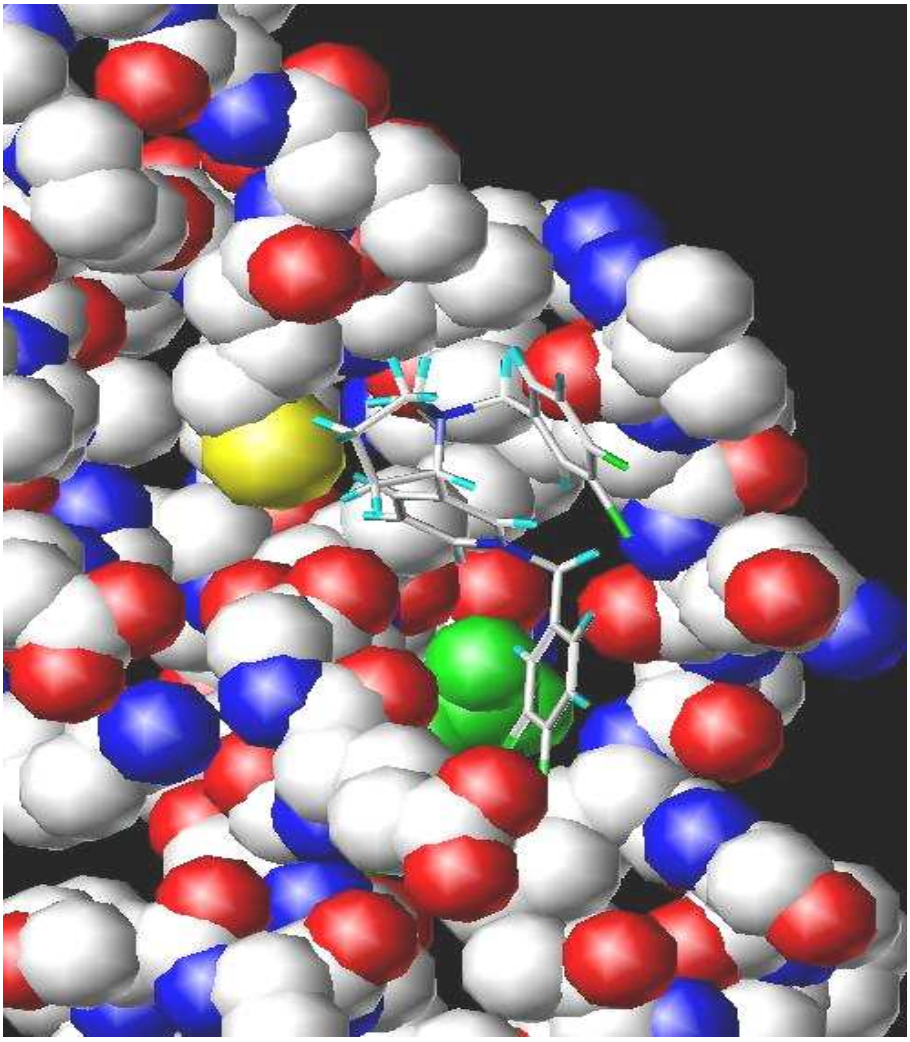


**Figure 4.12A Str118402, score -24.88 The size and configuration made it difficult for DOCK to place a ring group and amino group in proximity to D51. The multiple electronegative chlorine groups on the molecule add unwanted chemical interactions to an already awkward binding inside this pocket.**

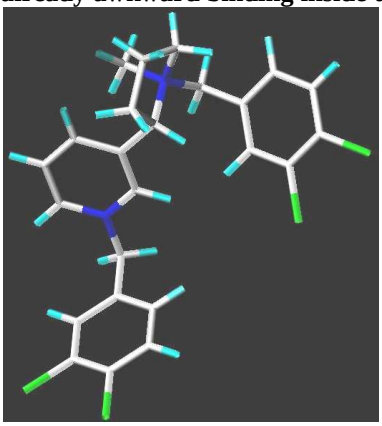


**Figure 4.12B Str118402 from NCI**

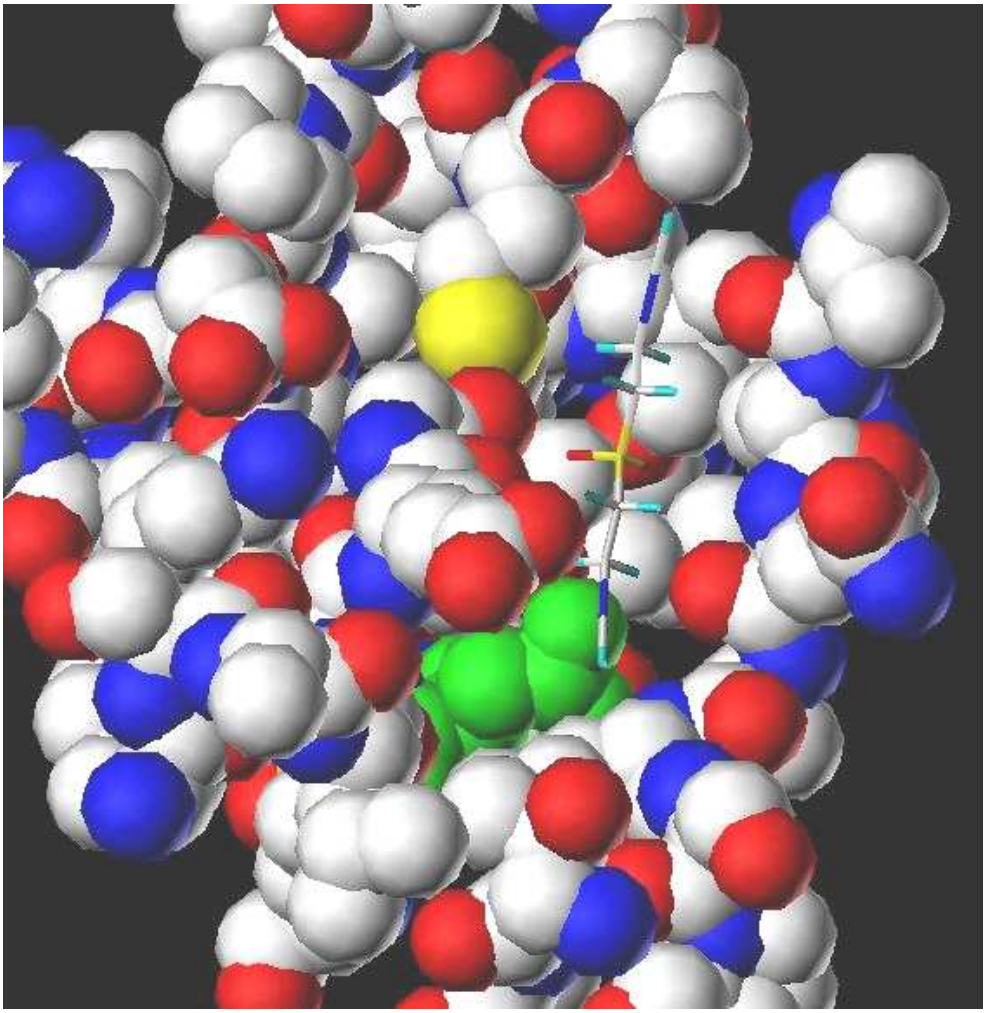


**Figure 4.13A Str54971- score -23.86, the only compound in the top 30 with absolutely no ring groups. Nitrogen, sulfur and oxygen groups are available and complementary to many wanted pocket interactions; the lack of ring group dispels any notion of van der Waals' interactions with Ile15 or Gly46.**

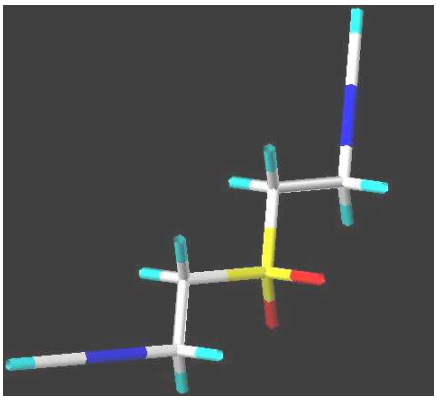


**Figure 4.13B Str54971 from NCI**

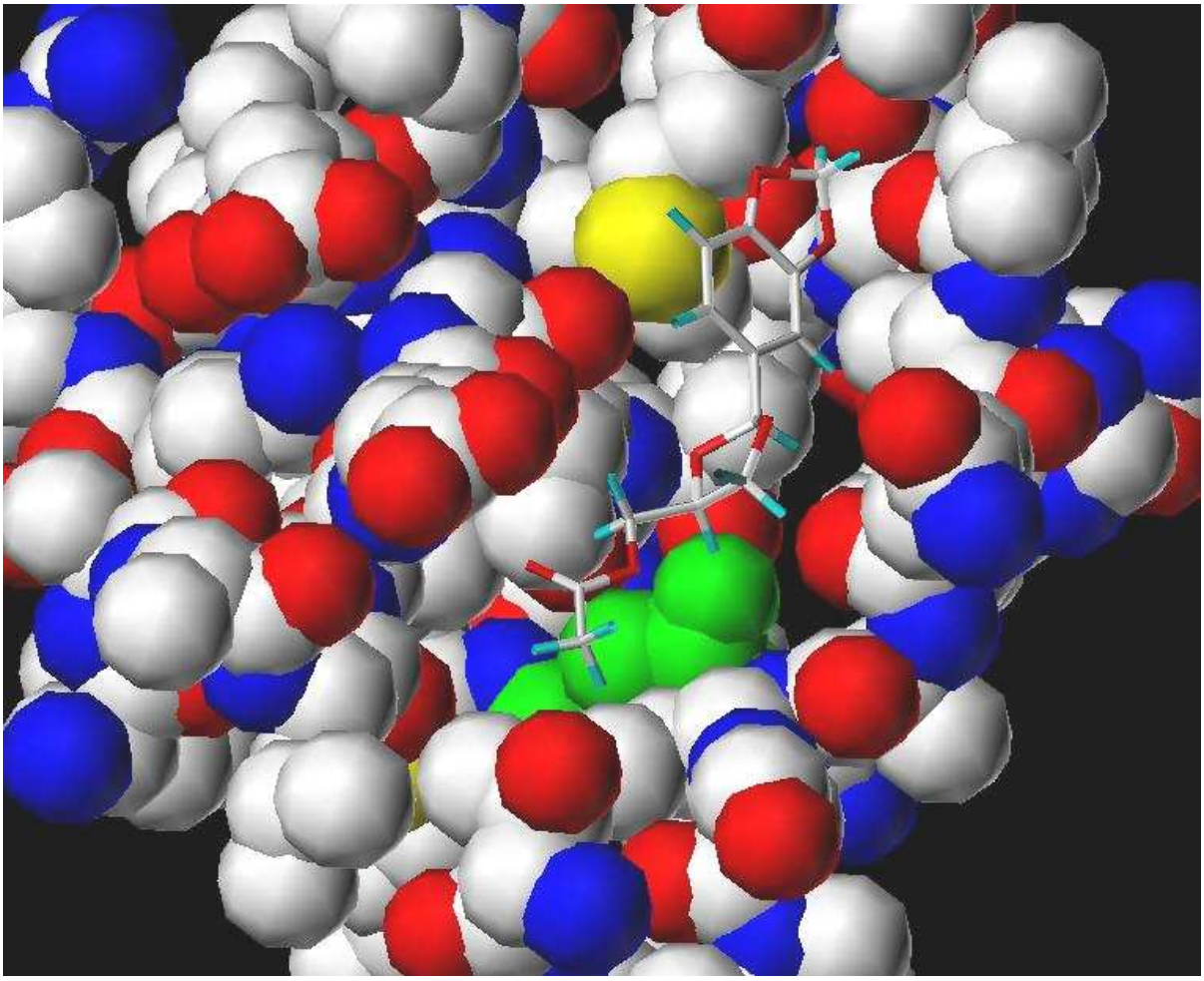


**Figure 4.14A Str89088 – score: -23.72, size shape and ring groups to be ideal candidate but lacks any complementary chemical features**

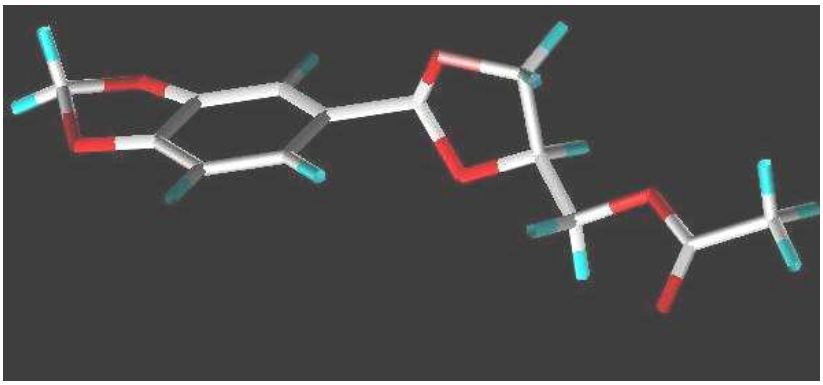


**Figure 4.14B Str89088 from NCI**

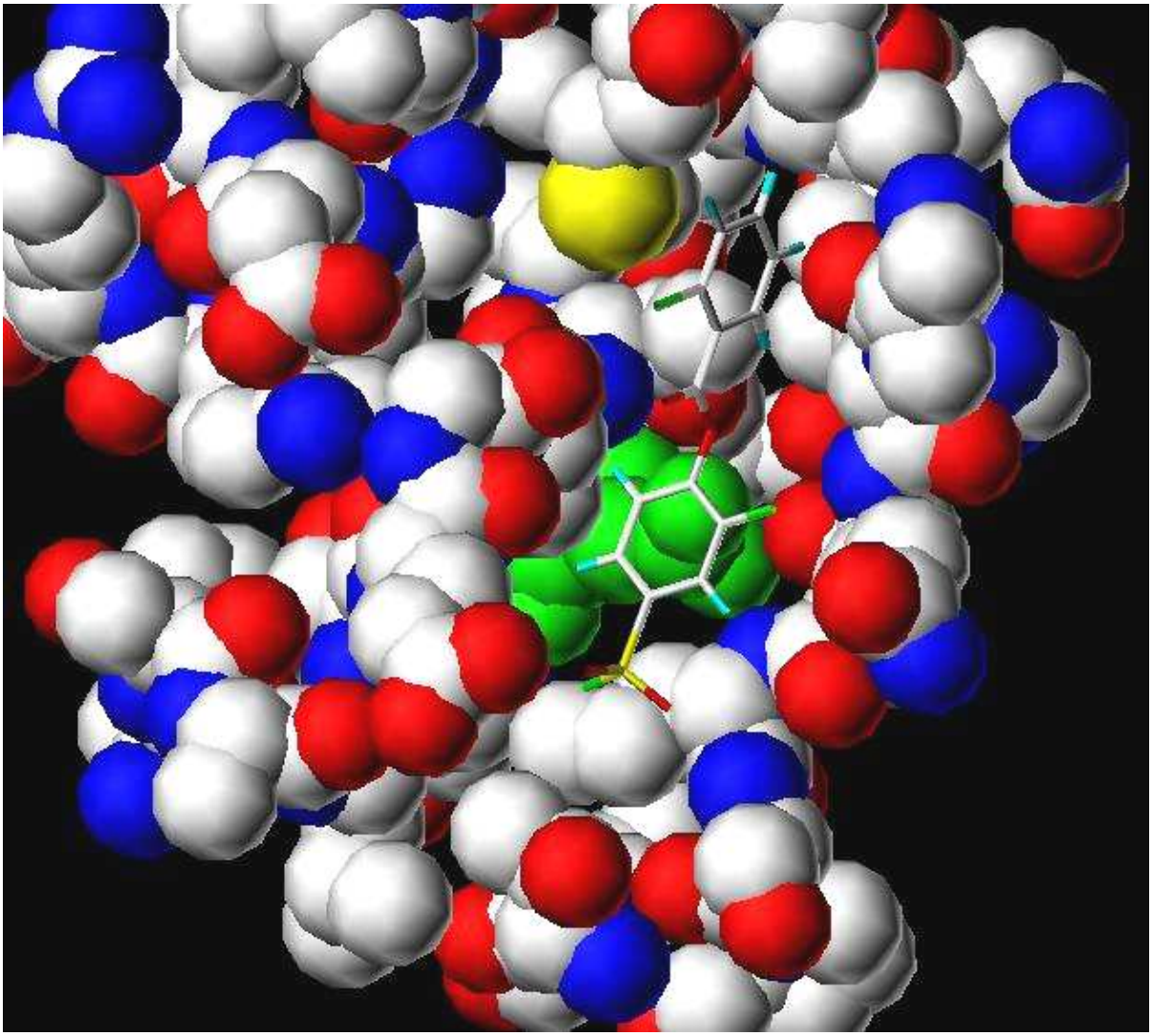


**Figure 4.15A Str72220: score -23.56 size, shape and benzyl groups to be ideal candidate but lacks any complementary chemical features**

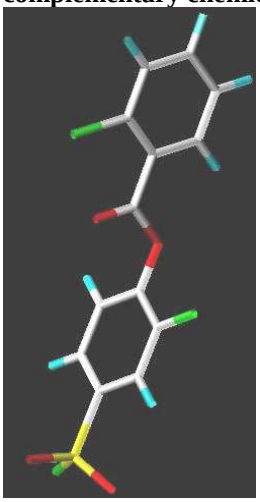


**Figure 4.15B Str72220 from NCI**

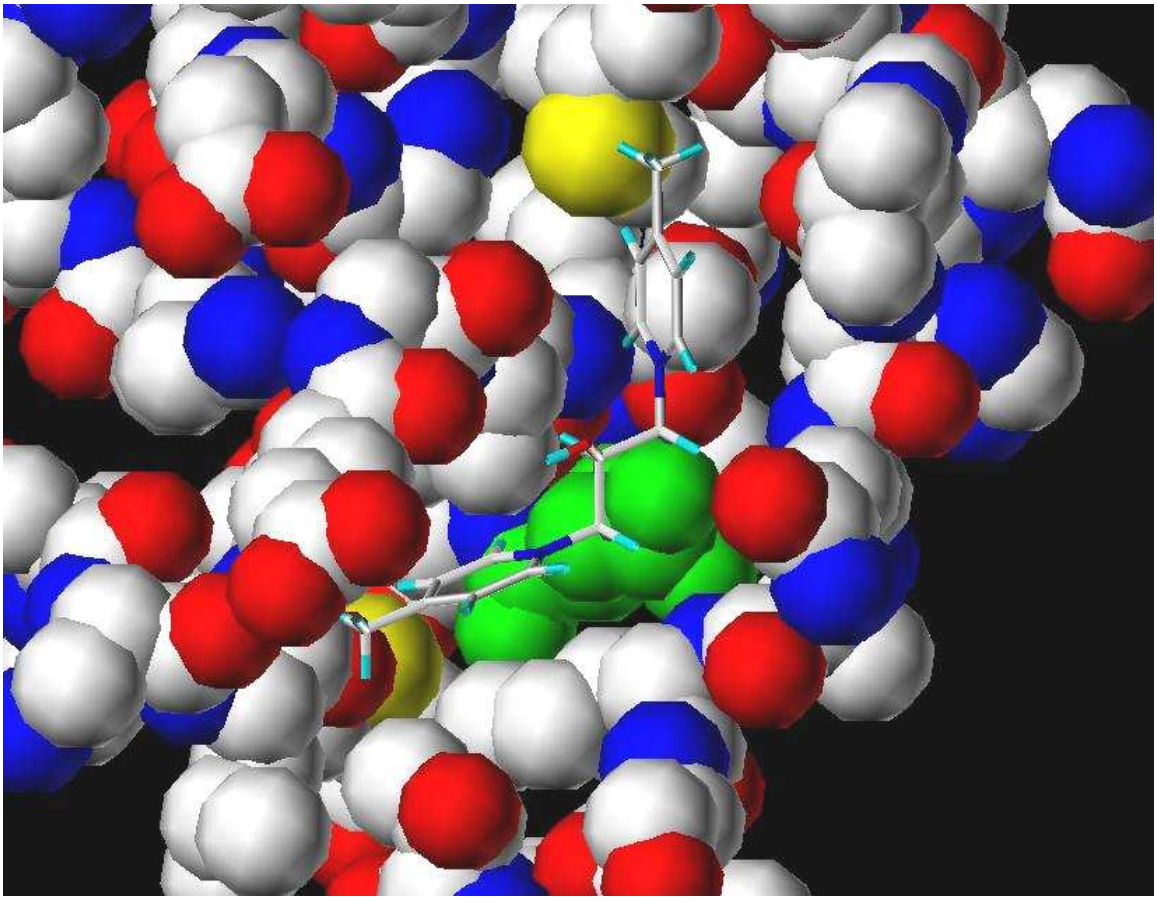


**Figure 4.16A Str92429: score -22.60 restricted by the side oxygen groups - has the distinct disadvantage of having oxygen in close proximity to D51.**

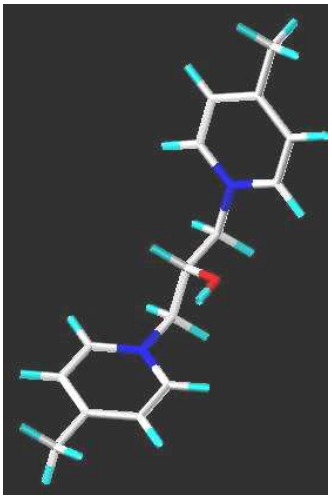


**Figure 4.16B Str92429 from NCI**

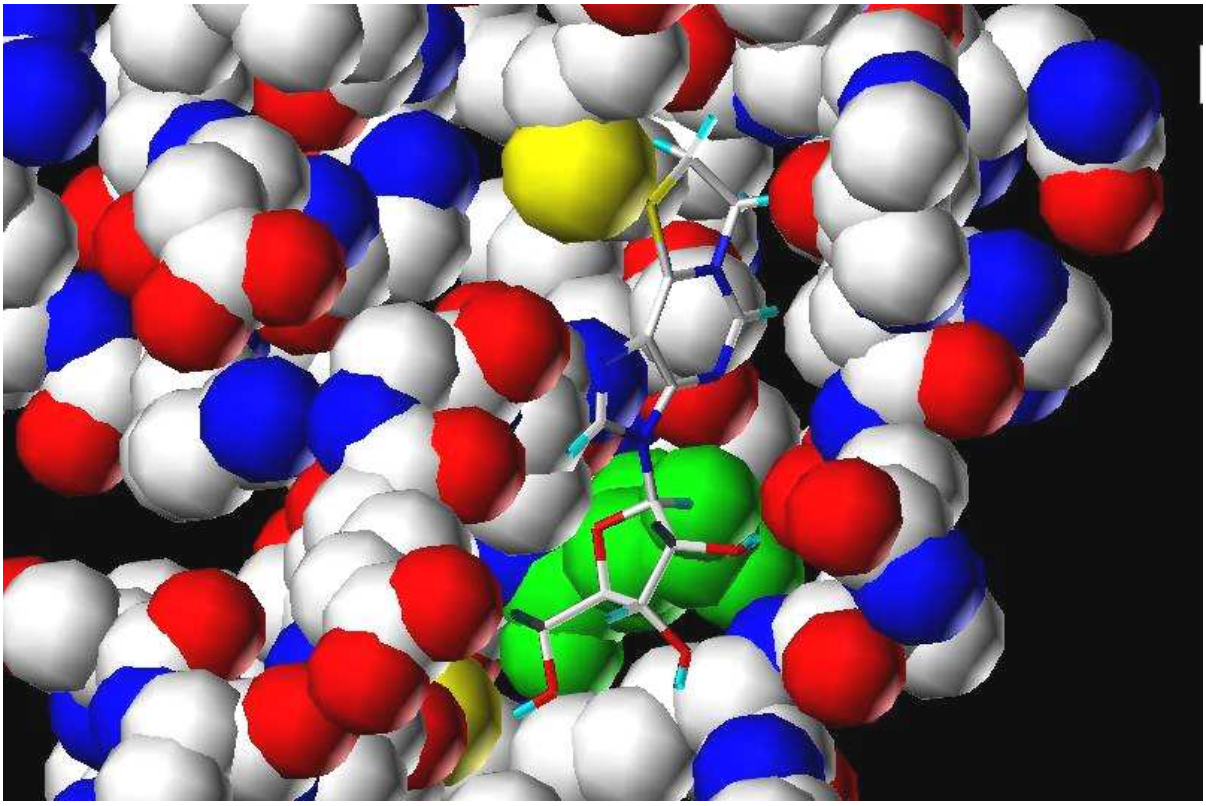


**Figure 4.17A Str143780: score -22.06**
**Hampered from lower energy score by the side oxygen groups. DOCK does not places amino groups in close proximity to D51**

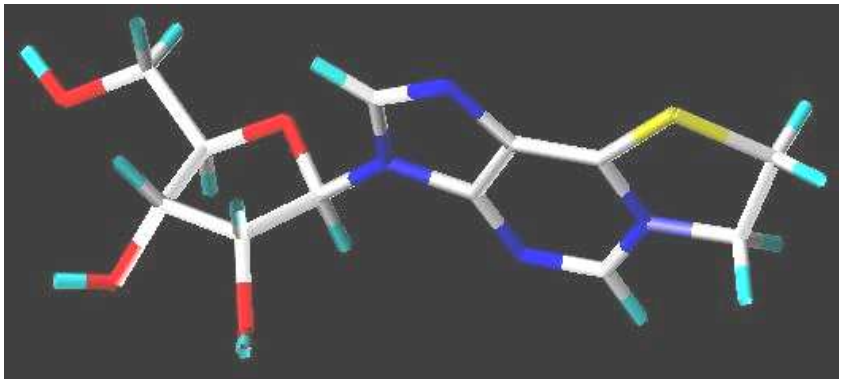


**Figure 4.17B Str143780, NCI**

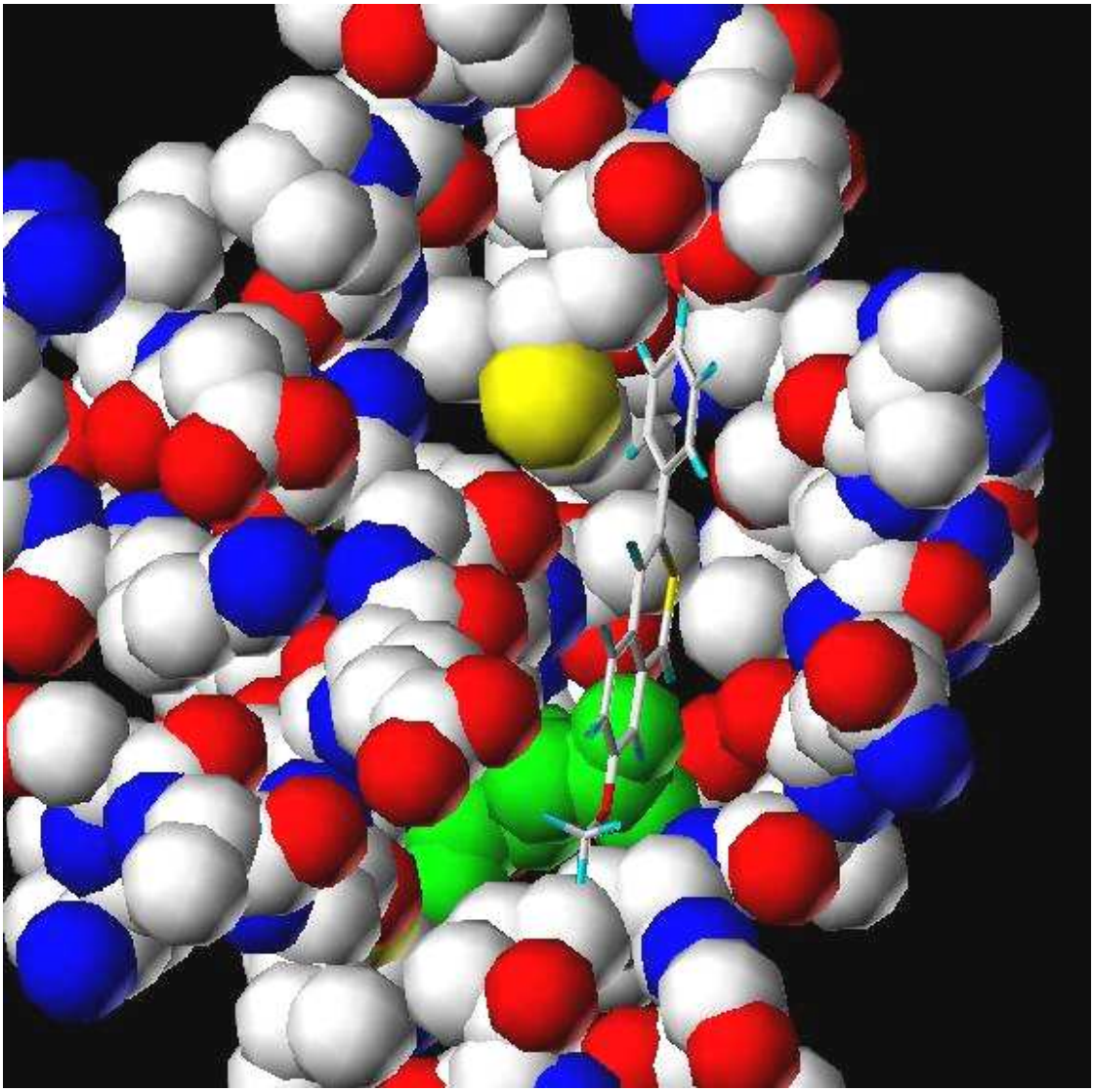


**Figure 4.18A Str75476 score -21.85 restricted from lower energy score by side oxygen groups**

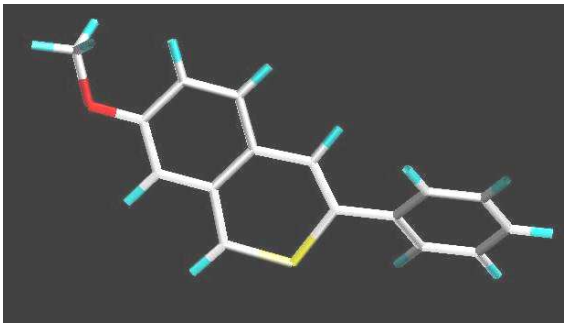


**Figure 4.18B Str75476 from NCI**

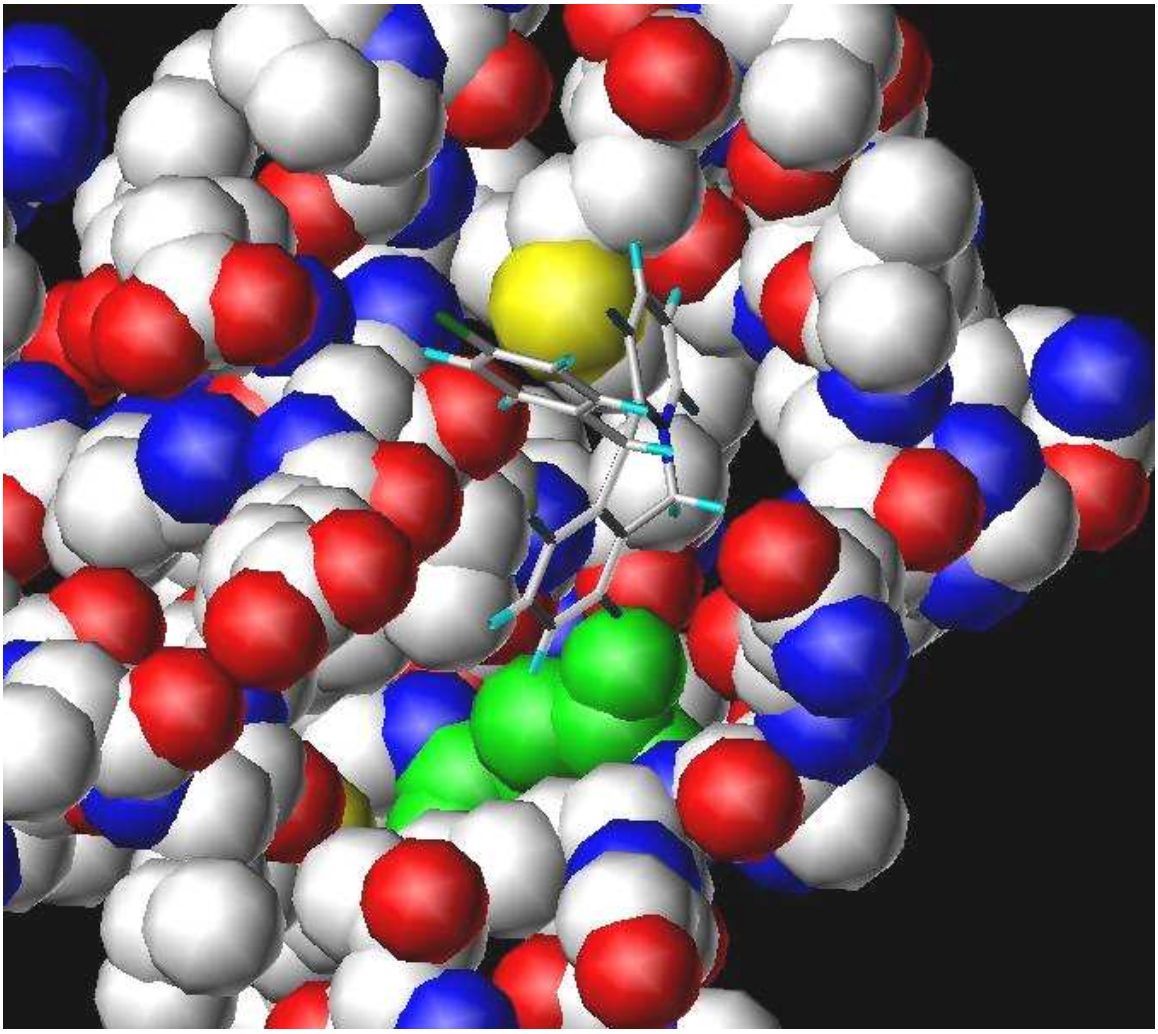


**Figure 4.19A Str45269: score -22.82, bulky awkwardness is only supplemented by having a multiple array of benzene groups for interactions with Gly46 and Ile15.**

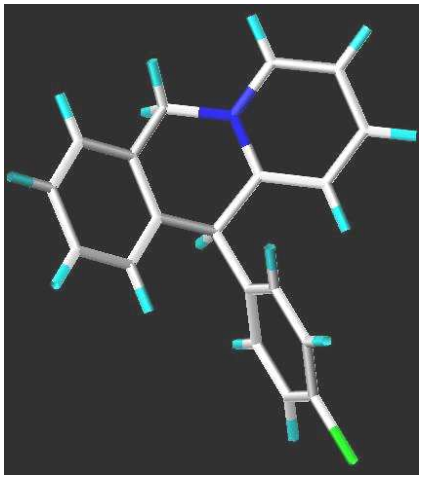


**Figure 4.19B Str45269 from NCI**

Cambridge U. screening: Figure 4.20A shows the molecule KOXBAA within the capsid pocket. This distorted molecule does have a ring amino group and would be a good compound if it did not have noticeable steric hindrance to fit the remainder of the molecule into the pocket. DOCK does not place the ring group in proximity of D51, Gly46 or Ile15. This type of bulkiness is also detected with the compound SOCVIP (figure 4.21B shown with solvent molecule) which does not have visually obvious interactions with any of the key residues or side-chains. SETJAC (figure 4.22B, shown with solvent molecule) has excellent ring groups and size distribution and is in line with the multi-ring, sulfur pharmacophore found frequently within the Cambridge database top scores. TMTFTC (figure 4.23B), BALNAD01 (figure 4.24B), TETTRI01 (figure 4.25B) and GAYHUJ (figure 4.26B) all posses multi- ring, sulfur properties. The ring groups sufficiently fill the pocket and are capable of interactions with the aforementioned key residues.

BUFNEV (figure 4.27B shown with solvent molecule) would be an excellent choice if it contained an amino on its ring group. In addition, DOCK places the single nitrogen of the compound in a position where it is unable to readily interact with D51 (figure 4.27A). Contrasting with BUFNEV, the ligand VEXPIX (figure 4.28A) has two amino groups, and one directly on a ring. However in an attempt to accommodate the chemical fit of amino to carboxyl, DOCK has to project a ring group away from the pocket to avoid steric hindrance, thus disturbing the other residue reactions with the amino ring group of the ligand (figure 4.28). PNPTCC01 (Figure 4.29B) and MTAZNI (Figure 4.30B) both are nickel-based compounds. While MTAZNI may be an excellent steric fit, compounds containing copper, nickel and vanadyl are far and few between, with only early stage results reported in the literature. Compounds containing these chemical features have been reported in blocking the gp120-CD4 entry,

however more testing is foreseen to determine if *and* how these compounds will excel at rendering HIV non-infectious (Vzorov *et al.,* 2003). Regardless of HIV-1 nickel studies, PNPTCC01 (figure 4.29A) is only a simple ring structure, incapable of occupying the entire pocket and maintaining simultaneous interactions with residues of importance. DBNTHR02 is a simple carbohydrate molecule (4.31B), with no discernable chemical properties to assist in binding within the pocket, although its pocket-filling size is palpable. VOWDOA (figure 4.32B) presents multiple amino and ring groups for interaction.

DOCK attempted to place the ring group of GADMIH (figure 4.33A) in the vicinity of the original ligand's N-terminal proline. Ignoring the amino groups present on the compound, DOCK shows its proclivity for steric fit, rather than chemical compatibility. This tendency is also seen in the placement of FAVPAT (figure 4.34A), and ANDREO (figure 4.35A). FAVPAT has conformational possibilities, but the placement of two oxygens in close range of D51 is questionable. ANDREO has been given a similar docking fate, with its non-ring oxygen having the same placement near critical residues.

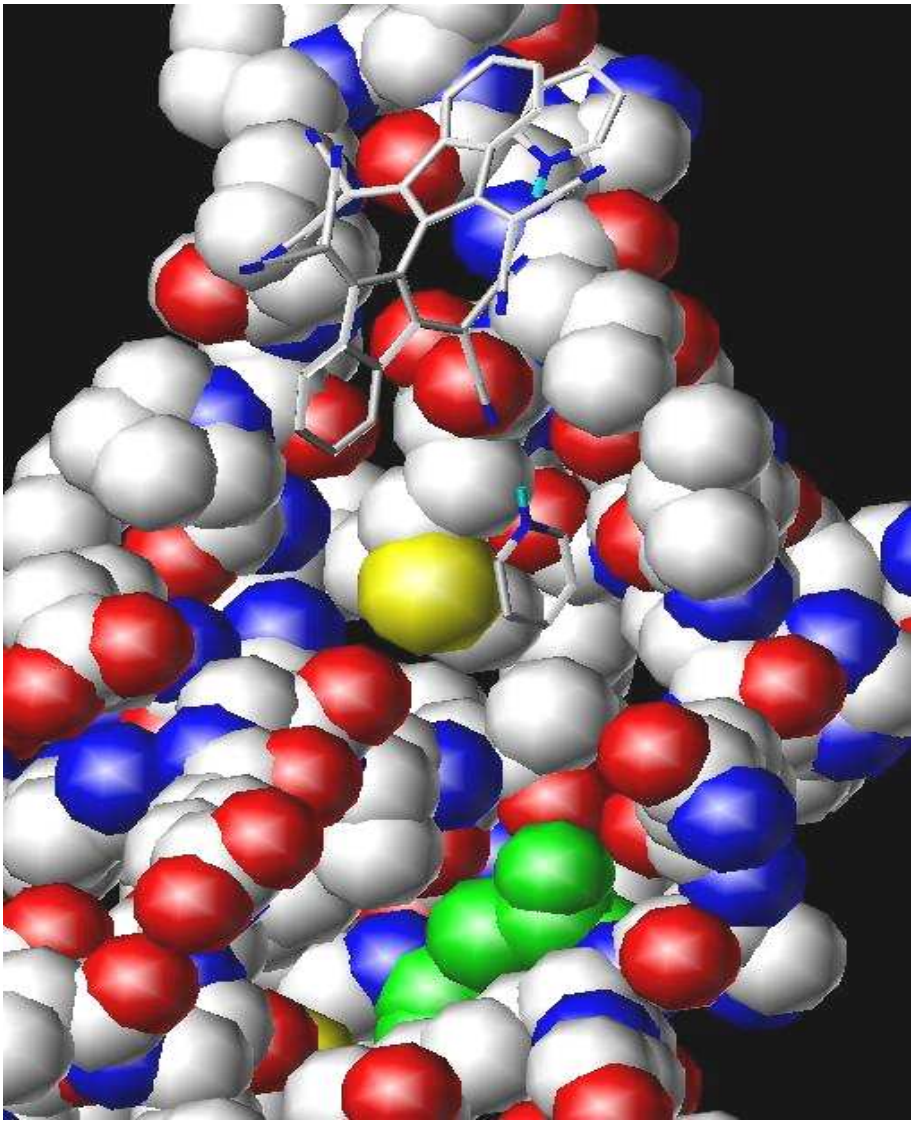


**Figure 4.20A KOXBAA score -25.77 noticeable steric hindrance makes awkward fit of molecule into the pocket**

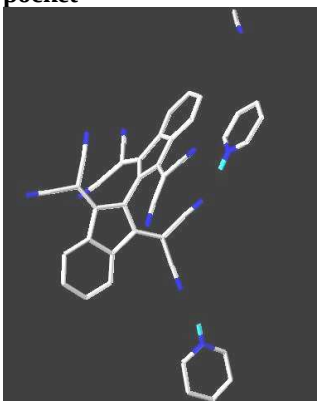


**Figure 4.20BKOXBAA from Cambridge**

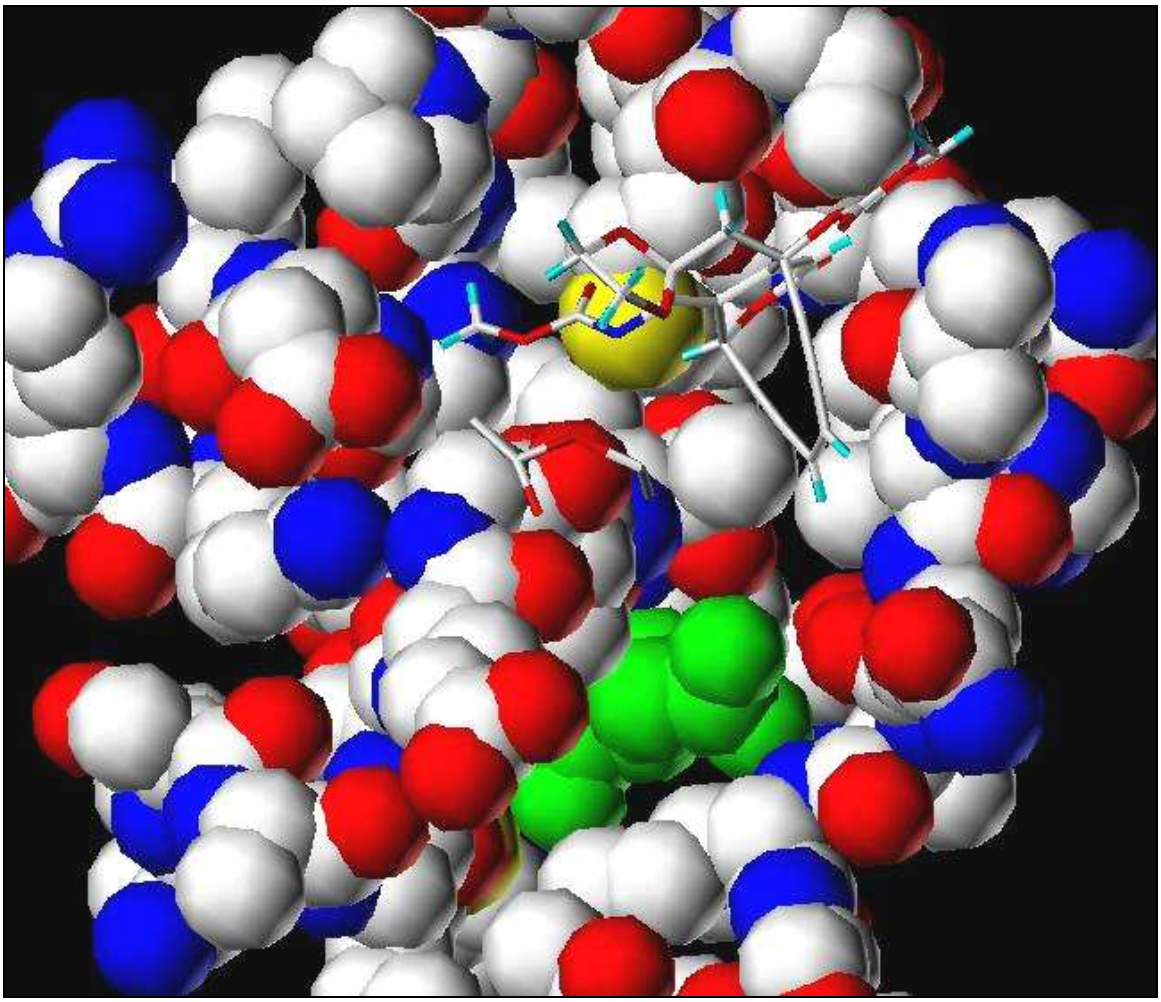


**Figure 4.21A SOCVIP, score -24.81 does not have visually obvious interactions with any of the key residues or side-chains**

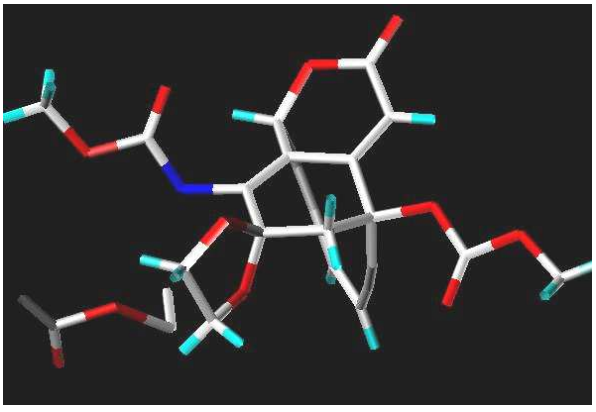


**Figure 4.21B SOCVIP**

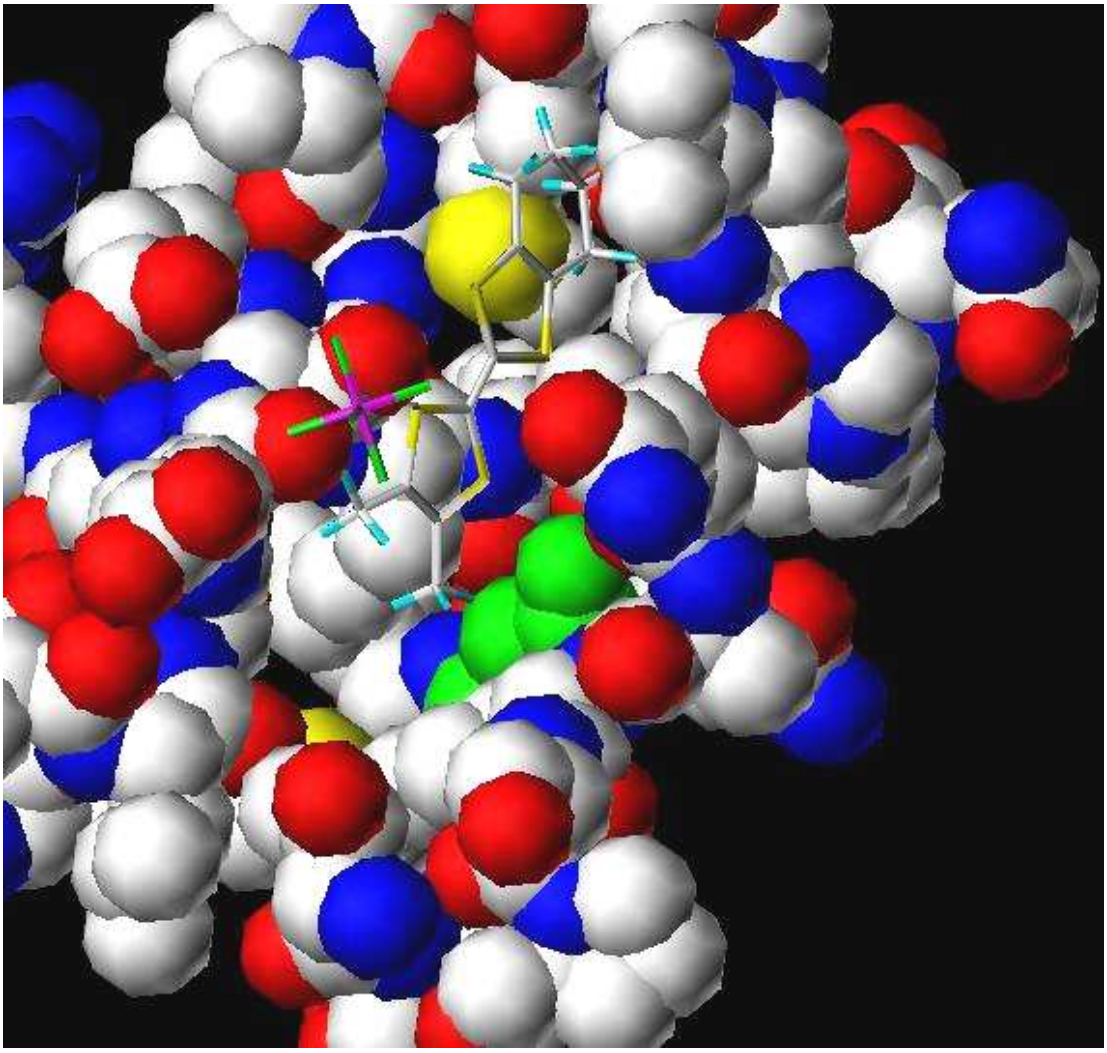


**Figure 4.22A SETJAC score -23.82 excellent ring groups and size distribution and is in line with the multi-ring, sulfur pharmacophore found frequently within the Cambridge database.**

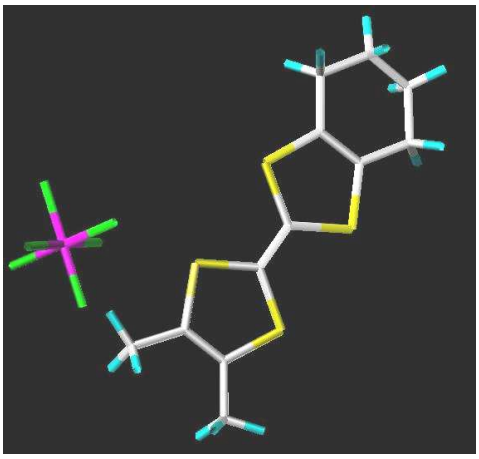


**Figure 4.22B SETJAC with solvent molecule**

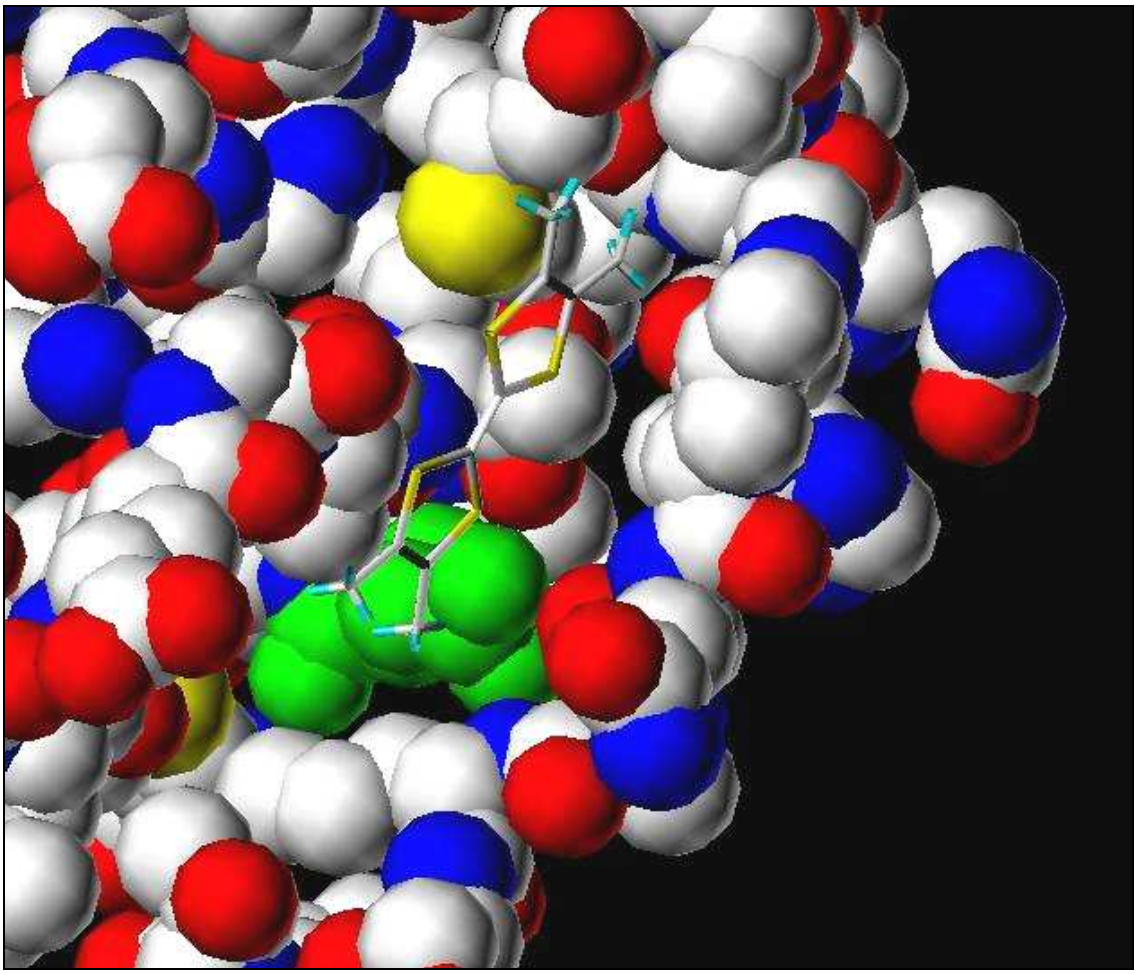


**Figure 4.23A – TMTFTC, score -24.86 multi-ring, multi-sulfur properties. The ring groups sufficiently fill the pocket and react with key residues.**

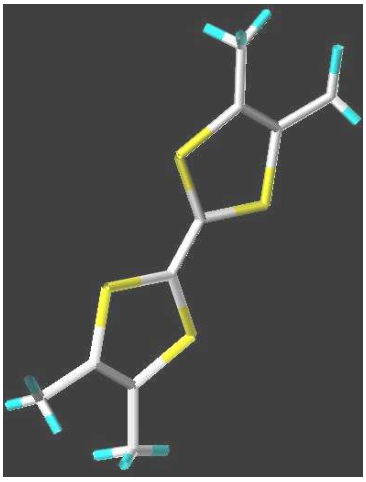


**Figure 4.23B – TMTFTC from Cambridge**

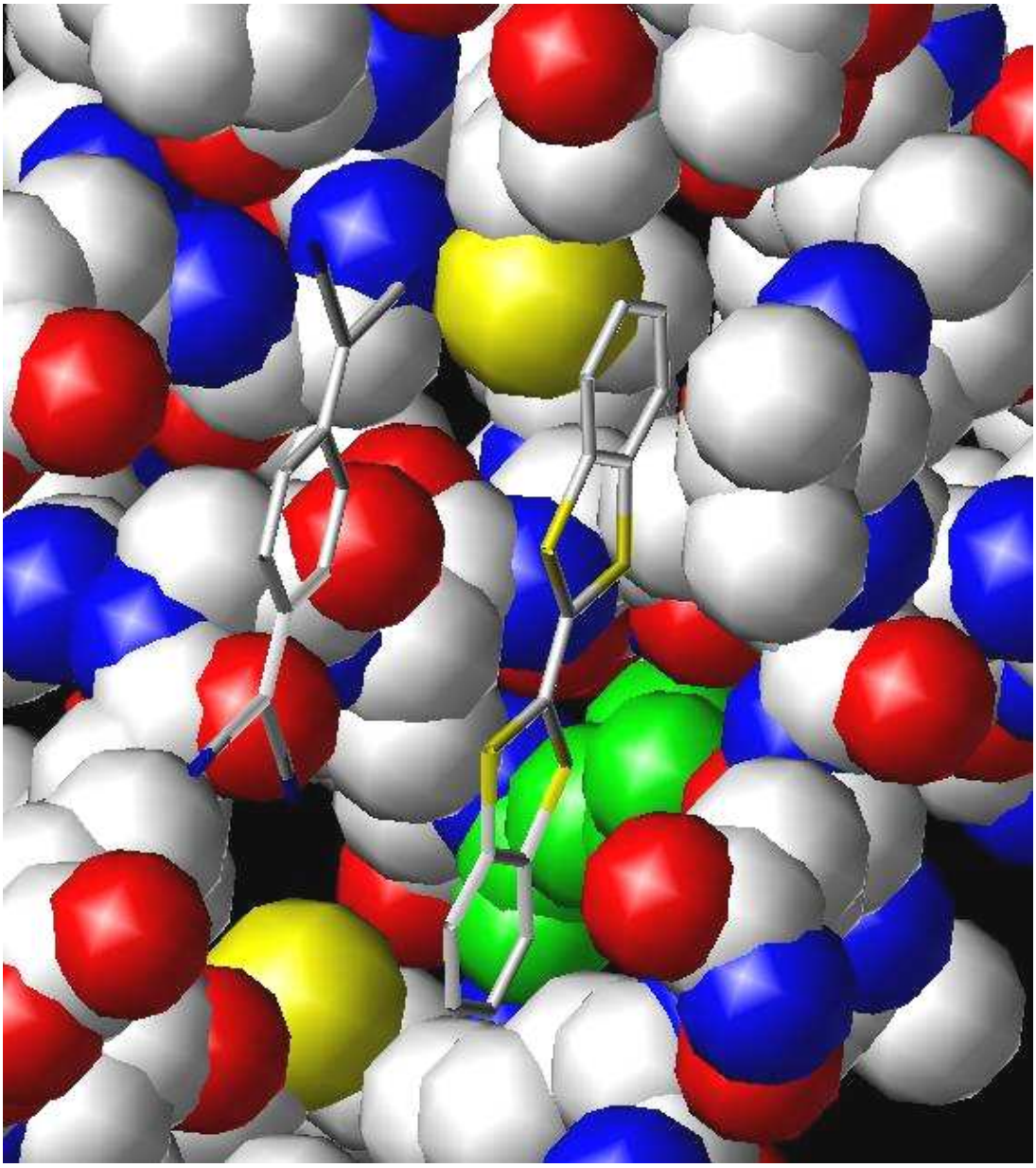


**Figure 4.24A – BALNAD01, score -22.40 interacts with all the key residues in the pocket with its multi-ring, multi-sulfur configuration. Shown with solvent molecule.**

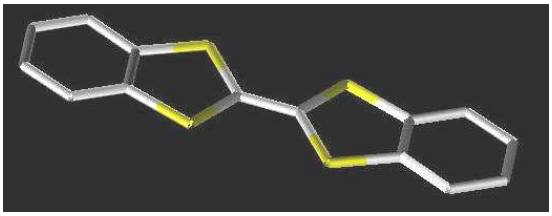


**Figure 4.24B - BALNAD01 from Cambridge**

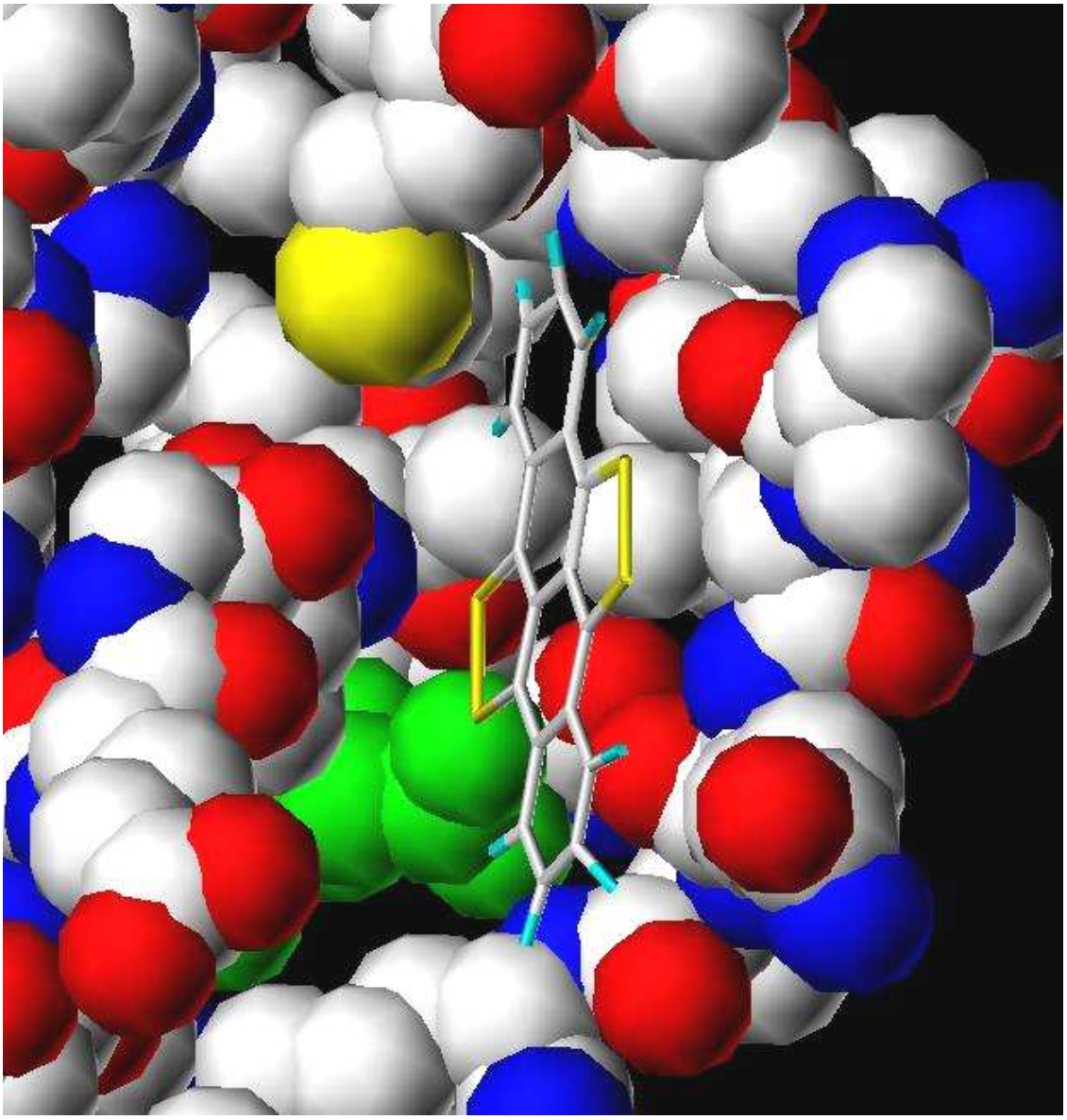


**Figure 4.25A – TETTR101, score -22.36 multi-ring, multi sulfur compound sufficiently fills the pocket and interact with all the key residues.**

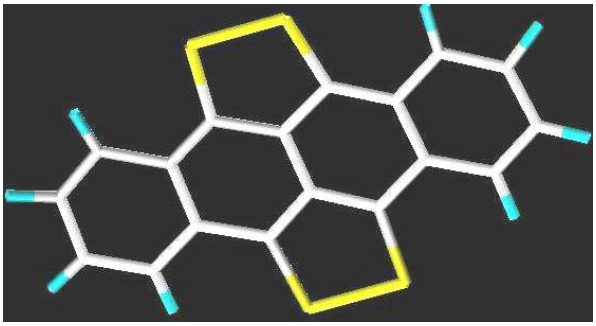


**Figure 4.25B - TETTRI01 from Cambridge**

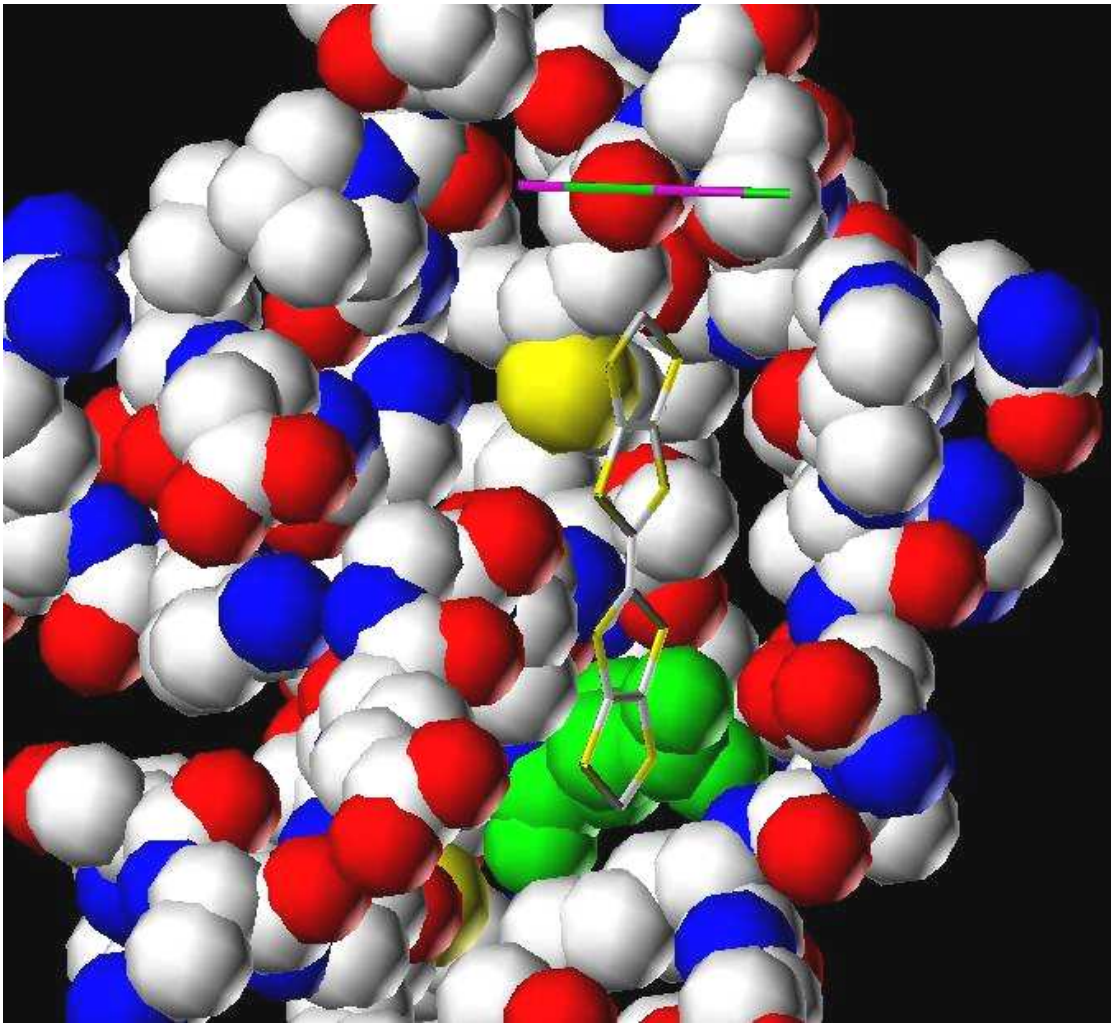


**Figure 4.26A GAYHUJ score -25.18, the top scoring compounds from the Cambridge database is the prototype of the multi-ring, multi sulfur pharmacophore (with solvent molecule).**

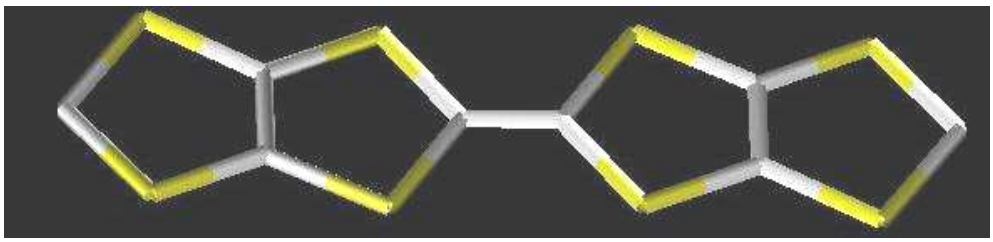


**Figure 4.26B GAYHUJ from Cambridge database**

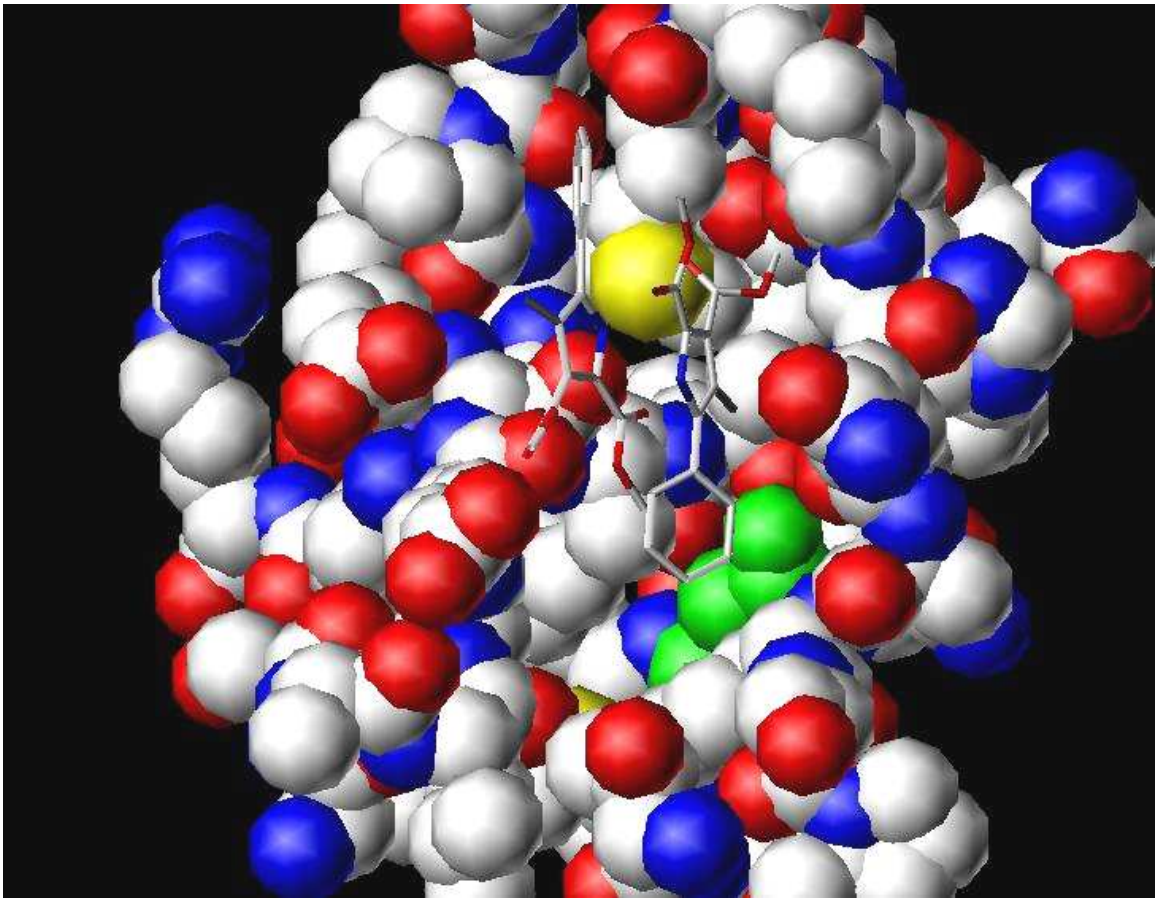


**Figure 4.27A - BUFNEV score -21.50 DOCK places the single nitrogen of the compound in a position where it is unable to readily interact with D51.**

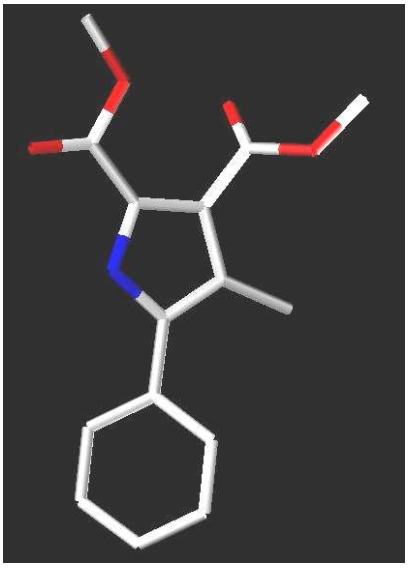


**Figure4.27B – BUFNEV from Cambridge Database**

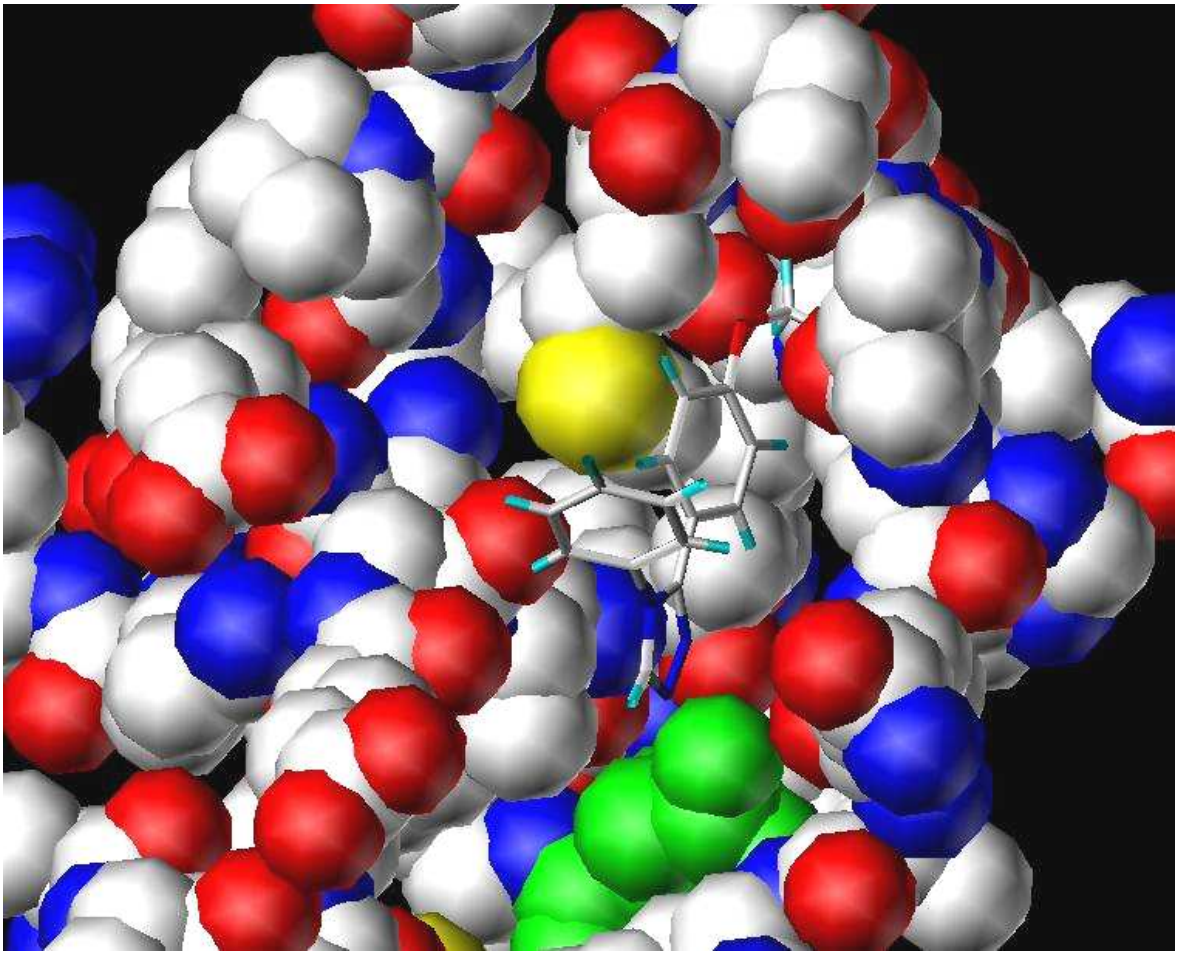


**Figure 4.28A – VEXPIX score -20.27 In an attempt to accommodate the chemical fit of amino to carboxyl, DOCK has to project a ring group away from the pocket to avoid steric hindrance, thus disturbing the other residue reactions with the amino ring group of the ligand.**

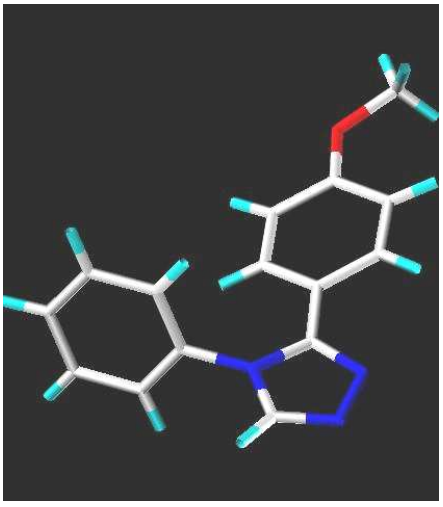


**Figure 4.28B – VEXPIX from Cambridge database**

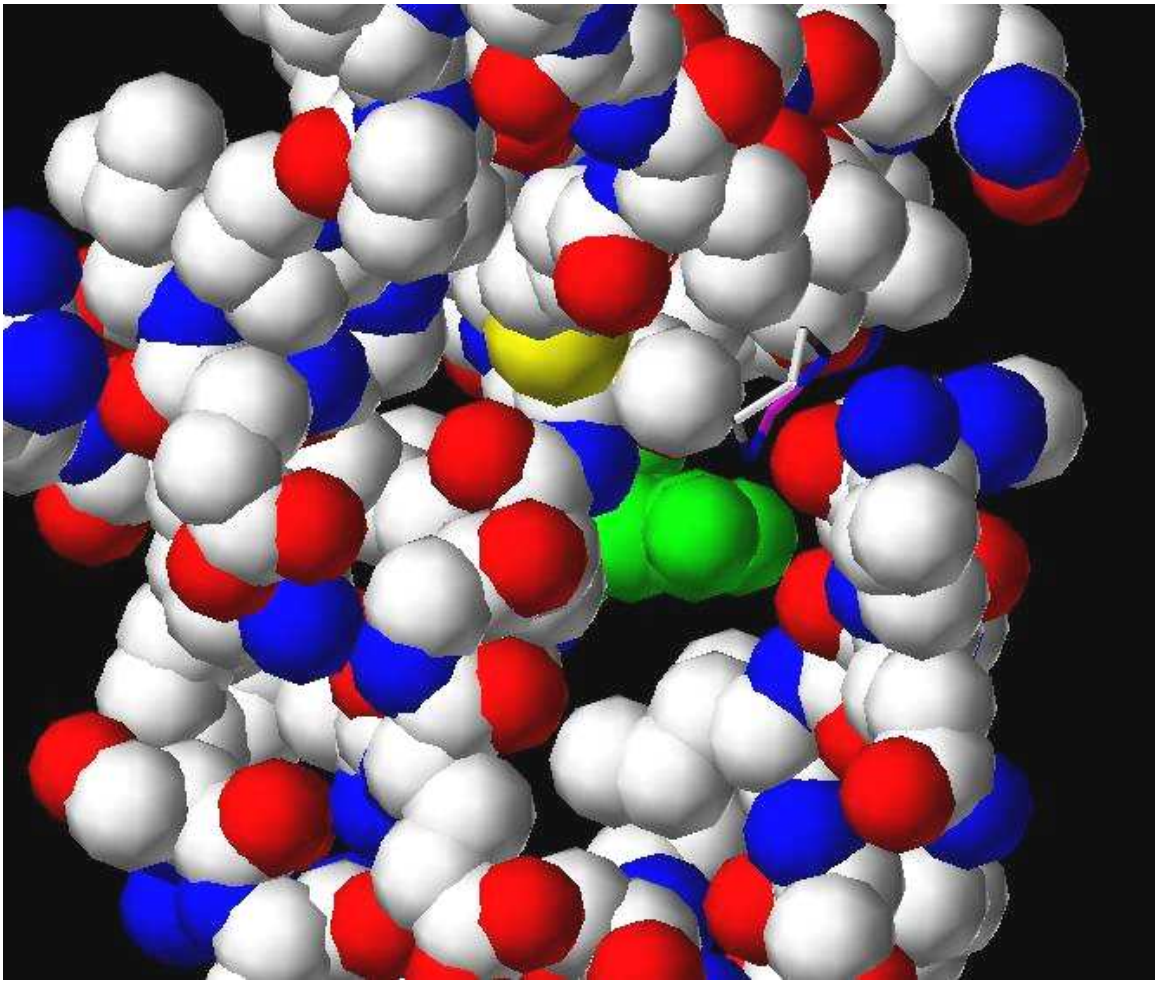


**Figure 4.29A -PNPTCC01, score -21.45, only a simple ring structure, incapable occupying the entire pocket and maintaining simultaneous interactions with residues of importance. Compounds containing nickel have been reported in blocking the gp120-CD4 entry.**

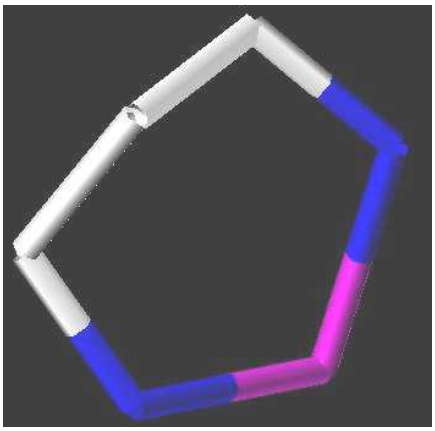


**Figure 4.29B PNPTCC01 from Cambridge database**

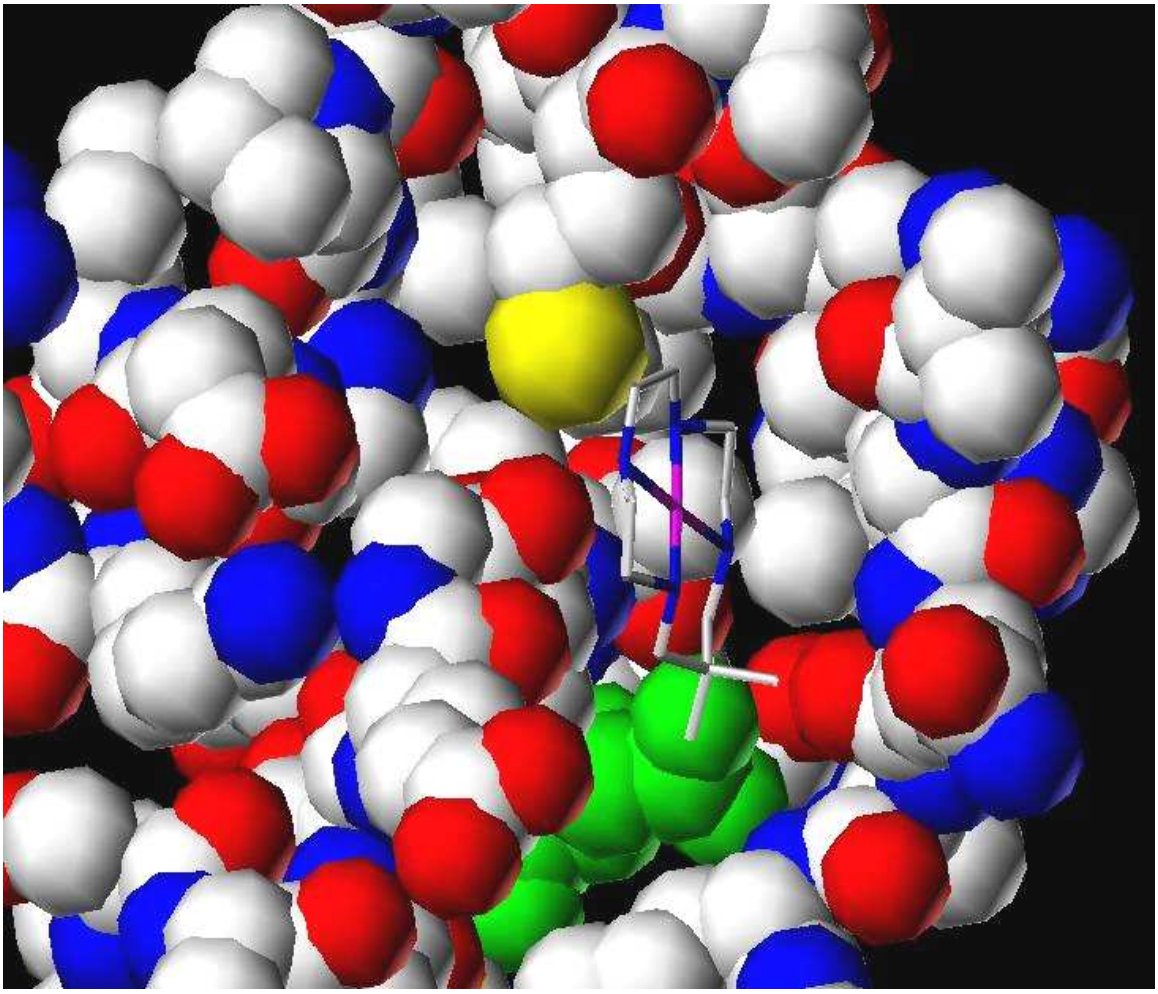


**Figure 4.30A MTAZNI score-24.96 Compounds containing nickel have been reported in blocking the gp120-CD4 entry.**

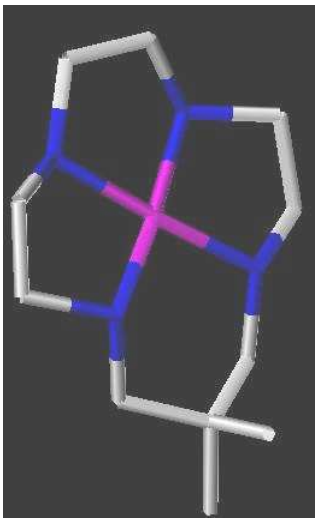


**Figure 4.30B- MTAZNI from Cambridge database**

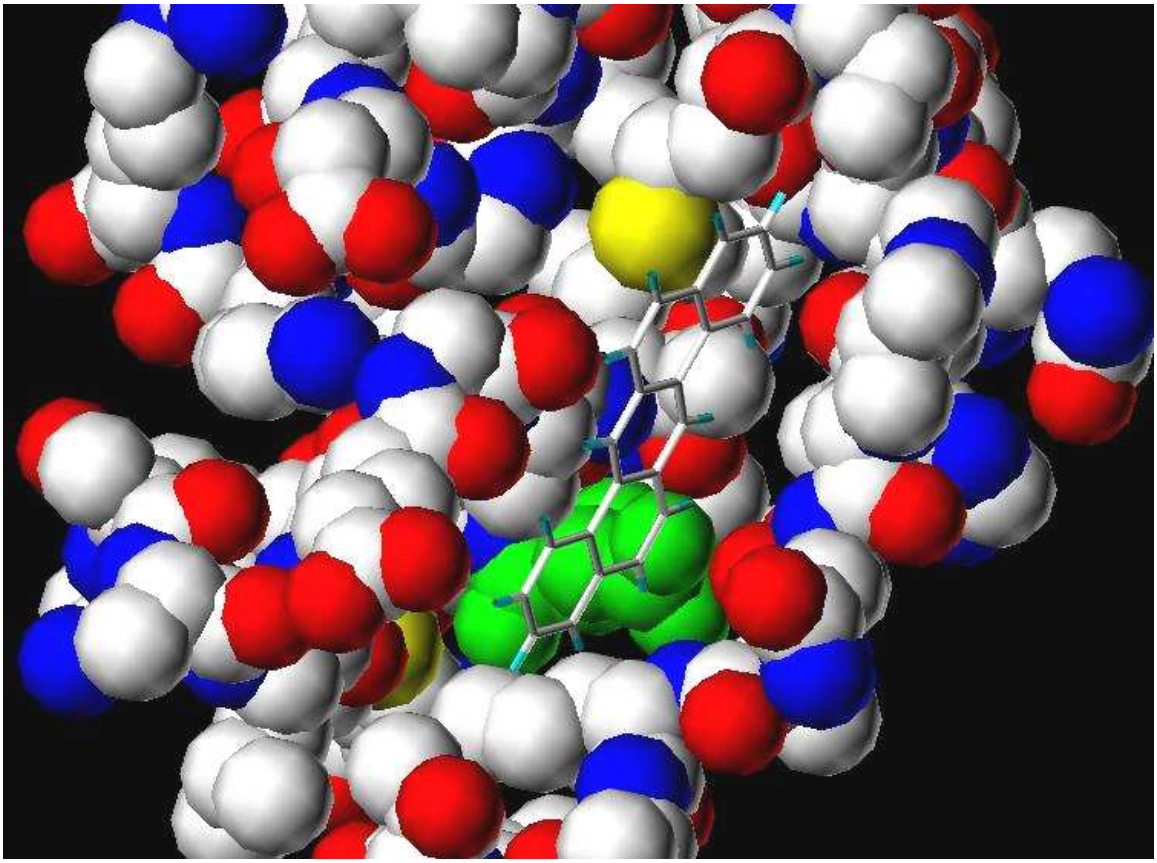


**Figure 4.31A -DBNTHR02 score -21.67 has no discernable chemical properties to assist in binding within the pocket, although its pocket-filling size is palpable**

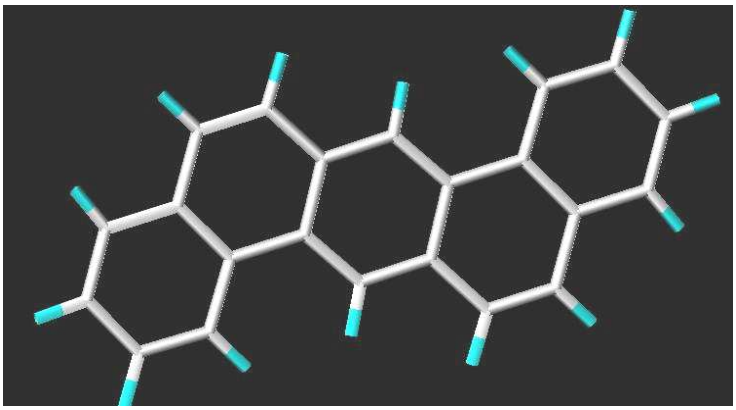


**Figure 4.31B - DBNTHR02 from Cambridge**

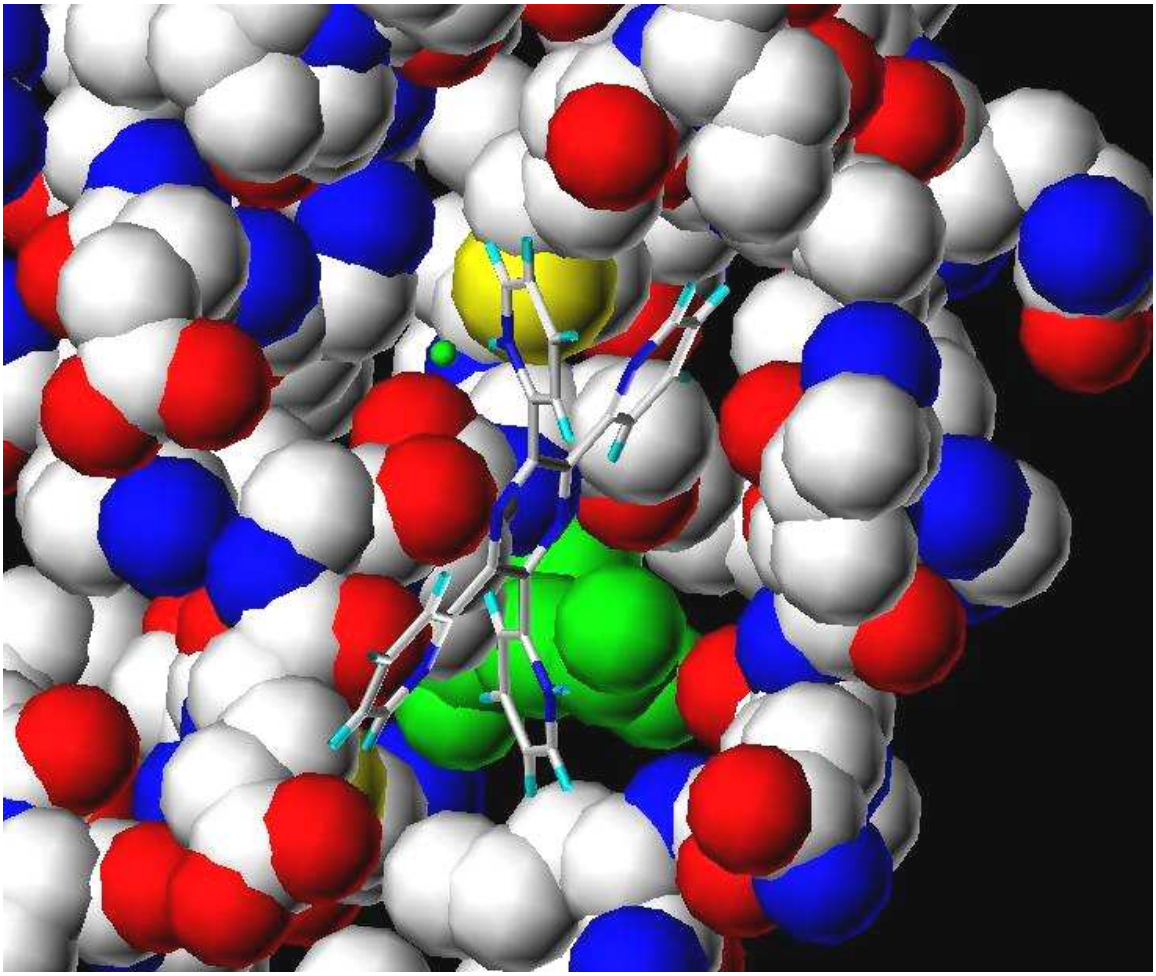


**Figure 4.32A –VOWDOA, score -21.58, presents multiple pyridine groups for interaction**

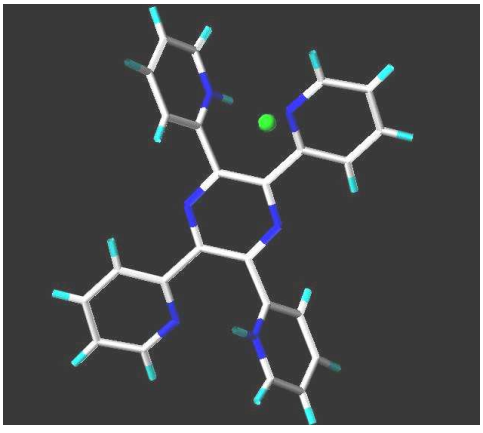


**Figure 4.32B – VOWDOA from Cambridge**

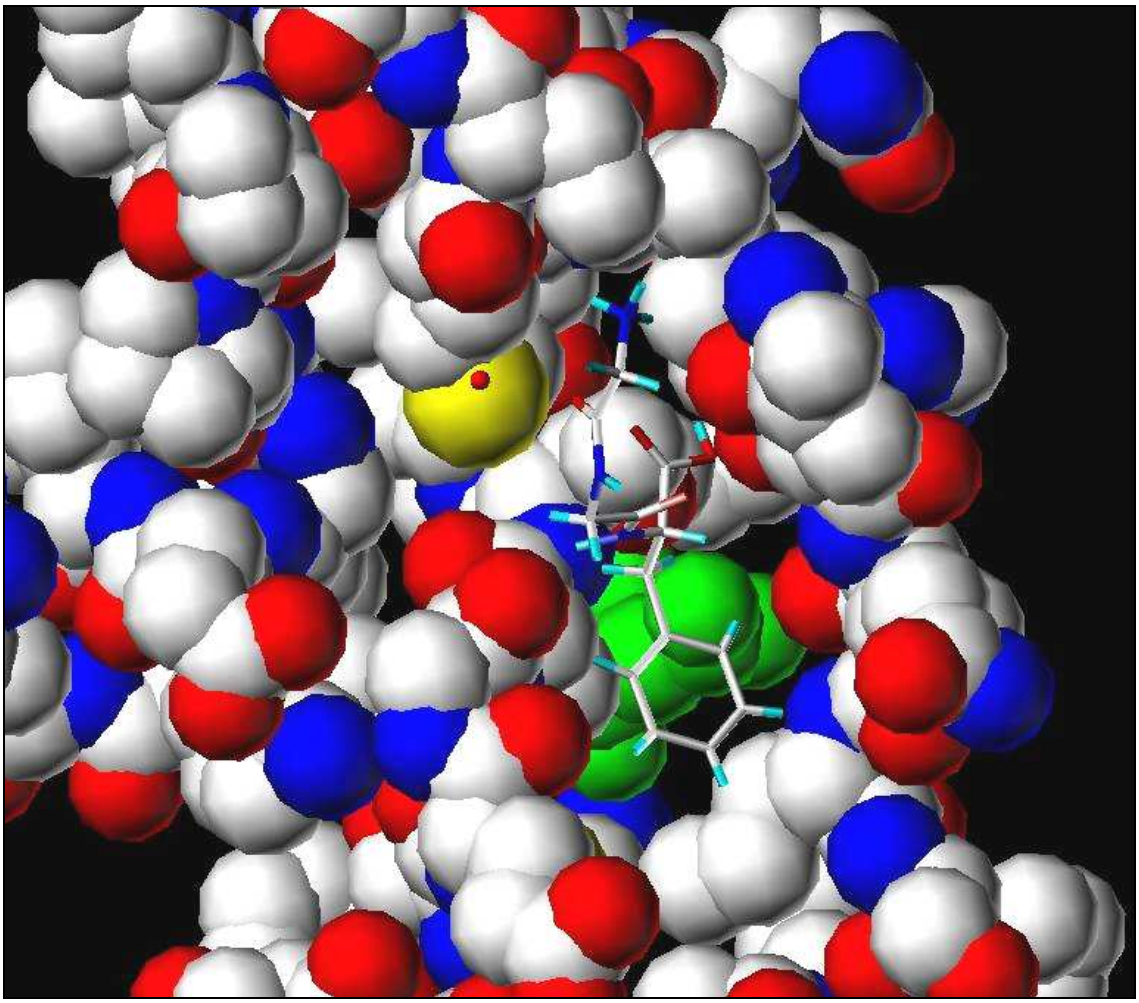


**Figure 4.33A -GADMIH, score -22.98, ignoring the amino groups present on the compound, DOCK shows its proclivity for steric fit, rather than chemical compatibility.**

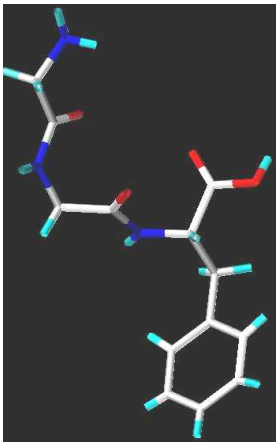


**Figure 4.33GADMIH from Cambridge database**

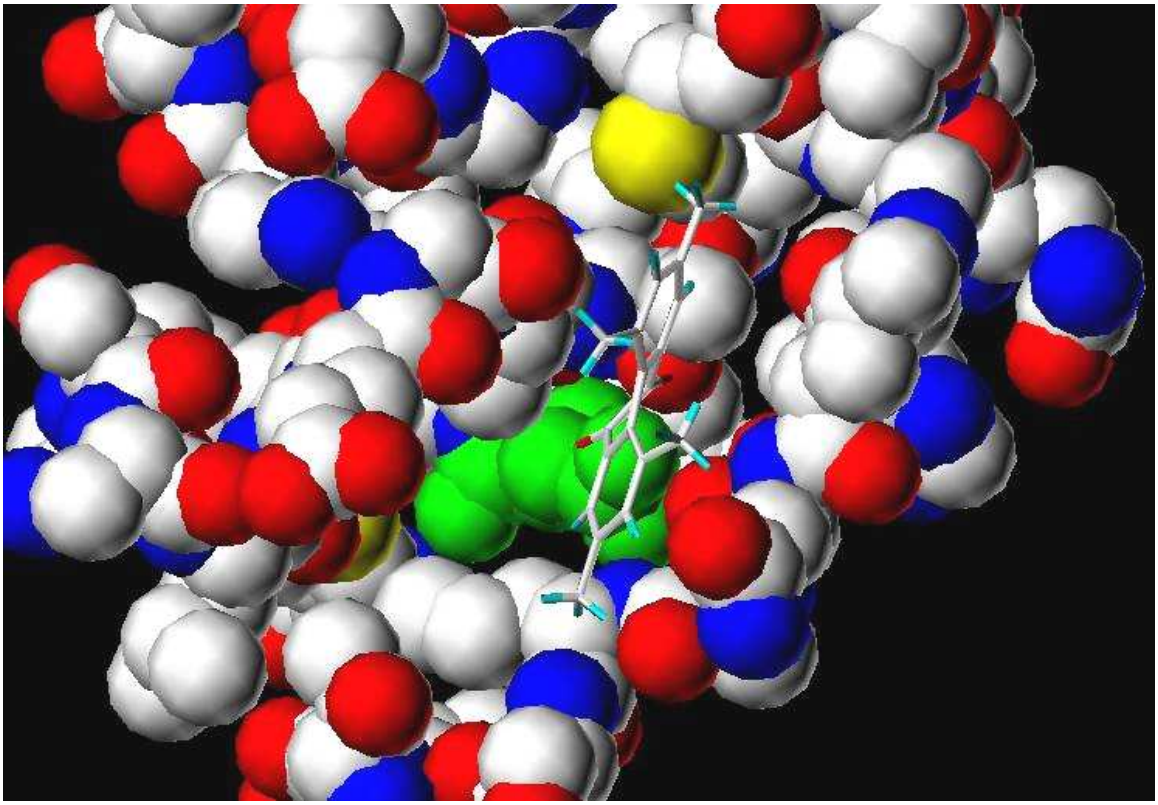


**Figure 4.34A FAVPAT score -22.02 FAVPAT has conformational possibilities, but the placement of two oxygens in close range of D51 is questionable**

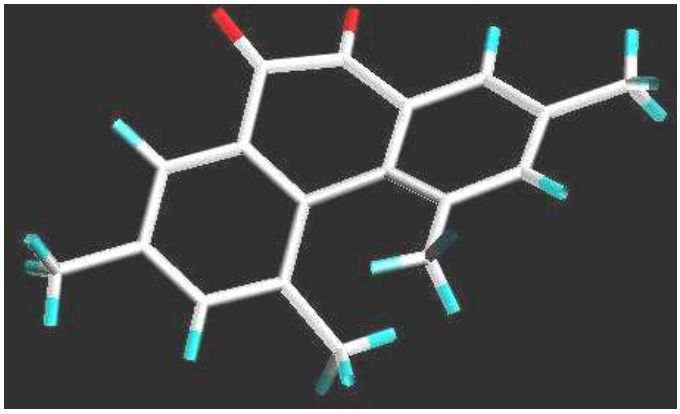


**Figure 4.34B – FAVPAT from Cambridge database**

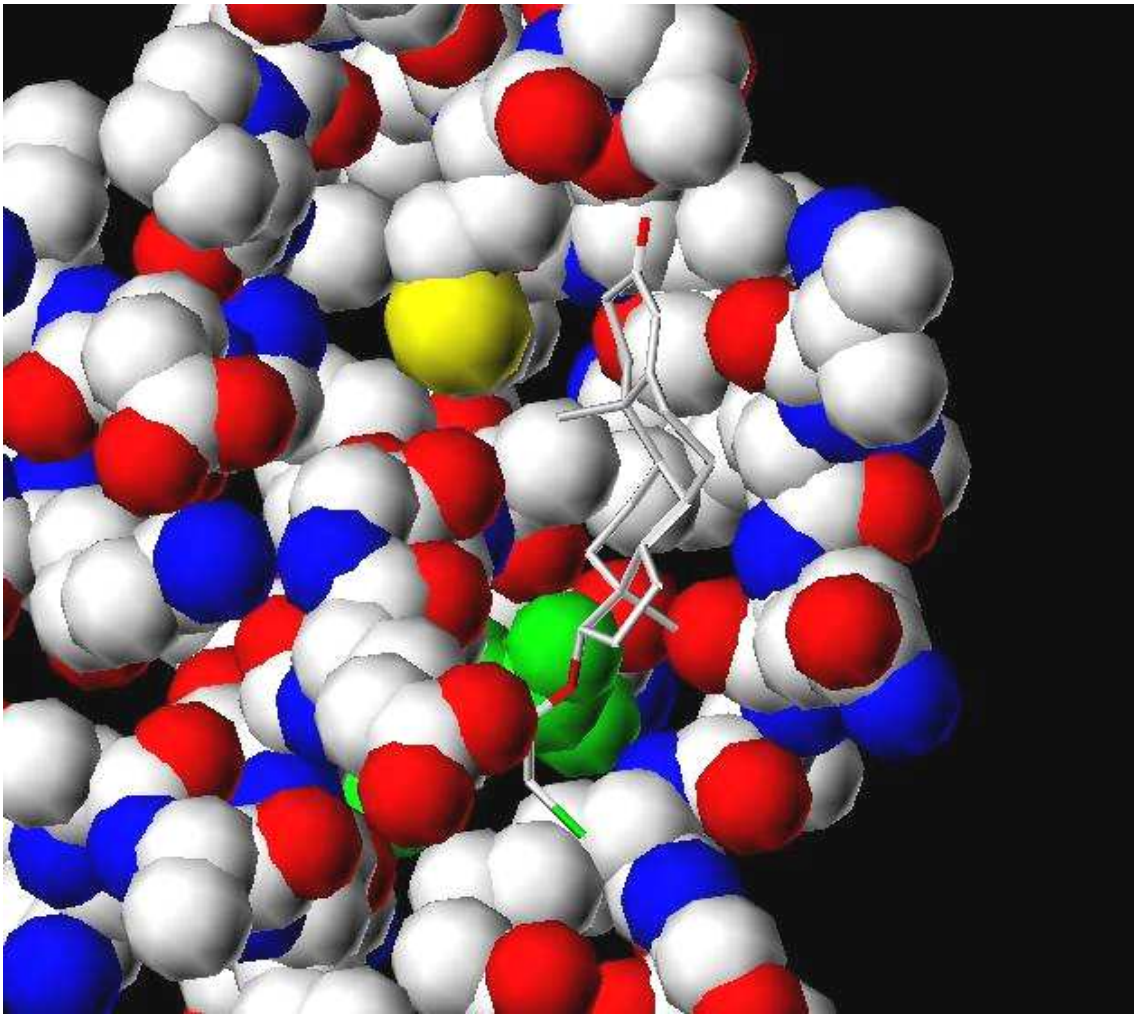


**Figure 4.35A - ANDREO score -21.30**
**ANDREO has been given the docking fate of placing its non-ring oxygen near critical residues.**

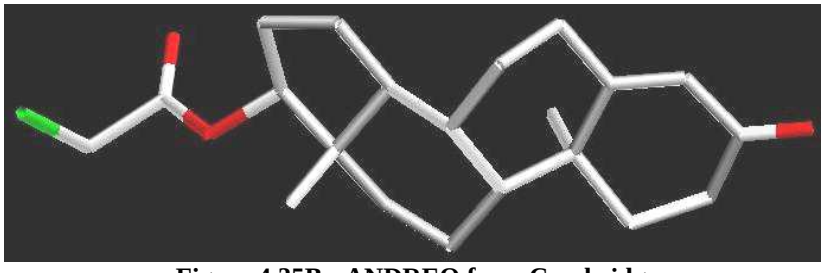


**Figure 4.35B - ANDREO from Cambridge**

**Results of Fragment Study**

The highest score of all database compounds was -51.07. Contrast this with the original 13 amino acid ligand registering a score of -7913.03, there is still considerable room for improvement. Ten, three and four residue fragments of the N-terminal were docked to understand how peptide fragments may energetically inhibit capsid formation. The 10 amino acid fragment (Pro1 through Met10) yielded an energy reading not statistically different from the original ligand -7904.81. The 3 residue fragment (Figure 4.36A) gave a score of -4311.36.

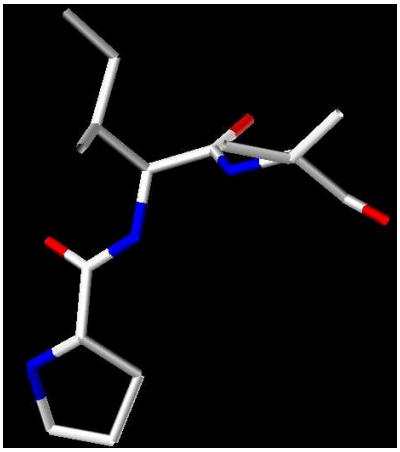


**Figure 4.36 – Pro1 through Val3**

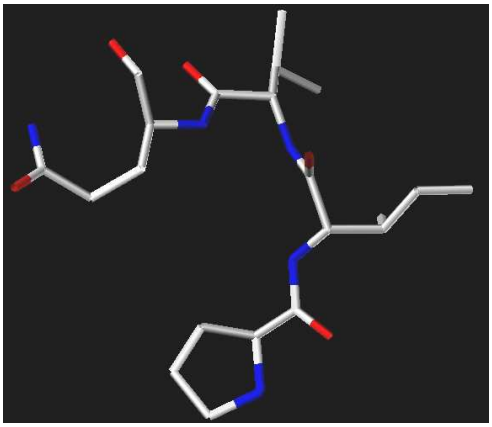


**Figure 4.37 – Pro1 through Gln4**

Figure 4.37 above shows the initial 4 amino acids as a single ligand. The returning docked score of -5350.54 shows the trend for better energy readings the closer the emulation of the entire 13 residue original ligand.

Adding or altering groupings from the top scores gave moderate but not overwhelming additional insight. Figure 4.38 shows the amino groups of the top scoring molecule, Str69897 being replaced by sulfur atoms. The returned energy score jumped to a significantly lower -171.96. However, similar alterations to the next two highest scores did not reveal a similar

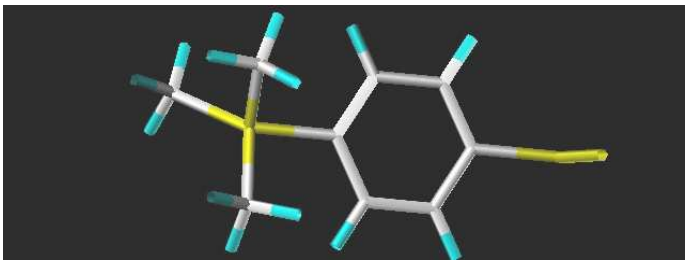
decrease in energetics. Str87814 with the same sulfur for nitrogen exchange actually had a raised energy score to -27.03. 

**Figure 4.38– Str69897**

Str29785 made a statistical improvement but not one of significance, lowering to -29.41.

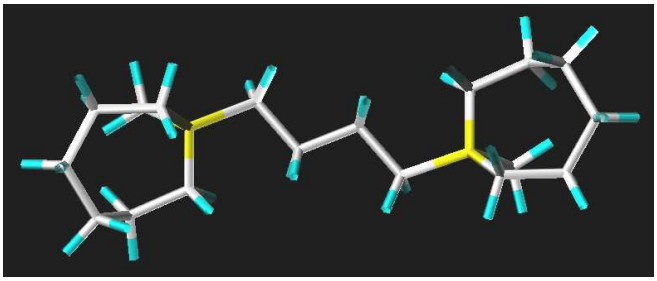


**Figure 4.39 – Str87814 with alterations**

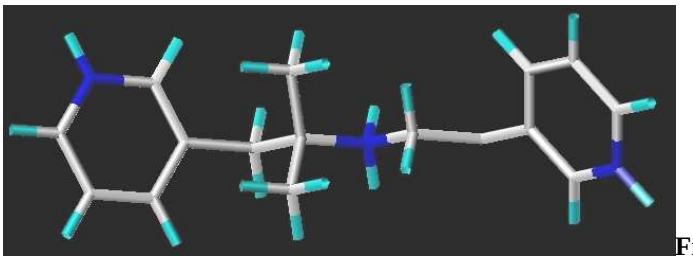


**Figure 4.40 – Str29785**

A few molecules were created (via Sybyl) as attempts to fit the pharmacophores that had been successful. Figures 4.41 and 4.42 show two compounds with successful scores (-20.41 and -17.71 respectively), thus illustrating that alterations to existing databases compounds will be more fruitful than creating new ones.

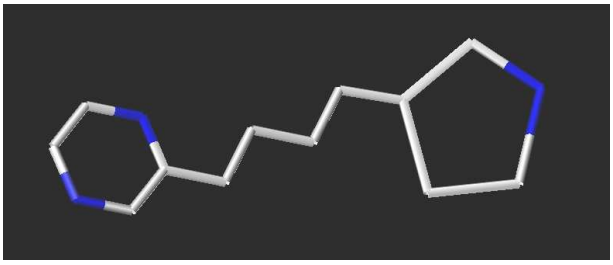


**Figure 4.41**

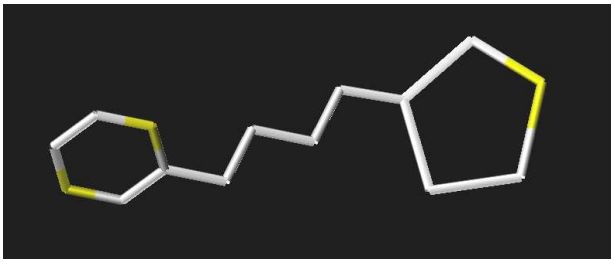


**Figure 4.42**

# Chapter V

## Conclusions and Future Work

Acquired Immune Deficiency Syndrome (AIDS) and its causative agent the Human Immunodeficiency Virus (HIV) have combined to utterly mesmerize the international realms of healthcare, life science, social science and politics. It is the most talked about and feared contagious disease of this generation. Death tolls and infection rates continue to climb, mostly in locations of depressed socio-economic conditions. The worldwide attention garnered by the medical and scientific communities, as represented in the form of published journal articles towards HIV is rivaled only by Cancer, Heart Disease and Stroke (Centers for Disease Control). However, despite the worldwide focus on HIV, current available therapies have only crossed two thresholds – the inhibition of viral enzymes protease and reverse transcriptase. While halting or slowing the efficiency of these two enzymes is critical to the inhibition of the overall viral life cycle, other targets exist upon which the field of drug discovery can capitalize.

In addition to the therapies currently available, work done in the field of drug discovery has yielded important leads and compounds in various stages of clinical phase trials. New leads are especially critical in vanquishing HIV targets where no therapies presently subsist, particularly the enzyme integrase and the structural proteins gp120 and gp41. Several improvements in the current therapies targeting protease and reverse transcriptase are well underway, including the potential of eliminating the mitochondrial toxicity of nucleoside reverse transcriptase inhibitors (NRTI) (Lewis *et al.,* 2003). As a topical approach, affordable, user-

friendly microbicides (in the form of gels, creams, foams, and suppositories) are under development with a target year of 2007. More than 60 of these compounds, some of which were initially developed as anti-retroviral agents, are currently in this surfactant approach pipeline and will be suitable for female-controlled vaginal insertion prior to intercourse (Stone, 2002). These therapies may be of extreme importance to underdeveloped areas of the globe, where use of to condoms or outright abstinence is of low frequency.

To date, computational methods have been used in a wide variety of anti-viral drug discovery projects in an effort to decrease the time required to screen the ever-increasing volumes of small molecule databases and libraries. DOCK 4.0, the software used in this particular dissertation, has been previously utilized directly in the creation of therapies currently available on the market. Programs with similar strategies and algorithms used by DOCK 4.0 have contributed to the list of compounds presently in the phase trial system.

The two small molecule databases used to screen via DOCK 4.0 have each been recognized as sources for matching drug targets with interactive compounds. The National Cancer Institute's website displays that its 3D database has been used to unearth novel compounds against HIV integrase, reverse transcriptase and protease. According to the Cambridge Crystallographic Database Center's website, as of the year 2000, over 10,000 journal citations referenced compounds derived from its database. If screening and docking conducts researchers towards any singular compound, they are available for registered users per the following websites:

Cambridge: http://www.ccdc.cam.ac.uk/products/csd/request/

NCI: http://dtp.nci.nih.gov/docs/misc/available_samples/dtp_indsamples.html/.

The HIV-1 capsid, a structural protein which surrounds the viral genome becomes an obvious target due to its protective function. The literature shows that science is still not clear about the interface of each dimer unit. Hence the complete account for the formation of the fully functional capsid electron dense barrier is unknown. However the amino terminus of this protein displays excellent potential as a drug target, bearing that an integral interaction between highly conserved residues can be targeted for disruption. The hypothesis is that a compound capable of fitting into and interacting with residues within the B-hairpin “pocket” created by the absence of the yet unfolded initial 13 residues will prevent final assembly formation of the capsid protein.

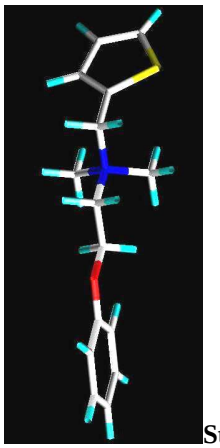
DOCK 4.0 and its sub-suite of programs were used to screen compounds from the two separate databases and rank them according to their ability to emulate the aforementioned N-terminal pocket interactions. Sybyl 6.7 was used to visualize ligand-receptor interactions and with the creation of pharmacophore-related compounds. Compounds in the top rankings were re-docked and evaluated individually.

Pharmacophores identified include sulfur-containing compounds, multi-ring structures and compounds with flexional torsion to maintain several interactions within the pocket.

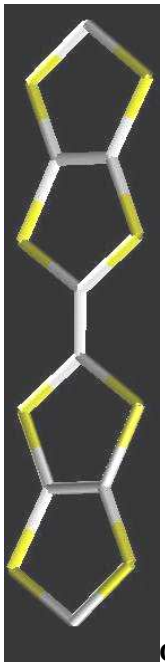
It was not the primary intention of this project to compare compounds from the National Cancer Institute’s database directly to those derived from Cambridge; however the scores, shape complimentarity, steric constraints and chemical composition all are superior in NCI derived compounds. Indeed, four of the top 5 overall scoring molecules (the exception is GAYHUJ seen in figure 5.1) are from NCI database.

The characteristics of compound Str69897 (figure 5.1) which has an amino group emulating the Pro1 ring and its subsequent interactions with Gly46 and Ile15, an amino group

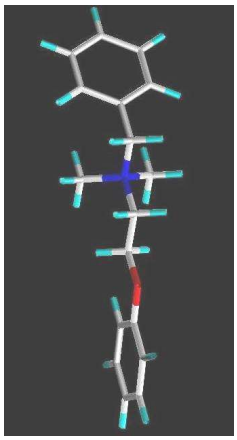
capable of interacting with D51 and Gln13 lacks the size to fill the rest of the pocket. The sulfur moieties found in some of the top Cambridge database may be interesting to test efficacy in follow up studies as a sulfur containing compound was also a top 5 NCI score. There is precedent for sulfur compounds in HIV research as the mechanism of 3-O-[3',3'-dimethylsuccinyl]-betulinic acid (DSB) has active inhibition of HIV-1 replication. The compound does not act on the protease itself, but on the Gag cleavage site between the capsid and the p2 protein (Zhou *et al.*, 2004). The chemical compositions of Str29785 and Str87814 lead them to be excellent candidates for multiple interactions. The goal is to interact precisely and directly with D51, as it is the most conserved residue of the 4 key residues within the pocket. Many current therapies (especially targeting HIV-1 protease and reverse transcriptase) become ineffective as mutations of targeted non-essential amino acids within the site are common. This obstacle may not be the case in N-terminal capsid endeavors.



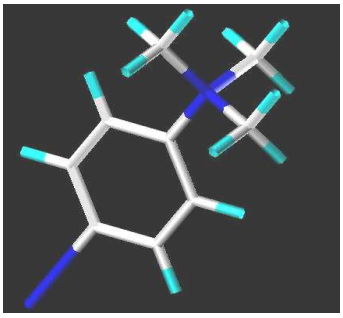
**Str9620**



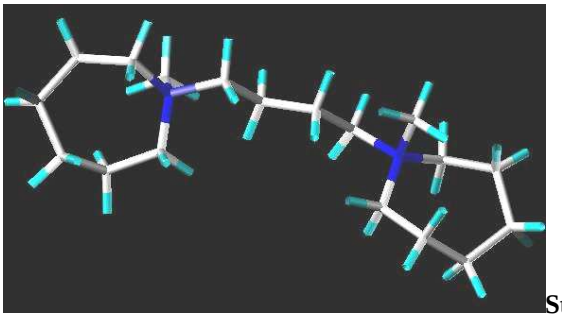
**GAYHUJ**



**Str29785**



**Str69897**



**Str87814**

**Figure 5.1 – The top five scoring compounds overall.**

The shape complementarity and torsional flexibility of any of these top5 molecules (with the exception of diminutive 69897) is acceptable as there is enough range of motion throughout the compound to fill into the pocket, but not so much as to be rigid and not stay within the pocket. Compounds such VOWDOA (figure 5.2), while containing the requisite amount of amino and ring groups, may have the bulkiness to not fit into this pocket in the first place.

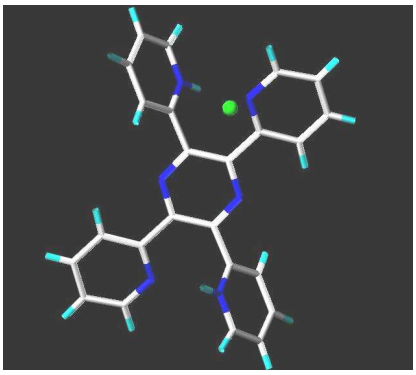


**Figure 5.2 VOWDOA**

Virtual docking is an extremely effective method of identifying pharmacophores versus a specific target. With expert personnel teaming to multi-task the UNIX-scripting, molecular modeling, medicinal chemistry and protein expression and purification, the ability to test the activity of small groups of directed compounds is quite compelling.

Future work involves the creation or acquisition of a few of these lead compounds. A medicinal chemist could create variations of any acquired compounds and test them versus the purified, wild-type capsid protein using spectrophotometric methods. Collaborations with a pharmaceutical company could expand this testing with more numerous pharmacophores.

# References


Abagyan R, Trotov M, Kuznetsov D. 1994. ICM- a new method for protein modeling and design: Applications to docking and structure prediction from the distorted native conformation. *Journal of Computational Chemistry* V**15**: 488-506

Ajay, Murcko MA. 1995. Computational methods to predict binding free energies in ligand-receptor complexes. *Journal of Med. Chem.* **38**:4953-4967

Amzel, LM. 1997. Structure-based drug design. *Current Opinions in Biotechnology* Volume**9**: 366-369

Aqvist J. 1996. Calculation of absolute binding free energies for charged ligands and effects of long-range electrostatic interactions. *Journal of Computational Chemistry* 17: 1587-1597

Artico, M *et al.*, 2000. Structure-based design, synthesis and biological evaluation of novel pyrrolyl aryl sulfones: HIV-1 non-nucleoside reverse transcriptase inhibitors active at nanomolar concentrations. *Journal of Medicinal Chemistry* 43(9): 1886-91.

Balkenhohl F, *et al.*, 1996. Combinatorial synthesis of small organic molecules. *Agnew Chem. International Ed Engl.* **35:** 2288-2337

Barklis, E. *et al.*, 1997. Structural analysis of membrane-bound retrovirus capsid proteins. *European Molecular Biology Organization* Volume 16:1199-1213

Beerenwinkel, N. *et al.*, 2003. Methods of optimizing antiviral combination *therapies Bioorganic & Medicinal Chemistry* Volume 11 Issue 17 pages 3589-3593

Blaney JM, Dixon M. 1993. A good ligand is hard to find: Automated docking methods. *In Perspectives in Drug Discovery and Design* Volume 1: 301-319

Bohm HJ, Stahl M. 1999. Rapid empirical scoring function is virtual screening applications. *Med. Chem. Res*. 9: 445-462.

BonHomme *et al.,* 2003.The pH dependence of HIV-1 capsid assembly and its interaction with cyclophilinA A. *Biophysical Chemistry Volume* 105 Issue 1:67-77

Braaten, D. Franke, E. and J. Luban. *J. Virol.* **70** (1996), pp. 3551–3560. Abstract-EMBASE | Abstract-MEDLINE | Abstract-Elsevier BIOBASE

Broomijmans N & Kuntz, I. 2003 Molecular Recognition and Docking Algorithms. *Annual Review of Biophysics and Biomolecular Structure* 32:335-73

Busetta B, Tickle IJ, Blundell TL. 1983. DOCKER, an interactive program for simulating protein receptor and substrate interactions. *Journal of Applied Crystallography* **16**: 432-437

Callebaut C, Krust B, Jacotot E, Hovanessian AG. 1993. T cell activation antigen, CD26, as a cofactor for entry of HIV in CD4+ cells. *Science.* 1993 Dec 24;262(5142):2045-50.

Cann, AJ, & Karn, J 1989, Molecular biology of HIV: new insights into the virus life cycle. *AIDS* Volume3 (Suppl. 1) S19-34

Carlson HA, Masukawa KM, Rubins K, Bushman FD, Jorgnesen WL, Lins RD, Briggs, JM. 2000. Developing a dynamic pharmacophore model for HIV-1 integrase. *Journal of Medicinal Chemistry* 43: 2100-2114

Centers for Disease Control and Prevention (CDC) 2002 MMWR 60:230-232

Chen, D. *et al.,* 1998. HIV entry and its inhibition. *Cell* Volume89:263-273.

Chimirri, A. *et al.,* 2001. Synthesis, biological activity, pharmacokinetic properties and molecular modeling studies of novel 1H, 3H-oxazolo[3, 4a] benzimidazoles: non-

nucleoside HIV-1 reverse transcriptase inhibitors. *Antiviral Chemical Chemotherapy* Volume 3:169-174

Connolly, ML., 1983. Solvent accessible surfaces of proteins and nucleic acids. *Science* 221: 709-713

Connolly, M. 1983. Analytical molecular surface calculation. *Journal of Applied crystallization* 16: 548-558.

Cooley LA and Sharon RL, 2003. HIV-1 cell entry and advances in viral entry inhibitor therapy *Journal of Clinical Virology* Volume 26, Issue 2:121-132

D' Cruz, OJ. & Ruckus, FM, 1999 Novel Derivatives of Phenethy-5-Bromopyrdylthiourea and Dihydroalkoxybenzyl-oxopyrimidine Are Dual-Function Spermicides with Potent Anti-Human Immunodeficiency Virus Activity. *Biology of Reproduction* Volume 60:1419-1428.

De Clercq, E. *et al*., 1994. Highly potent and selective inhibition of human immunodeficiency virus by the bicyclam derivative JM3100. *Antimicrobial Agents in Chemotherapy* Volume 38 pp668-674

De Clercq, E. *et al.*, 2002. New Developments in anti-HIV chemotherapy. *Biochemica et Biophysica Act* 1587:258-275.

Deres, K., C.H. Schroeder, A. Paessens, S. Goldmann, H.J. Hacker, O. Weber, T. Kramer, U. Niewohner, U. Pleiss, J. Stoltefuss, E. Graef, D. Koletzki, R.N.A. Masantschek, A. Reimann, R. Jaeger, R. Grob, B. Beckermann, K.H. Schlemmer, D. Haebich and H. Rubsamen-Waigmann, Inhibition of hepatitis B virus replication by drug-induced depletion of nucleocapsids. *Science* **299** (2003), pp. 893–896.

DesCharlis RL, Sheridan RP, Dixon JS, Kuntz ID. 1986. Docking flexible ligands to macromolecular receptors by molecular shape. *Journal of Medicinal Chemistry* 29: 2149-53.

Dickson, C., *et al.,* 1984. Protein Biosynthesis and Assembly p513-648. In R. Weiss, N. Teich. H. Varmus, and J. Coffin (ed.*) RNA tumor viruses.* Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.

Dorfman T, *et al.*, 1997. Active site residues of cyclophilinA are crucial for its incorporation into human immunodeficiency virus type 1 virions. *Journal of Virology* Volume 71 pages 7110-7113.

Erlich L, Agresta, B. and Carter, C. 1992. Assembly of Recombinant Human Immunodeficiency Virus Type 1 Capsid Protein In Vitro. *Journal of Virology* Volume 66:4874-4883

Ewing TJA, Makino S, Skillman AG, Kuntz ID. 2001. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *Journal of Computer Aided Drug Design* **15:**411-428

Fahmy A, Wagner G. 2002. TreeDock: a tool for protein docking based on minimizing van der Waals energies. *Journal of American Chemical Societies* 124 (7): 1241-50.

Finn PW, Kavarki LE, Latombe LC, Matwani R, Shelton C, Yao A. RAPID: Randomized Pharmacophore Identification for Drug design. *Computational Geom. ACM* 97: 324:333.

Filikov, AV, 2000. Identification of ligands for RNA targets via structure-based virtual screening: HIV-1 TAR. *Journal of Computation Aided Molecular Design* Volume 6:593-610

Fischer WB, 2003. Vpu from HIV-1 on an atomic scale: experiments and computer simulations. *Federation of European Biochemical Studies* 553(1):39-46

Fitzon, T. *et al.,* 2000.Proline Residues in the HIV-1 $NH_2$-Terminal Capsid Domain: Structure Determinants for Proper Core Assembly and Subsequent Steps of Early Replication. *Virology* 268:294-307.

Folch, MD *et al.,* 2003.Infectious Diseases, Non-zero-sum thinking, and the Developing World *The American Journal of the Medical Sciences* 326(2): 66-72.

Fontenot, JD, Tan X and Phillips DM. 1998. Structure-based design of peptides that recognize the CD4 binding domain of HIV-1 gp120. *Acquired Immune Deficiency Syndrome* Volume 12:1413-1418.

Forshey, B.M., U. von Schwedler, W.I. Sundquist and C. Aiken. 2002. Formation of a human immunodeficiency virus type 1 core of optimal stability is crucial for viral replication. *J. Virol.* **76** pp. 5667–5677.

Fradera X, Knegtel RMA, Mestres J, 2000. Similarity-Driven Flexible Ligand Docking *Proteins: Structure, Function and Genetics* **40**: 623-636

Franke, EK., Hui EH., Luban Y. 1994. Specific incorporations of cyclophilinA A in HIV-1 virions. 1994. *Virology* 372(6504): 359-62

Gabb J, Jackson RM, Sternberg MJE. Modeling protein docking using shape complementarity, electrostatics and biochemical information. *Journal of Molecular Biology.* 272: 106-120

Gabriel, JL and Mitchell, WM, 1996. Functional design of potential inhibitors of human immunodeficiency virus (HIV) binding to CD4+ target cells: a molecular model of gp120 predicts ligand binding. *Drug Design and Discovery* Volume 2:103-114.

Gamble TR., F. Vajdos, S. Yoo, D.K. Worthylake, S.M. Houseweart, W.I. Sundquist and C.P. Hill, Crystal structure of human cyclophilinA bound to the amino-terminal domain of HIV-1 capsid. *Cell* **87** (1996), pp. 1285–1294.

Gamble, TR., S. Yoo, F.F. Vajdos, U.K. von Schwedler, D.K. Korthylake, H. Wang, J.P. McCutcheon, W.I. Sundquist and C.P. Hill, Structure of the carboxyl-terminal dimerization domain of the HIV-1 capsid protein. *Science* **278** (1997), pp. 849–853.

Geijenbeek *et al.,* 2000. DC-SIGN, a dendritic cell-specific HIV-1 binding protein that enhances trans-infected of T cells. *Cell* volume100 pp587-597

Gelderblom, HR., *et al.,* 1987. Fine structure of human immunodeficiency virus (HIV) and immuno-localized of structural proteins. *Virology* Volume156: 171-176.

Gilson MK, Given JA, Bush BL, McCammon JA. 1997. The statistical-thermodynamic basis for computation of binding affinities: A critical review. *Biophysics Journal* **72**: 1047-1062.

Gitti, RK., B.M. Lee, J. Walker, M.F. Summers, S. Yoo and W.I. Sundquist, Structure of the amino-terminal core domain of the HIV-1 capsid protein. *Science* **273** (1996), pp. 231–235.

Good AC, Ewing TJ, Gschwend DA, Kuntz ID. 1995. New Molecular shape descriptors: application in database screening. *Journal of Computer Aided Molecular Design* 9(1): 1-12.

Golke H, Hendlich M, Klebe G. 2000. Knowledge-based scoring functions to predict protein-ligand interactions. *Journal of Molecular Biology* 295:337-356.

Goodford, PJ. 1985. A computational procedure for determining energetically favorable binding sites on biologically important macromolecules. *Journal of Medicinal Chemistry* Volume 28: 849-857.

Goodsell DS, Olson AJ. 1990. Automated docking of substrates to protein by simulated annealing. *Proteins* **8**:195-202.

Gross, I. *et al.*, 1998. N-terminal Extension of Human Immunodeficiency Virus Capsid Protein Converts the In Vitro Assembly Phenotype from Tubular to Spherical Particles. *Journal of Virology* Volume 72:4798-4810.

Gross *et al*., 2000.A conformational switch controlling HIV-1 morphogenesis. *European Molecular Biology Organization Journal* Volume**19**: 103-113

Gschwend DA, Kuntz ID. 1996. Journal Computer Aided Molecular Design 10: 123.

Gordon K, & Balasubramanian, S., 1999. Recent advances in solid-phase chemical methodologies. *Current Opinions in Drug Discovery* **2**:342-349

Halfon, P. *et al.,* 2003. Kinetics of disappearance of resistance mutations and reappearance of wild-type during structured treatment interruptions *Acquired Immune Deficiency Syndrome* Volume **17(9):** 1351-1361.

Halperin I, Buyong M, Wolfson H, Nussinov R, 2002. Principles of Docking: An overview of Search Algorithms and a Guide to Scoring Functions. *Proteins: Structure Function and Genetics* **47**: 409-443

Haque, TS., *et al.,* 1999. Potent, low molecular weight no-peptide inhibitors of malarial aspartyl protease plasmepsin II. *Journal of Medicinal Chemistry* Volume**42**: 1425-1440

Hendrix DK, Kuntz ID. 1998. Surface solid angle-based site points for molecular docking. *Pac Symp.. Biocomputers* 317-326

Henderson, LE *et al.*, 1992. Gag proteins of the highly replicative MN strain of human immunodeficiency virus type 1: Post-translational modifications, proteolytic processings, and complete amino acid sequences. *Journal of Virology* Volume 66: 1856-1865.

Hindle SA, Rarey M, Bunning C, Lengaue T. 2002. Flexible docking under pharmacophore type constraints. Journal of Computer Aided Molecular Design 16(2): 129-149.

Hou T, Wang J, Chen L, Xu X. 1999. Automated docking of peptides and proteins by using genetic algorithms combined with tabu search. *Protein Engineering* 12: 639-647.

Houston, JG. & Banks, M., 1997. Developments in automated and miniaturized screening technology. *Current Opinions in Biotech* volume**8**:734-740

Hoglund, S., *et al.*, 1990. Analysis of the assembly of the HIV core by electron microscope tomography, p149-157. *In* L. H. Pearl (ed.) Retroviral protease: control of maturation and morphogenesis. Stockton Press, New York.

Huang H, Chopra R, Verdine GL, Harrison SC. 1998. Structure of a covalently trapped catalytic complex of HIV-1 reverse transcriptase: Implications for drug resistance. Science 282(5394): 1669-1675.

Jacobson *et al.*, 2000 Single dose safety, pharmacological, and antiviral activity of the human immunodeficiency virus (HIV) type 1 entry inhibitor PRO 542 in HIV-infected adults. *Journal of Infectious Disease* volume 182:326-329

Janin J, Chervils J. 1993. Protein docking algorithms: simulating molecular recognition. *Current opinions in structural biology* 1993. **3**:265-269

Jetz., *et al.,* (2002). High rate of recombination throughout the human immunodeficiency virus type 1 genome. *Journal of Virology* Volume 74;1234-1240.

Jiang F, Kim SH. 1991. "Soft Docking" matching of molecular surface cubes. *Journal of Molecular Biology.* 219: 79-102

Johnson ME *et al.,* 1994. Conformational rearrangements required of the V3 loop of HIV-1 gp120 for proteolytic cleavage and infection. *Federation of European Biochemical Studies* volume 337(1):4-8

Jones G, Willett P, Glen RC, Leach AR, Taylor R. 1997. Development and validation of a genetic algorithm for flexible docking. *Journal of Molecular Biology* **267**:727-748.

Jones, IM and Yuko, Morikawa.1998. The molecular basis of HIV Capsid Assembly *Reviews in Medical Virology* Volume 8:87-95

Jordan, R., Gold L., Cummins, C, and Hyde C. 2002. Systematic review and meta-analysis of evidence for increasing numbers of drugs in antiretroviral combination therapy. *British Medical Journal* V324 pp1-10

Joseph-McCarthy D, Thomas BE, Belmarsh M, Moustakas D, Alvarez JC. 2003. Pharmacophore-based molecular docking to account for ligand flexibility. Proteins: Structure, Function and Genetics Volume 51: 172-188.

Joseph-McCarthy D, Alvarez JC. 2003. Automated Generation of MCSS-Derived Pharmacophoric DOCK site points for searching multi-conformational databases. *Proteins: Structure, Function and Genetics* Volume 51: 189-202.

Kick, EK., *et al.,* 1997. Structure-based design and combinatorial chemistry yield low nanomolar inhibitors of cathepsin D. *Chemical Biol* Volume 4: 297-307.

Kirkpatrick DL, Watson S, Ulhaq S. 1999. Structure based drug design: combinatorial chemistry and molecular modeling. *Comb. Chem. High Throughput Screening* 4:211-221

Kearsly SK, Underwood DJ, Sheridan RP, Miller MD. 1994. Flexibases: a way to enhance the used of molecular docking methods. *Journal of Computer Aided Molecular Design* **8**:565-82

Koh, John T. 2003; Making virtual screening a reality. *PNAS* Volume 100 No. 12 pp 6902-6903.

Komai T. *et al.*, 1997. Development of HIV-1 protease expression methods using the T7 phage promoter system. *Applied Microbiology Biotechnology* 47:241-245

Kramer B, Rarey M, Lenguaer 1999. Evaluation of the FlexX incremental construction algorithm for protein-ligand docking. *Proteins: Structure, Function and Genetics* 37: 228-241.

Kuby, J. 1994. HIV structure and infectious process pp523-555. Immunology. W.H. Freeman & Co. New York, NY. Second edition. Library of Congress.

Kuntz ID, Blaney JM, Oatley SJ, Langridge R. 1982. A geometrical approach to macromolecular interactions. **161**: 269-88

Lamb, ML., 2001. Design, Docking, and Evaluation of Multiple Libraries Against Multiple Targets. *Proteins* **42**:296-318

Leach AR, Kuntz ID. 1992. Conformational analysis of flexible ligands in macromolecular receptor sights. *Journal of Computational Chemistry* **13**:730-748

Leibowitz N, Fligelman Z, Nussinov R, Wolfson HJ. 2001. An automated multiple structural alignment and detection of a common substructural motif. *Proteins* 43: 235-245.

Lengauer T, Rarey M. 1996. Computational methods for biomolecular docking. Current opinions in structural biology 6: 402-406.

Leuthardt A and Roesel J. 1993. Cloning, expression and purification of a recombinant poly-histidine-linked HIV-1 protease. *Federation of European Biochemical Societies* Volume 326:275-280.

Lewis W, Day BJ, Copeland WC. Mitochondrial toxicity of NRTI antiviral drugs: an integrated cellular perspective. *Nat Rev Drug Discov*. 2003 Oct; 2(**10**):812-22

Lin SL, Nussinov R, Fischer D, Wolfson HJ. 1994. Molecular surface representation by sparse critical points. *Proteins*. 18:94-101

Lin SL, Nussinov, R 1996. Molecular recognition via face-center representation of a molecular surface. *Journal of Molecular Graphics*. 14: 78-90

Liu M, Wang S. 1999. MCDOCK: a Monte Carlo simulation approach to the molecular docking problem. *Journal of Computational Aided Molecular Design* Volume 13: 435-51.

Liu S, Fan, S and Sun, Z. 2003. Structural and functional characterization of the human CCR5 receptor in complex with HIV gp120 envelope glycoprotein and CD4 receptor by molecular modeling studies. *Journal of Molecular Modeling* (online) 2003 Aug 29.

*Luban, J., et al., 1993. Human immunodeficiency virus type 1 Gag protein binds to cyclophilinA A and B.* Cell *Volume 73, Issue 6, Pages 1067-1078*

Luty BA, Wasserman ZR, Stouten PFW. 1995. A molecular mechanics/grid method evaluation of ligand-receptor interactions. *Journal of Computational Chemistry* **18**: 94-101

Makino S. Ewing TJ, Kuntz ID. 1999. DREAM++: flexible docking program for virtual combinatorial libraries. *Journal of Computational Aided Molecular Design* 13(5): 513-532.

Marchand *et al.,* 2003. Metal-dependent inhibition of HIV-1 integrase by beta diketo acids and resistance of the soluble double-mutant (F185K/C280S) *Molecular Pharmacology* Volume **3:**600:609.

Marrone, TJ., 1997. Structure-based drug design: computational advances. *Annual Review of Pharmacology and Toxicology* Volume **37**: 71-90

Mason, JS, Beno., BR. 2000. Library design using BCUT chemistry-space descriptors and multiple four-point pharmacophore fingerprints: simultaneous optimization and structure based diversity. *Journal of Molecular Graphical Models* (4-5): 438-451.

Maurin, C. Bailly, F., Cotelle, P., (2003) Structure-activity relationships of HIV-1 integrase inhibitors-enzyme-ligand interactions. *Current Medicinal Chemistry* Volume (18): 1795-1810.

McPhee F, Good A, Kuntz ID, Craik C. 1996. Engineering human immunodeficiency virus 1 protease heterodimers as macromolecular inhibitors of viral maturation. *PNAS Online.* Volume **93**, Issue 21, 11477-114781

Melnick L *et al.,* 1998. An *Escherichia coli* Expression assay and screen for Human Immunodeficiency Virus Protease Variants with Decreased Susceptibility to Indinavir. *Antimicrobial Agents and Chemotherapy* Volume 42:3256-3265

Meng EC, Stoichet BK, Kuntz ID. 1992. Automated docking with grid-based energy evaluation. *Journal of Computational Chemistry* 13:505-524

Mervis, RJ., *et al.,* 1988. The Gag gene products of human immunodeficiency virus type 1: Alignment within the Gag open reading frame, identification of post-translational

modifications, and evidence for alternative Gag precursors. *Journal of Virology* Volume 62: 3993-4002.

Miller MD, Sheridan RP, Kearsy SK. 1999. SQ. A program for rapidly producing pharmacophorically relevant molecular superpositions. Journal Med. Chem. 42: 1505-1514.

Miller MD, Kearsley SK, Underwood DJ, Sheridan RP. 1994. FLOG: a system to select 'quasi-flexible' ligands complementary to a receptor of known three-dimensional structure. *Journal of Computer Aided Molecular Design* Volume **8**: 153-174.

Mitchell JBO, Laskowski RA, Alex A, Thornton JM. 1999. BLEEP – a potential of mean force describing protein-ligand interactions: Generating the potential. Journal of Computational Chemistry Volume 20: 1165-77.

Moebius U et al., 1992.The human immunodeficiency virus gp120 binding site on CD4: Delineation by quantitative equilibrium and kinetic binding studies of mutants in conjunction with a high-resolution CD4 atomic structure. *Journal of Experimental Medicine* Volume 2:507-517

Momany, C. *et al.,* 1996.Crystal structure of dimeric HIV-1 capsid protein. *Nature structural biology.* Volume 3 number 9:763-770

Meugge I, Martin YC. 1999. A general and fast scoring function for protein-ligand interactions: A simplified potential approach. *Journal Med. Chem*. 42: 791-804

Murray CW, Baxter CA, Frenkel AD. 1999. The sensitivity of the results of molecular docking to induced effects: application to thrombin, thermolysin and neuraminidase. *Journal of Computer Aided Molecular Design* **13**:547-562

Murray CW, Clark DE, Auton TR. 1997. PRO-SELECT: combining structure based drug-design and combinatorial chemistry for rapid lead discovery. *Journal of Computer Aided Molecular Design* **11**: 193-207

Nermut, M.V. and Hockley, D.J. .1996. Comparative morphology and structural classification of retroviruses. *Curr. Top. Microbiol. Immunol.*, **214**, 1-24

Norel R, Lin SL, Wolfson HJ, Nussinov R. 1994. Shape complementarity at protein-protein interfaces. *Biopolymers* 34: 933-940.

Ott, DE., *et al.,* 1997. Analysis and localization of cyclophilin a found in the virions of human immunodeficiency virus type 1 MN strain. *AIDS Research and Human Retroviruses.* Volume 9 Issue 11 1003-1006

Pang Y-P, Perola E, Xu K, Prendergast FG. 2001. EUDOC: a computer program for identification of drug interaction sites in macromolecules and drug leads from chemical databases. *Journal of Computational Chemistry* **22**: 1750-1771

Pattabiraman N, Levitt M, Ferrin TE, Langridge R. 1985. Computer graphics in real-time docking with energy calculation and minimization. *Journal of Computational Chemistry* 6: 432-436.

Perola, E., *et al.,* (2000) Successful virtual screening of a chemical database for farnesyltransferase inhibitor leads. *Journal of Medicinal Chemistry* Volume **43**(3):401-408.

Rao, GS, Bhatnagar, S., Ahuja, V. 2002. Structure-based design of a novel peptide inhibitor of HIV-1 integrase: a computer modeling approach. *Journal of Biomolecular Structural Dynamics* Volume 1:31-38

Rarey M, Wefing S, Lengauer T. 1996. Placement of medium-sized molecular fragments into active sites of proteins. *Journal of Computer Aided Drug design* **10**:41-54

Reicin, A.S., A. Ohagen, L. Yin, S. Hoglund and S.P. Goff. 1996. The role of Gag in human immunodeficiency virus type 1 virion morphogenesis and early steps of the viral life cycle. *Journal of Virology* **70** pp. 8645–8652.

Reynolds CA, King PM, & Richards WG. 1992. *Molecular Physics* Volume 76, 251-275

Rigoutsos I, Platt D, Califano A. 1996. Flexible 3-D substructure matching and novel conformer derivation in very large databases of 3-D molecular information. IBM Research Division. Yorktown Heights NY. Watson Research Center.

Rimsky, LT, Shugars, DC, Matthews, TJ, 1998. Determinants of human immunodeficiency virus type 1 resistance to gp41-derived peptides. *Journal of virology* 72(2): 986-993.

Ritchie DW, Kemp GJL. 2000 Protein docking using spherical polar Fourier calculations. *Proteins* 39:178-194

Roche O, Kiyama R, Brooks CL. 2001. *Journal of Medicinal* Chemistry 44: 3592

Rose JR, Babe LM, Craik CS, Defining the level of Human Immunodeficiency Virus Type 1 (HIV-1) Protease Activity Required for HIV-1 Particle Maturation and Infectivity. *Journal of Virology* Volume69:2751-2758.

Rossmann, MG., 1988. Antiviral agents targeted to interact with viral capsid proteins and a possible application to human immunodeficiency virus. *Proc. Natl. Acad. Sci. U.S.A.* **85** pp. 4625–4627.

Sadowski J, Rudolph C, Gasteiger J. 1990. Automatic generation of three-dimensional atomic coordinates for organic molecules. *Tetrahedron Computer Methodologies* 3: 537-547

Schwedler, UK,. Stemmler, T.L. Klishko, V.Y 1998. Proteolytic refolding of the HIV-1 capsid protein amino-terminus facilitates viral core assembly. *EMBO J.* **17**, pp. 1555–1568

Siddiqui MI, Kataria S, Ahuja V, Rao GS A peptide inhibitor of HIV-1 protease using alpha, beta- dehydro residues: a structure based computer model. *Indian Journal of Biochem Biophys.* **2001** Feb-Apr;38(1-2):90-5.

Silvestri R, De Martino G, La Regina G, Artico M, Massa S, Vargiu L, Mura M, Loi AG, Marceddu T, La Colla P. 2003. Novel indolyl aryl sulfones active against HIV-1 carrying NNRTI resistance mutations: synthesis and SAR studies. *J Med Chem*. 2003 Jun 5; **46(12):** 2482-93.

Smith, TJ, M.J. Kremer, M. Luo, G. Vriend, E. Arnold, G. Kamer, M.G. Rossmann, M.A. McKinlay, G.D. Diana and M.J. Otto, The site of attachment in human rhinovirus 14 for antiviral agents that inhibit uncoating. *Science* **233** (1986), pp. 1286–1293.

Stoichet BK, Leach AR, Kuntz ID. 1999. Ligand solvation in molecular docking. *Proteins.* **34**: 4-16.

Stone, Alan (2003) MICROBICIDES: A NEW APPROACH TO PREVENTING HIV AND OTHER SEXUALLY TRANSMITTED INFECTIONS, *Nature Reviews Drug Discovery* **1**, 977 -985 (2002); doi:10.1038/nrd959

Sun, Y., *et al.,* 1998. CombiDOCK: structure-based combinatorial docking and library design. *Journal of Computational Aided Molecular Design* **12**:597-604

Suto, MJ. 1999. Developments in solution-phase combinatorial chemistry. *Current Opinions in Drug Discovery* Volume**2**: 377-384.

Tame, JRH. 1999. Scoring functions: a view from the bench. *Journal of Computer Aided Molecular design* 13: 99-108

Tang, C. Tang C, Loeliger E, Kinde I, Kyere S, Mayo K, Barklis E, Sun Y, Huang M, Summers MF**.** 2003. Antiviral Inhibition of the HIV-1 Capsid Protein. *Journal of Molecular Biology* Volume 327:1013-1020.

Tang, S., T. Murakami, B.E. Agresta, S. Campbell, E.O. Freed and J.G. Levin. 2001. Human immunodeficiency virus type 1 N-terminal capsid mutants that exhibit aberrant core morphology are blocked in initiation of reverse transcription in infected cells. *Journal of Virology.* **75** (2001), pp. 9357–9366.

Taylor RD, Jewsbury PJ, Essex JW. 2002. A review of protein-small molecule docking methods. *Journal of Computer Aided Molecular Design* 16: 151-166

Thali, M., *et al.,* 2002. Functional association with cyclophilin A with HIV-1 virions. *Nature* Volume 372: 363-365

Tomioka N, Itai A, Iitaka Y. 1987. A method for fast energy estimation and visualization of protein-ligand interaction. *Journal of Computer Aided Molecular Design*. **1**:197-210.

TotRov R, Abagyan R. 1998. Flexible ligand-protein docking by global energy optimization in internal coordinates. *Proteins* (Suppl.) 1:215-20.

Trosset J, Scheraga HA. 1999. Prodock: Software package for protein modeling and docking. *Journal of Computational Chemistry*. 20:412-427.

Turkington, Carol, Non-nucleoside reverse transcriptase inhibitors. *Gale Encyclopedia of Medicine*

Verdonk ML, Cole JC, Hartshorn, MJ, Murray CW, Taylor R. 2003. Improved Protein-ligand docking using GOLD. *Proteins: structure, function and Genetics* 52:609-623

Vieth M, Hirst JD, Kolinski A, Brooks CL, 1998. *Journal of Computational Chemistry* 19: 1612

Vzorov AN, Marzilli LG, Compans RW, Dixon DW. 2003. Prevention of HIV-1 infection by phthalocyanines *Antiviral Res.* 2003 Jul;**59(2)**:99-109.

Walters WP, Stahl MT, Murcko MA. 1998. Virtual Screening, an overview. *Drug Discovery Today* **3**:160-178

Wang R, Lu Y, Wang S., 2003. Comparative evaluation of 11 scoring functions for molecular docking. *Journal of Medicinal Chemistry* Volume 12: 2287-2303

Wang J. *et al*., 2001. Use of MM-PBSA in Reproducing the Binding Free Energies to HIV-1 RT of TIBO Derivatives and Predicting the Binding Mode to HIV-1 RT of Efavirenz by Docking and MM-PBSA. *Journal of American Chemical Society* 1223:5221-5230

Weber *et al.,* 2003. A novel TaqMan real-time PCR assay to estimate ex vivo human immunodeficiency virus type 1 fitness in the era of multi-target (*pol* and *env*) antiretroviral therapy *Journal of General Virology* Volume 84:2217-2228

Wilk, T and Fuller S, 1999. Towards the structure of the human immunodeficiency virus: divide and conquer? *Current Opinion in Structural Biology* Volume 9:231-243

Wills, JW & Craven, RC, 1991. Form, function and use of retroviral Gag proteins. *AIDS* Volume 5:639-654.

Wu G, Robertson DH, Brooks C, Veith M. 2003. Detailed Analysis of Grid Based Molecular Docking: A Case study of CDOCKER – A CHARm-Based MD Docking Algorithm. *Journal of Computational Chemistry* Volume 24(13): 1549-1562

Zauhar, Randy J. Guillermo Moyna, LiFeng Tian, ZhiJian Li, William J. Welsh Shape Signatures: A New Approach to Computer-Aided Ligand- and Receptor-Based Drug Design *J. Med. Chem.,* **46** (26), 5674 -5690, 2003

Zhang LY, Gallicchio E, Friesner RA & Levy RM. 2001. *Journal of Computational Chemistry* Volume 22, 591-607.

Zhou J, Yuan X, Dismuke D, Forshey BM, Lundquist C, Lee KH, Aiken C, Chen CH. Small-molecule inhibition of human immunodeficiency virus type 1 replication by specific targeting of the final step of virion maturation. *J Virol.* 2004 Jan;**78(2)**:922-9.

Zhou Z, Madrid, M, Madura, J. 2002. Docking of Non-nucleoside Inhibitors: Neotripterifordin and its Derivatives to HIV-1 Reverse Transcriptase *Proteins* 49:529-542.

Zhu J, Fan H, Liu H, Shi Y. Structure-based ligand design for flexible proteins: application of new F-DycoBlock *J Comput Aided Mol Des*. 2001 Nov;**15(11**):979-96

Zollner, B. *et al*., Primary genotypic resistance of HIV-1 to the fusion inhibitor T-20 in long-term infected patients. *AIDS* Volume 15, pp935-936.

Zybarth G, *et al.*, 1994. Proteolytic activity of novel human immunodeficiency virus type 1 proteinase proteins from a precursor with a blocking mutation at the N-terminus of the PR domain. *Journal of Virology Volume* 68(1):240-250.

# Appendix A

## Computational Methods

The entire capsid protein was downloaded from Rutgers University's Research Collaboratory for Structural Bioinformatics's Protein Data Bank (http://www.rcsb.org/pdb/). Structure Explorer search for molecule "1afv" resulted in the HIV-1 Capsid Protein (p24) Complex with Fab25.3, March 1997 (C. Momany *et al*,). The downloaded ".pdb" file was then transferred to the SGI Mainframe "xena" (/rx/home/houstonj/xena) for manipulation by Sybyl, a user-friendly molecular modeling package with capabilities for molecular mechanics, conformational searching, minimization, semi-empirical (interface to MOPAC) and ab initio MO calculations (interface to Gaussian98), molecular graphics, active analog approach (Created Sept 17, 1998. Copyright 1991-1998, Tripos Inc., All Rights Reserved).

The subtraction of the first 13 residues was performed using the ".jot" files incorporated in the SGI mainframe (simple cut and paste *or* highlight and delete). A separate file was created using the initial 13 residues of the capsid and named "ligand.pdb'. The Subdirectory /Dissertation was created within xena/home/houstonj to house all receptor, ligand, input and output files.

DOCK 4.0 is actually a suite of sub-programs, each of which is briefly albeit awkwardly described in the accompanying loose-leaf manual. Progressing though these programs in an organized manner channels the user through the four basic stages: Ligand Preparation, Site Characterization, Scoring Grid Calculation and finally, Docking.

**The Overall Program Sequence is:**

- get_near_res
- invertPDB
- autoMS
- sphgen
- showsphere
- showbox
- grid -i
- dock –i

***Ligand Preparation***

Initial set-up requires a program to create a file containing all atoms of all residues in the receptor that have their Alpha-carbon within a user-specified distance of the ligand. The program 'get_near_res' does just that: *G*ets all the *N*earby *R*esidues and allots them to a file. In addition, it creates a list file that gives the closest Alpha-carbon to ligand distance for each residue written.

The command was typed at the UNIX prompt. The user was prompted for subsequent answers. The *small file* is the 'ligand.pdb' file created by the subtraction of the initial 13 residues from the original PDB download. The *file to find nearby residues* is the 'receptor.pdb' file. Note: if both files are not in the PDB format (possibly they are '.mol2' format or a format stemming from Rasmol) then converting each file into PDB must be accomplished previous to get_near_res using the program 'convsyb'. The user was then prompted to set the number of Angstroms. The choice was made for **ten,** well within the usual range of between 8 and 15. Option '1' was chosen to judge *any* atom within the angstrom cutoff set in previous answer. The option to include all hydrogens in the calculations was declined.

The name chosen for the resulting *outputs* files was 'keep_exclude'. The name derives from the tendency to misinterpret exactly *which* files *will* be used for further manipulations.

‘Keep_exclude’ signifies the portion of the receptor that is closest to the ligand and will indeed be required in later sub-programs. Note: once this program (get_near_res) reads your input files, the resulting screen data should have a larger amount of ‘atoms read in’ for the protein file in comparison to the ligand file. Use the ‘charge’ command to ensure ‘keep_exclude.pdb’ has a make-up in which residue differences are detectable from the original receptor. Verification should also be done using visualization with Sybyl.

***Site Characterization***

InvertPDB: In addition to the creation of a sub-component of the receptor to be used in docking algorithms and scoring functions, the need for a file including the *rest* of the receptor that will *not* be used in DOCK calculations still exists. The sub program ‘invertPDB’ is a shell script to extract all atoms in a larger PDB file that are not in a smaller PDB file, where the latter is a subset of the former (DOCK manual). The purpose for the creation of the file and hence, the function of this program (‘invertPDB’ signifies the *inverse* PDB atoms of the ones generated in get_near_res) is that an output is needed in for calculations of the surface of the receptor, to be accomplished by a later DOCK sub-program. The command line reads:

- xena 193%  invertPDB “receptor_name”.pdb  keep_exclude.pdb > exclude.pdb

This gave output file ‘exclude.pdb’, which was utilized in the surface calculations of the capsid protein.

The script AutoMS converts PDB files to QCPE MS input format, runs a QCPE MS surface calculation, converts the resulting surface to UCSF MS format with reformatms, creates a Sybyl dot file of the surface (if ms2dot is available), and prepares an INSPH file for running

SPHGEN (DOCK manual). Two files were needed for input into this sub-program: the receptor.pdb file and the exclude.pdb file created from the invertPDB command. The program will automatically search for the file 'exclude.pdb' within the current working subdirectory. AutoMS symbolizes the command to automatically calculate the molecular surface of the protein and making sure not to include the exclude.pdb atoms in generation of a sphere file. The command line reads:

- xena 194% autoMS receptor.pdb

The defaults on autoMS included: 3.0Angstroms for surface density, and 1.4Angstroms for probe radius.

"…what's good about DOCK is that it uses spheres; what's bad about DOCK is that is uses spheres…" – unknown source, DOCK manual

Spheres were generated to fill the target site and each sphere center represented a putative ligand atom position. The use of spheres in general is an attempt to limit the enormous number of possible orientations within the active site. Spheres touch the surface of the molecule without intersecting that surface – just like ligand atoms. Overall, DOCK spheres are allowed to intersect other spheres – their volumes overlap.

For receptors, a negative image of the surface invaginations was created; for a ligand, the program created a positive image of the entire molecule. Spheres were constructed using the molecular surface first described by Richards (1977) and calculated with a program created by Connolly (1983). Each sphere touches the molecular surface at two points and has its radius along the surface normal of one of the points. For the receptor, each sphere center is "outside" the surface, and lies in the direction of a surface normal vector. For a ligand, each sphere center

is “inside” the surface, and lies in the direction of a surface normal vector. Spheres are calculated over the entire surface, producing about one sphere for every surface point. Since this leads to a very dense representation, there are filters in place to ensure that only the largest sphere associated with each receptor atom is kept. Each invagination of the receptor surface is characterized by a “set” of overlapping spheres. The sets are called clusters by DOCK, and are sorted out by descending size of the number of spheres in a set. Each one of the clusters is an attempt by sphgen to generate a sphere set that truly represents the original ligand. The largest set is the all-inclusive cluster 0 (zero) the largest *individual* cluster set is 1. Cluster number 2 is the next to largest, leading up to clusters n – which is the cluster set with the smallest number of spheres. Spheresets in the higher number ranges usually have declining efficiency at representing the pocket, therefore sphereset 2 is generally more likely to be used than say, sphereset 11. The sub-program showsphere helps to visualize these clusters and will be discussed later – after the *sph*eres have actually been *gen*erated.

The input deck for *sph*ere *gen*eration (*sphgen*) was the file created by autoMS: INSPH. This file was modified using the ‘.jot’ program.

The variables of note are the molecular surface file ‘*msfil’*, and the adjustments files ‘*radmin’* and ‘*radmax*’. Although the defaults were used, manipulating *msfil* alters the molecular surface calculated by autoMS. All adjustments must be made in a Fortran-compatible format for further usage by sphgen. Adjusting the sphere radius in angstroms is done with radmin and radmax. Radmax has a default of 4.0A that was adjusted to 4.5A to increase the potential overlapping spheres in the Pro1 region. Radmin was kept at its default of 1.4, with spheres with a smaller radii being discarded from consideration.

The command SPHGEN is typed as lowercase as the line reads:

- xena 263% sphgen

Note: If the set-up through the previous sub-programs was done correctly up to this point, sphgen takes 12-20 seconds of CPU time to complete sphere generation. SPHGEN output has two key components: OUTSPH and "receptor.sph". The final line of the OUTSPH file (by cat or jot) showed exactly how many cluster set were created by the sphgen run. "Clustering is complete" followed by the *five numbers of clusters* was the output listed by OUTSPH, with cluster set 0 (zero) utilized for creating the final sphere set used in subsequent grid and dock runs.

The receptor.sph file was converted into a .pdb file and in order to be viewed by Sybyl by the sub-program showsphere.

- xena 212% showsphere

Showsphere requested the name of the sphere cluster file: receptor.sph. It then requested which of the available subsets (from zero to n) the user wished to turn into a pdb file. Cluster zero was chosen. The request was denied to generate an ms file. The name of the pdb file was a reflection of the cluster number derivation – 'spheresetzero.pdb'. Showsphere completes it tasks by returning the name of the cluster set and which sphere comprise that particular cluster.

Spheresetzero.pdb however includes all of the spheres generated by the sphgen calculations. Upon SYBYL visual inspection, *none of the further subsets of clusters gave an accurate description of the ligand in the appropriate pocket, despite the manipulation of*

*variables within subprograms leading to this point*. Therefore, the sphereset to be used in further grid and dock procedures was done manually within Sybyl.

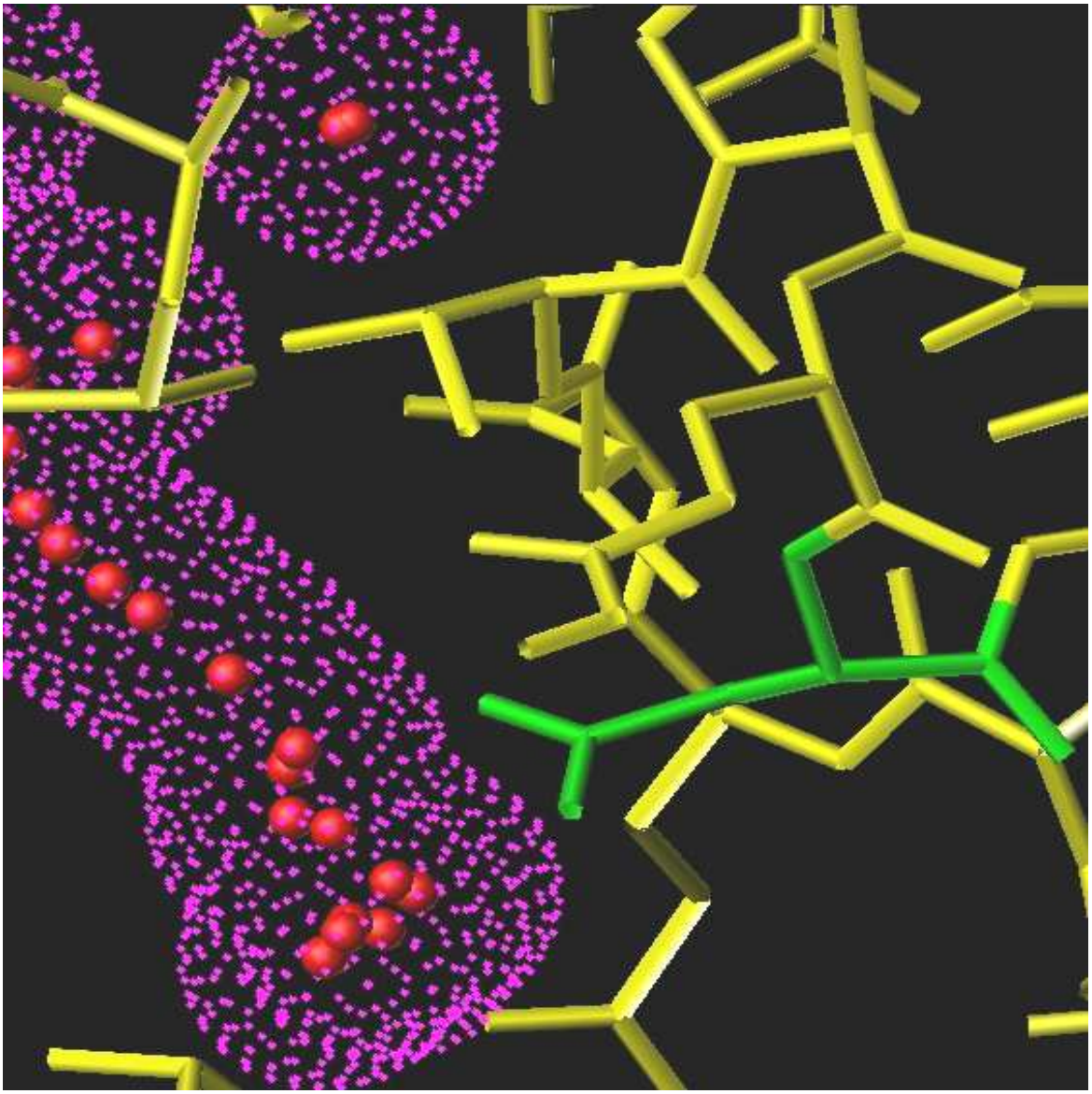


**Figure A.2 – van der Waals' radii around the calculated spheres. Asp51 in close proximity.**

Figure A.2 has these same spheres within the pocket, surrounded by a dot surfaces file representing a van der Walls' radius (magenta) around each sphere (generated from Sybyl with

*View/Dot Surfaces/vdw Dot Surface* and naming/highlighting the output dot file at the user's discretion). Molecular details of the sphere's vdw radius within the pocket are visualized more concretely with the magnified view shown in Figure 5. The proximity and electron donating potential of Asp51, as well as the aggregate number of spheres in the pocket space vacated by Pro1 are unmistakably discernable. The illustration of spheres representing *only* the selected portion of the ligand that fits tightly into the pocket is also demonstrated in figures 3.3-3.6. The stick figure of the ligand in figure 3.6, against the backdrop of the space-fill keep_exclude region, clearly contains segments that are not affiliated with any red-colored spheres.

Making the final sphere representation by hand is tedious; however it is a *crucial* step in the overall docking procedure in order to clearly represent the original ligand *within* the pocket of choice.

Protocol for generating the final sphere representation of the ligand:

- The keep_exclude.pdb file is read into Sybyl
- The all-inclusive sphere file (allspheres.pdb) is then read into Sybyl.
- The ligand file is read into Sybyl
- With the combined files clearly viewable on screen, the sequential window pull-down commands： Build-Edit/Delete/Atom should create an Atom Expression box.
- Within this Atom Expression box, M2 represents the sphere file and is highlighted with the right mouse button.
- Right clicking on spheres that are not representative of the ligand in the N-terminal pocket, while periodically rotating the molecule complex eliminates them (once OK is hit) from the file. This method is preferable over '.jot' manipulations, as the deleted spheres will not be individually or collectively visualized while performing this sensitive operation.

The final, adjusted sphere file was given the name 'workspheres.pdb'.

***Scoring grid calculation***

Showbox is an interactive sub-program that gives the user the capability to visualize the location and size of the grids that will be calculated by the program '*grid*'.

- xena 133% showbox

The initial prompt is whether the box should be constructed automatically and encompasses all of the spheres in that cluster. This user answered yes and by doing so was prompted to specify the number of Angstroms as a value for how closely the box faces may approach a sphere center. Three was the value. The cluster file name was 'workspheres.pdb' and the re-created sphere file of 0 was the cluster number answer. The final program prompt is for the name of the output file: 'site_box.pdb'. The output file was immediately available for visualization in Sybyl.

GRID creates the files that are required for high-speed evaluation/scoring in DOCK. The only type of scoring that was utilized in this project was the energy-based scoring (contact and chemical scoring are the others). This sub-program also calculates whether a ligand atom is in severe steric overlap with a receptor atom – a bump grid. This problem is identified when any *.bmp extension file is created. This is set up in such a manner that prior to scoring; each orientation is processed with the bump filter to reject ones that penetrate deep into the receptor. Orientations that pass through this bump filter are then allowed to be evaluated for scoring.

The calculations involving grid are required previous to the final step of docking. If done properly, Grid will take 2-5 minutes to accomplish 100% of its processing (although the creators of DOCK allow for up to 45 minutes of CPU time).

The command line:

- xena 31% grid –i  Final

With ‘Final’ being the prefix the user specifies that all output files will have. Final will also be the name of the grid file, once completed (see figure 3.8).

Figure A.3 shows the UNIX-based prompting once the command ‘grid’ is implemented. Noticeable from this list is that contact and chemical scoring grids have both been denied. An energy based scoring function was the only scoring function used during the course of this project, as this is the best way to ascertain potential lead compounds on a pharmacophore bases. The majority of these responses are from the defaults given by the program *e.g.*, using 10Angstroms as the energy cut-off distance. Answers that deviated from the defaults included a ‘yes’ to output_molecule and ‘yes’ to the bump filter.

Figure A.4 illustrates the final output file after the UNIX command grid –i Final. This file is used as a reference for the user to understand all variable inputs for the docking run. Of note are the energy parameters detailing the translational, rotational and convergence of the ligand.

```
_____________General_Parameters_____________
compute_grids                  [NO yes] y
grid_spacing                   [0.3]
output_molecule                [NO yes] y

_________________Scoring_Parameters_________________
contact_score                  [NO yes] n
chemical_score                 [NO yes] n
energy_score                   [NO yes] y
energy_cutoff_distance         [10.0]
atom_model                     [UNITED all]
attractive_exponent            [6]
repulsive_exponent             [12]
distance_dielectric            [YES no]
dielectric_factor              [4.0]
bump_filter                    [NO yes] y

_____________________File_Input_____________________
receptor_file                  [receptor.mol2] receptor.mol2
box_file                       [site_box.pdb] site_box.pdb
vdw_definition_file            [/xena/home/dock/4.0.1/parameter/vdw.defn]

_____________________File_Output_____________________
score_grid_prefix              [Final]
receptor_out_file              [receptor_out.mol2]
```

**Figure A.3 – UNIX based prompting in the grid –i command**

The output *grid* file 'Final':

```
flexible_ligand                yes
orient_ligand                  no
score_ligand                   yes
minimize_ligand                yes
multiple_ligands               no
random_seed                    0
anchor_search                  no
torsion_drive                  yes
clash_overlap                  0.5
conformation_cutoff_factor     5
torsion_minimize               no
write_conformations            no
intramolecular_score           yes
intermolecular_score           yes
gridded_score                  yes
grid_version                   4
contact_score                  no
chemical_score                 no
energy_score                   yes
energy_cutoff_distance         10
distance_dielectric            yes
dielectric_factor              4
attractive_exponent            6
repulsive_exponent             12
atom_model                     u
vdw_scale                      1
electrostatic_scale            1
energy_minimize                yes
initial_translation            1
initial_rotation               0.1
maximum_iterations             100
energy_convergence             0.1
maximum_cycles                 1
score_grid_prefix              Final
vdw_definition_file            /xena/home/dock/4.0.1/parameter/vdw.defn
flex_definition_file           xena/home/dock/4.0.1/parameter/flex.defn
flex_drive_file          xena/home/dock/4.0.1/parameter/flex_drive.tbl
```

**Figure A.4 the output file Final**

### *Multi-Server/Client Docking from a Database*

- xena 77% Dock –i Final

This last command of the Dock sequence creates the input deck. As a test, the docking of one user-choice compound (versus the target defined with Site Characterization and using the

scoring function calculated with GRID) was completed. The command line read: dock –i Test1.inp. Figure A.5 displays the prompts the user must satisfy in order to begin docking.

```
      UUUUUUUUU     CCCCCCC     SSSSSSS    FF/   FFF/
      UU/    UU/  CC/    CC/  SS/    SS/  FF/ FFF/
     UU/    UU/  CC/    CC/  SS/         FFFFF/
    UU/    UU/  CC/    CC/  SS/          FF/ FF\
   UU/    UU/  CC/    CC/  SS/    SS/  FF/   FF\
 UUUUUUUUU/    CCCCCCC/    SSSSSSS/   FF/     FF\


University of California at San Francisco, DOCK 4.0.1
DOCK 4.0.1 was released on May 17, 1998.

____________________Job_Information____________________
launch_time                    Wed Feb  4 16:25:26 2004
host_name                      xena
memory_limit                   536870912
working_directory              /xena/home/houstonj/DockHere
user_name                      houstonj

_________________General_Parameters_________________
flexible_ligand                no
orient_ligand                  yes
score_ligand                   yes
minimize_ligand                no
multiple_ligands               yes
parallel_jobs                  no
random_seed                    0

_____________Orient_Ligand_Parameters_____________
match_receptor_sites           yes
random_search                  no
automated_matching             yes
maximum_orientations           500
write_orientations             yes
rank_orientations              yes
rank_orientation_total         1

_________________Scoring_Parameters_________________
intermolecular_score           yes
gridded_score                  yes
grid_version                   4
bump_filter                    yes
bump_maximum                   1
contact_score                  no
chemical_score                 no
energy_score                   yes
atom_model                     u
vdw_scale                      1
```

```
electrostatic_scale             1

_____________Multiple_Ligand_Parameters_____________
ligands_maximum                 1
initial_skip                    0
interval_skip                   0
heavy_atoms_minimum             0
heavy_atoms_maximum             100
rank_ligands                    yes
rank_ligand_total               100
restart_interval                100

_____________________File_Input_____________________
ligand_atom_file                /rx/xena/home/houstonj/DockHere/VEXPIX2.mol2
receptor_site_file              work4spheres.pdb
score_grid_prefix               dock
vdw_definition_file             /rx/xena/home/dock/4.0.1/parameter/vdw.defn
quit_file                       Test_1.quit
dump_file                       Test_1.dump

_____________________File_Output_____________________
ligand_energy_file              VEXPIX2_nrg.mol2
info_file                       Test_1.info
restart_file                    Test_1.rst
```

**Figure A.5 the DOCK program sequential prompting of user**

The completion of the prompts using grid or dock commands always creates a file bearing the name of the input deck. In this case, figure 3.10 represents the output file (although no 'output' suffix is attached). Of note in figures A.5 and A.6:

- Flexible ligand option is declined. Later, in follow up docking procedures among the top compounds selected, this feature will be utilized. However, the effort of a CPU to score every ligand as flexible would be gigantic. This project involved scoring ligands as rigid, but examining 500 orientation of each ligand within the pocket.
- Therefore, the orient ligand feature is accepted and the ability to write and rank each ligand, returning only the top orientation of the 500 is vital.
- Minimize ligand option is declined – this can be done in Sybyl if desired to save CPU time.

- Multiple ligands option is accepted. Even though this is a single compound docking run, we need a top 100 ranking of the orientation and the rank ligand prompt is located within the multiple ligand sub-prompts.
- Again, the only type of scoring function used is the energy-based kind. This means declining contact and chemical based functions under Scoring Parameters sub-prompt.
- VEXPIX2 is the single molecule of choice. A fairly random selection.

The subsequent file created (figure A.6) is Test1.inp, with any necessary manipulations to the file accomplished using the test program 'Jot'.

```
compute_grids                yes
grid_spacing                 0.3
output_molecule              yes
contact_score                no
chemical_score               no
energy_score                 yes
energy_cutoff_distance       10
atom_model                   u
attractive_exponent          6
repulsive_exponent           12
distance_dielectric          yes
dielectric_factor            4
bump_filter                  yes
bump_overlap                 0.75
receptor_file                receptor.mol2
box_file                     showbox_output
vdw_definition_file          /rx/xena/home/dock/4.0.1/parameter/vdw.defn
score_grid_prefix            dock
flexible_ligand              no
orient_ligand                yes
score_ligand                 yes
minimize_ligand              no
multiple_ligands             yes
intermolecular_score         yes
gridded_score                yes
grid_version                 4
vdw_scale                    1
electrostatic_scale          1
random_seed                  0
match_receptor_sites         yes
random_search                no
ligand_centers               no
automated_matching           yes
maximum_orientations         500
write_orientations           yes
```

```
ligand_atom_file                /rx/xena/home/houstonj/DockHere/VEXPIX2.mol2
receptor_site_file              good_spheres.pdb
```

```
ligand_out_file                 ligand_out.mol2
receptor_out_file               receptor_out.mol2
bump maximum                    1
bump_maximum                    1
energy_minimize                 yes
initial_translation             1
initial_rotation                0.1
maximum_iterations              100
energy_convergence              0.1
maximum_cycles                  5
cycle_convergence               1
energy_termination              1
ligand_energy_file              VEXPIX2_nrg.mol2
parallel_jobs                   no
ligands_maximum                 1
initial_skip                    0
interval_skip                   0
heavy_atoms_minimum             0
heavy_atoms_maximum             100
rank_ligands                    yes
```

**Figure A.6 - the input deck: Test_1.inp**

Docking of the molecule VEXPIX2 via the input decks led to the output seen in figure A.7. DOCK read all pertinent information, calculated energies and returned the *single* best energy score of the 500 orientations examined. CPU time allocated is also listed in the file designated by the user as 'Test_1.info' (next to last line Figure A.7).

```
Reading general grid info from dock.bmp
Reading bump grid from dock.bmp
Reading energy grids from dock.nrg
  VDW grids use a 6-12 Lennard-Jones potential with a united atom model.
  Reading attractive VDW energy grid.
  Reading repulsive VDW energy grid.
  Reading electrostatic energy grid.

_________________________Docking_Results_____________________________
Name         : VEXPIX2
Description  : Generated from the CSD
Orientations tried                                 :       5677
Orientations scored                                :        500

Best intermolecular energy score                   :     -20.62
RMSD of best energy scorer (A)                     :       5.21
```

```
Elapsed cpu time (sec)                              :       0.89


Writing restart information to disk.
Writing top scoring molecules to disk.

Finished processing molecule in 0.92 seconds.
```

**Figure A.7 Test_1.info**

However, a successfully docked single molecule within the target, while an effective verification of the previous sub-programs and mechanisms, was not the focus of this experiment. Docking multiple ligands in multiple orientations over parallel servers, with each server distributing compounds to multiple clients requires supplemental programming.

***Script-based Input decks***

For multi-ligand, server-client docking, the command dock –i (with the file name of user choice) remains the same. The majority of the prompting responses involve the same instructions - with key exceptions. Figure A.8 shows the immediate result of the dock –i command as seen within a UNIX shell. Important differences are seen in capital letters as the values normally placed directly into the master input deck are now pulled from script files. This allows the user to have one master input deck, one script that interacts with that master file and a simple 6 line input file (Camb.inp; Figure A.9) to as a feeder for both. It is a mechanism to automate the entire docking process, with as little alterations by the user as possible.

Summarizing the overall script-based docking sequence; it is execution of the script itself (the '.sh' file, figure A.10) that looks for the information in the small input file (figure A.9) and transfers that information to the master input file (seen in figure A.8). A close examination of the script file displays the initial 'read' commands searching for the information provided in the

small input file. Later in the script, ‘echo’ and ‘sed’ commands relay the input to the master.inp file.

```
compute_grids                   yes
grid_spacing                    0.3
output_molecule                 yes
contact_score                   no
chemical_score                  no
energy_score                    yes
energy_cutoff_distance          10
atom_model                      u
attractive_exponent             6
repulsive_exponent              12
distance_dielectric             yes
dielectric_factor               4
bump_filter                     yes
bump_overlap                    0.75
receptor_file                   receptor.mol2
box_file                        showbox_output
vdw_definition_file             DOCK_ROOT/parameter/vdw.defn
score_grid_prefix               jerry
flexible_ligand                 no
orient_ligand                   yes
score_ligand                    SCORE_LIGAND
minimize_ligand                 no
multiple_ligands                yes
intermolecular_score            yes
gridded_score                   yes
grid_version                    4
vdw_scale                       1
electrostatic_scale             1
random_seed                     0
match_receptor_sites            yes
random_search                   no
ligand_centers                  no
automated_matching              yes
maximum_orientations            MAX_ORIENTATIONZ
write_orientations              yes
ligand_atom_file                LIGAND_ATOM_FILE
receptor_site_file              good_spheres.pdb
ligand_out_file                 SERVER_OR_CLIENT_out.ptr
receptor_out_file               receptor_out.mol2
bump maximum                    1
bump_maximum                    1
energy_minimize                 yes
initial_translation             1
initial_rotation                0.1
maximum_iterations              100
energy_convergence              0.1
```

```
maximum_cycles                  5
cycle_convergence               1
energy_termination              1
ligand_energy_file              SERVER_OR_CLIENT_nrg.ptr
parallel_jobs                   yes
ligands_maximum                 1000
initial_skip                    0
interval_skip                   0
heavy_atoms_minimum             0
heavy_atoms_maximum             100
rank_ligands                    yes
rank_ligand_total               100
restart_interval                100
quit_file                       SERVER_OR_CLIENT.quit
dump_file                       SERVER_OR_CLIENT.dump
info_file                       SERVER_OR_CLIENT.info
restart_file                    SERVER_OR_CLIENT.rst
rank_orientations               yes
rank_orientation_total          100
chemical_definition_file        DOCK_ROOT/parameter/chem.defn
chemical_score_file             DOCK_ROOT/parameter/chem.defn
ligand_chemical_file            SERVER_OR_CLIENT_chm.mol2
contact_cutoff_distance         4.5
contact_clash_penalty           50
ligand_contact_file             SERVER_OR_CLIENT_cnt.mol2
parallel_server                 SERVER_YES_NO
server_name                     SERVER_NAME
client_name                     SERVER_OR_CLIENT
client_total                    NUM_CLIENTS
```

**Figure A.8 Master input deck with Script references**

```
/rx/xena/home/dock/4.0.1
master.inp
Camb
8
/rx/xena/home/dock/4.0.1./database/Cambridge_database.mol2
500
```

**Figure A.9 small input deck required for '.sh' file – Camb.inp**

- Line 1 is the DOCK_ROOT.
- Line 2 is the name of the input deck this information will be sent.
- Line 3 is the prefix name all subsequent files will retain.
- Line 4 represents the total number of clients. Each client receives molecules from the database under the direction of the server. Client #2 processes a molecule only when client#1 is busy. Client #3 receives a molecule to process only when clients 1&2 are buys. For client #8 to process a molecule from the server, clients 1 through 7 must have all been busy at the exact same moment.

The parallel, server-client approach dramatically decreased overall CPU time as each SGI mainframe contained two processors.

- Line 5 is the location of the database.
- Line 6 is the number of orientations examined for each ligand.

```
#!/usr/bin/ksh
#
# this file requires that-
#    the text SERVER_NAME be placed in the position of the server's name
#    the text CLIENT_NAME be placed in the position of the client's name
#
#  ask for the DOCK_ROOT environment just in case
read DOCK_ROOT?"DOCK_ROOT?"
#  ask for the original server input file
read server_basis?"server input file for basis?"
#  ask for a name designation that will be applied to all file names with
server or client added
read basis_name?" name designation?"
# ask for how many clients to generate
read tot?" number of clients to generate?"
# ask for the ligand database input file
read ligand_atom_file?" ligand database, full directory and name?"
# ask for the maximum orientations
read max_orientations?" maximum orientations?"
#
# generate server first
echo $basis_name $server_basis $tot
server_name=$basis_name"_server"
echo generating $server_name
sed s/SERVER_NAME/$server_name/ $server_basis |  sed s/NUM_CLIENTS/$tot/ | \
    sed s#LIGAND_ATOM_FILE#$ligand_atom_file# | sed s#DOCK_ROOT#$DOCK_ROOT# >
$server_name.inp_tmp
cp $server_name.inp_tmp $server_name.inp_tmp_clients

sed s/SERVER_OR_CLIENT/$server_name/ $server_name.inp_tmp_clients | sed
s/SCORE_LIGAND/no/  | \
    sed s#MAX_ORIENTATIONZ#1#  > $server_name.inp_tmp
#
# make job submission script
echo "#!/bin/csh -f" > $basis_name"_job.sh"
echo $DOCK_ROOT"/bin/dock -i " $server_name.inp " -o " $server_name".output
&" >> $basis_name"_job.sh"

# make quit file
echo "#!/bin/csh -f" > $basis_name"_quits.sh"
echo "# To stop all jobs, uncomment out the server" >> $basis_name"_quits.sh"
echo "#touch "$server_name.quit >> $basis_name"_quits.sh"
# make restart file
echo "#!/bin/csh -f" > $basis_name"_restarts.sh"
echo cp $server_name".output" $server_name".output_before_restart" >>
$basis_name"_restarts.sh"
echo $DOCK_ROOT"/bin/dock -r -i " $server_name.inp " -o "
$server_name".output &" >> $basis_name"_restarts.sh"
```

```
#
#
echo generate clients now
#
icnt=1
while [ $icnt -le $tot ]
do
   echo $icnt
   client_name=$basis_name"_client_"$icnt
   echo generating $client_name
   sed s/SERVER_OR_CLIENT/$client_name/ $server_name.inp_tmp_clients | sed
s/SERVER_YES_NO/no/ | \
          sed s#MAX_ORIENTATIONZ#$max_orientations# | \
          sed s/SCORE_LIGAND/yes/ > $client_name.inp

   echo $DOCK_ROOT"/bin/dock -i " $client_name.inp " -o "
$client_name".output &"  >> $basis_name"_job.sh"

   echo "client_name_"$icnt"      " $client_name >> $server_name.inp_tmp
# add to quit file
   echo "touch client_name_"$icnt.quit >> $basis_name"_quits.sh"
# add to restart file
echo cp $client_name.output $client_name.output_before_restart  >>
$basis_name"_restarts.sh"
echo "dock -r -i " $client_name.inp " -o " $client_name".output &" >>
$basis_name"_restarts.sh"

   icnt=`expr $icnt + 1 `
done

sed s/SERVER_YES_NO/yes/ $server_name.inp_tmp > $server_name.inp
rm  $server_name.inp_tmp $server_name.inp_tmp_clients

chmod +x $basis_name"_quits.sh"
chmod +x $basis_name"_restarts.sh"
chmod +x $basis_name"_job.sh"
```

**Figure A.10 Script file Camb_dock.sh**

Once the multi-ligand, multi-orientation docking was completed, the essential task of re-examining each of the top hits became the next hurdle. The script shown in figure A.11 has the function of extracting one single user-selected compound from a database. This particular script read from the Cambridge University database however adjustments are readily completed. It reads the initial line of the database and each subsequent molecule, assigning the value of 1 to the counter FLAG. As long as the script does not locate the compound of choice, FLAG is re-

assigned the values 0 and 1 within the loop. The searching continues until the string “ANDREO” is found. At that point, when the string $1 is actually equal to the search phrase, FLAG become equal to 2 and the molecular information of that particular compound is printed, to become a usable file for docking either as a follow-up (the parameters for spheres or grid may have been altered) or as a test of an dock input deck.

```
# awk script to
#
#
{
  if ($1 == "@<TRIPOS>MOLECULE")
     {
     FLAG = 1
     HEADER = $0
#     printf "test\n"
      }
  else if (FLAG == 1)

          {
           if ($1 == "ANDREO")
              {
     printf "@<TRIPOS>MOLECULE\n"
              FLAG = 2
#              print $HEADER
              print
              }
           else
              {
              FLAG = 0
              }
          }
   else if (FLAG == 2)

      {
       if ($1 != "@<TRIPOS>MOLECULE")
           {
           print
           }
       else
           {
           FLAG=0
           }
        }
    else FLAG = 0
 }
```

**Figure A.11 – awk script to extract molecule of choice from database**