Using Supercomputing to Conduct Virtual Screen as Part of the Drug Discovery Process in a Medicinal Chemistry Course

David Toth
University of Mary Washington
1301 College Avenue
Fredericksburg, VA 22401

dtoth@umw.edu

Jimmy Franco
Merrimack College
315 Turnpike Street
North Andover, MA 01845

jimmy.franco@merrimack.edu

ABSTRACT
The ever-increasing amount of computational power available has made it possible to use docking programs to screen large numbers of compounds to search for molecules that inhibit proteins. This technique can be used not only by pharmaceutical companies with large research and development budgets and large research universities, but also at small liberal arts colleges with no special computing equipment beyond the desktop PCs in any campus' computer laboratory. However, despite the availability of significant quantities of compute time available to small colleges to conduct these virtual screens, such as supercomputing time available through grants, we are unaware of any small colleges that do this. We describe the experiences of an interdisciplinary research collaboration between faculty in the Chemistry and Computer Science Departments in a chemistry course where chemistry and biology students were shown how to conduct virtual screens. This project began when the authors, who had been collaborating on drug discovery research using virtual screening, decided that the virtual screening process they were using in their research could be adapted to fit in a couple of lab periods and would complement one of the instructors' courses on medicinal chemistry. The resulting labs would introduce students to the virtual screening portion of the drug discovery process.

General Terms
Supercomputing, Computational Chemistry Education, Drug Discovery, Medicinal Chemistry

Keywords
Docking, Virtual Screening, AutoDock Vina, PyMOL

1. INTRODUCTION
Identifying novel chemotherapeutics has become increasingly challenging and expensive. For every 10,000 compounds evaluated in animal trials, only 10 will make it to clinical trials. The average cost to bring a drug to market is estimated to be about 800 million dollars [1]. Thus the need for more efficient methods of identifying compounds has become increasingly important. One of these methods is virtual screening. The increasing amount of available computing power and the number of protein structures that have been solved have made this an increasingly attractive approach. As of April 10, 2012, there were 80,710 structures in the Protein Data Bank (PDB), which offer a plethora of possibilities for conducting virtual screens [2]. The number of solved structures will only increase as thousands of structures are deposited annually in the PDB.

Many of the chemistry and biology curriculums lack sufficient computational instruction to prepare the next generation of scientists. Proficiency in computational science has become increasingly important. Many industrial companies including big pharmaceutical companies such as Pfizer, Genentech, Eli Lilly & Co and Johnson & Johnson have begun using methods like virtual screening to improve their efficiency in the drug discovery process. Thus student graduating with experience using computational tools and methods will be much more employable.

In this paper we describe our experiences with students using a supercomputer to conduct a virtual screen using AutoDock Vina to identify inhibitors for a number of diseases [3]. The docking program calculates the binding affinity of each of the compounds in a library of compounds specified by the user. The compounds are sorted by binding affinity using Microsoft Excel and subsequently the top hits can be visualized in PyMOL [4]. Visualizing the compounds in PyMOL allows the student to confirm that the predicted binding conformation would induce the required inhibitory affect. The project described here can be incorporated to into a large drug discovery project. The compounds identified as hits from the docking

program could subsequently be screened in a wet laboratory.

Having undergraduate students work on drug discovery in an academic environment is now feasible with the minimal computational power available on any campus and the supercomputer time one can obtain with grants. This type of applied project stimulated interest amongst our students, as they were able to envision what the impact of the project would be if they found a good potential inhibitor. This project also allowed us to highlight the interdisciplinary nature of the modern drug discovery process, which relies on computer science, chemistry, and biology. The project outlined here creates a platform for a drug discovery research project.

2. RELATED WORK

While there have been several articles published about using virtual screens in a curriculum, none to our knowledge have used supercomputing [5]. An advantageous attribute of this project is the ability to adapt this project to any number of diseases or disorders. Along with a large variety of targets the concept of using super computing power can be adopted to a wide variety of simulations and modeling programs [6, 7].

A recent article by Sutch et al. described an activity focused on a structure based drug design [8]. One of the programs used in that activity to conduct the virtual screen was MEDock, which is a simple docking program. Unfortunately, this simplicity also imposes several imitations on MEDock's versatility. It only allows for areas of 300 atoms to be evaluated in a virtual screen. It also limits the number jobs that can be submitted. The largest problem with this program is that it perpetuates the black box thinking of virtual screening. Students need much less insight into the program to be able to successfully screen compounds, thus requiring less understanding of the science behind the project. Also in the project described herein students use PyMOL, which is a commonly used program for visualizing macromolecules in both academia and industry.

Other articles describing small molecule interactions with drug targets have focused on the specifics of the compounds' conformations and chemical properties in relation to protein, but do not address the greater issue of the drug discovery process or virtual screening [9, 10]. None of the previous lab activities we have found in the literature required the student to engage in the computer science aspect as much as this activity does. Most of the activities used programs that are less versatile but have a graphical user interface. While this can be very advantageous for large classes, it does allow the students to conduct the exercise with out much understanding of the docking program.

3. THE GOALS AND ACTIVITIES

There were a number of goals for the labs, including getting students to learn the important role that computers can play in the drug discovery process. Students were also supposed to learn how to use a docking program, gain experience using software other than the commercial off-the-shelf software they use on a daily basis, and get exposure to the Linux operating system and a command line interface. Other goals were to gain an appreciation for how much supercomputers can speed up the virtual screening process and understand that supercomputing time can be obtained at no cost even by small academic institutions that do not have the financial resources to buy a supercomputer or time on a supercomputer. Students also were shown how to use PyMOL, a protein visualization program. Finally, students also were shown some new data analysis skills with Excel.

Students learned a little about supercomputers as part of the lab. The most striking thing that students learned is that the virtual screening process is significantly faster using a supercomputer, because they can screen many molecules at once, rather than only a few at a time, when using a single CPU core per molecule. They also learned that as opposed to desktop computer or a server where you can just start tasks whenever you want, on a supercomputer, you must submit your task to the queuing system and the queuing system controls when your task is run. The students learned that the queuing system using a number of factors to determine when a task should be run, including the number of CPU cores needed and the amount of time requested for the task. Therefore, they understood that while using more CPU cores might finish the virtual screen faster once the task was started, requesting many more resources might delay when the task was started and could ultimately result in the virtual screen being completed later than if they requested fewer CPU cores. Students also learned how to check the queue on the supercomputer to see whether their task was waiting or being run. Figure 1 shows a screen capture from the lab manual where the students would check the queue.

The laboratory was conducted in three phases over the course of two days due to the laboratory time being 1 hour and 15 minutes. However, we note that the activity would fit very well in a traditional 3 hour laboratory. Prior to the laboratory students had to choose a protein that was known to be a good drug target from the PDB. To identify a protein student conducted a literature search using SciFinder, PubMed or Google Scholar. Students were instructed to identify a protein that had been previously shown to be a good drug target, either through chemical inhibition, knockout study, or methods that demonstrated the proteins potential as a drug target. Secondly, students had to very verify that 3D structure had been solved. This was easily done by searching the PDB site for the structure. With only these two constraints, students have a large number of targets to choice from. The variability of the drug target selection was purposely done was to allow the students to take ownership over the project as well to force the student to think critically about the target. The

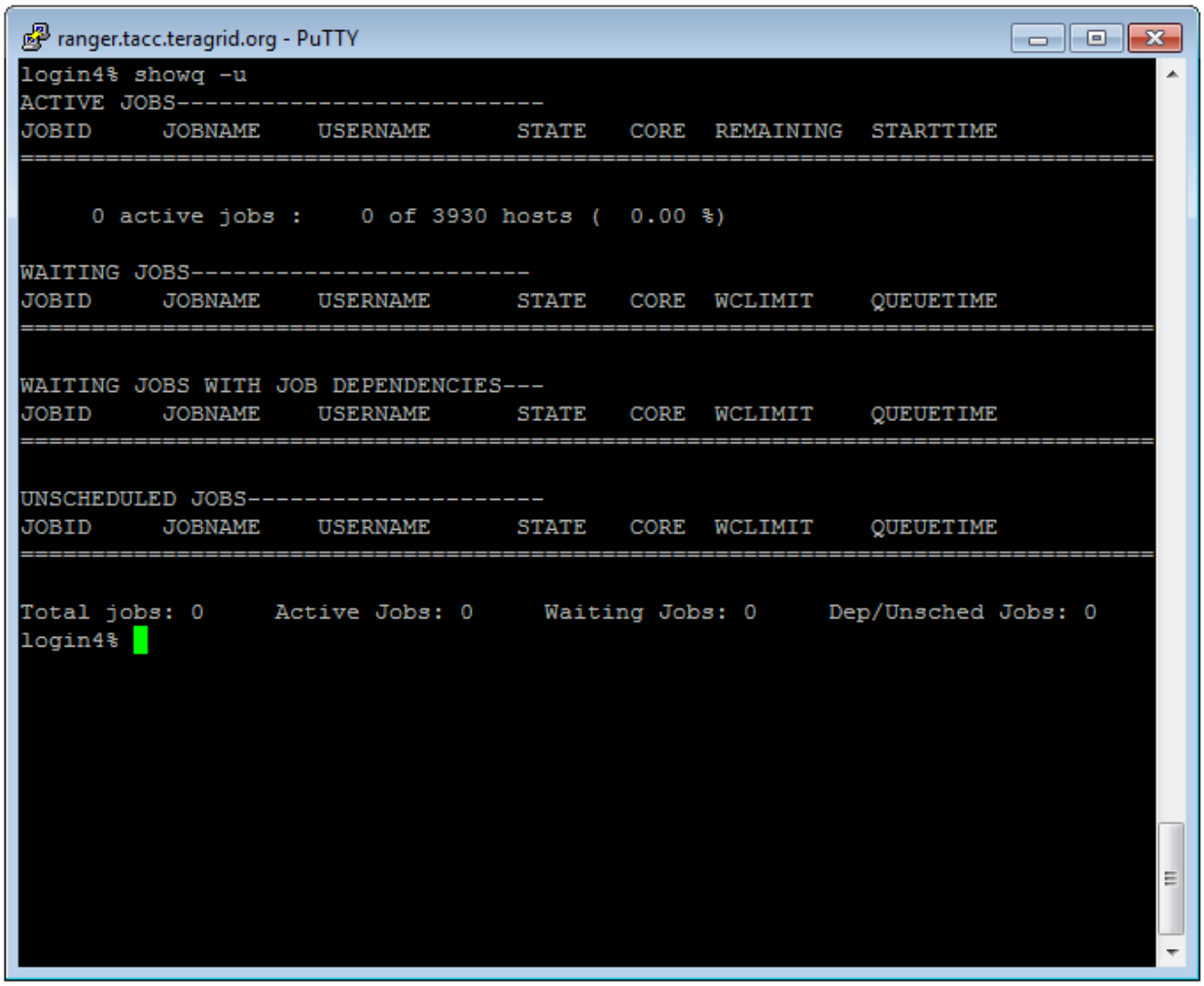

Figure 1: A screen capture from the laboratory manual showing the output from querying the supercomputer's queue.

structures were then converted to pdbqt files (the file format that AutoDock Vina uses) and the search grid was set using AutoDock Tools [11]. Although students were required to choose a protein, convert it, and find the search grid, three pre-converted proteins structures and search grids have been included in the supplemental materials with this paper (Sample_Targets.doc) for readers wanting to test the lab without having to first find a protein, convert it to a pdbqt file, and find the search grid. Those three proteins are targets for Alzheimer's disease, Cancer, and HIV.

On the first day, students carried out Phase one of the laboratory. Before beginning the laboratory, students were given a 26-page full-color laboratory manual that included numerous screen captures to help them with the laboratory. Figures 2 and 3 show portions of the lab manual showing the students how to lo on to a remote computer with ssh and scp. The lab manual has been included in the supplemental materials with this paper (Laboratory_Manual.docx). Students began the laboratory by downloading a protein pdbqt file from the course Blackboard site and then a secure shell (ssh) program and a secure copy (scp) program from the Internet. Next, each student was given a distinct username and password to log onto a server on campus. Students logged on to the server, transferred the protein file to the server, and using the docking program, tested how well the protein bound to a potential drug molecule. We showed the students how to use AutoDock Vina, an open source software package from the Scripps Institute [3].

Students continued working on the laboratory on the second day by starting with Phase two. In Phase two, students created a shell script to automate the screening of multiple compounds. While Phase two of the laboratory manual included instructions to perform the analysis of the data from the virtual screening, to save time, students were assigned to do that portion of the laboratory at home. Students then did Phase three of the laboratory. In Phase three, students logged on to a supercomputer located across

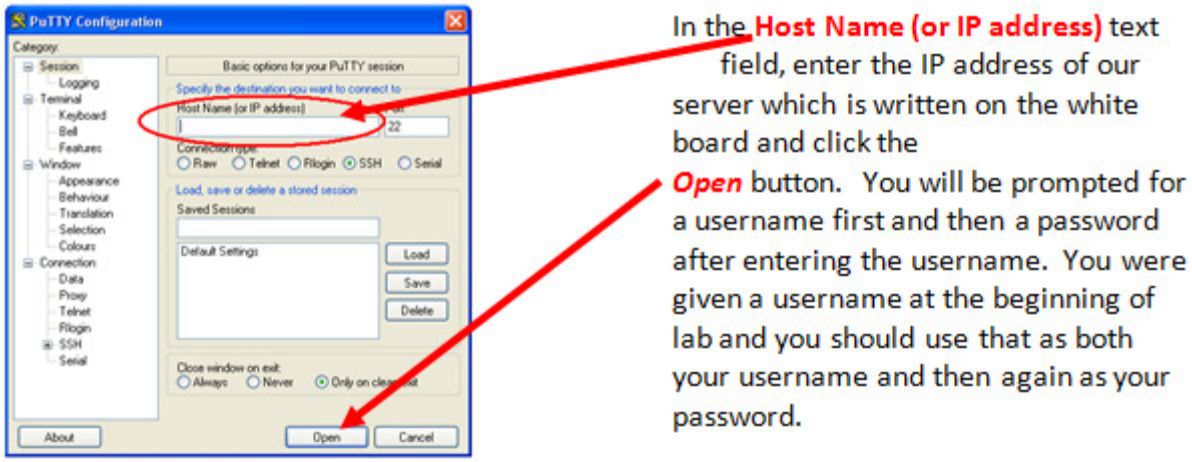

Figure 2: Portion of the laboratory manual showing how to log on to a supercomputer or a remote server using ssh.

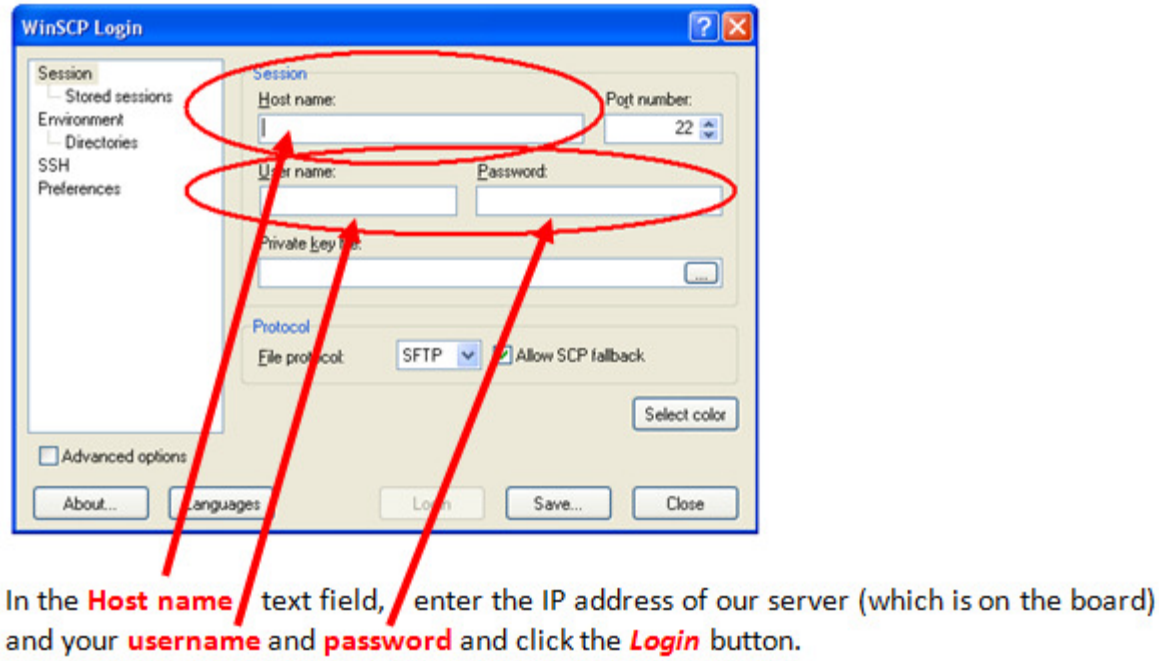

Figure 3: Portion of the laboratory manual showing how to log on to a supercomputer or a remote server using scp.

the country and uploaded a protein file. Finally, they edited a shell script that they could use to automate the virtual screening, and submitted the job to the batch scheduling system on the supercomputer.

Finally, the students were asked to complete their projects in their assigned groups. The output files from the virtual screens were posted on blackboard for the convenience of the students. The results of each group's screen were posted as text files. Each group converted their text file to an Excel file so they could quickly identify the top binding affinity compounds; those compounds are termed hits.

AutoDock Vina identifies several binding conformations for each of the compounds screened and outputs those values, as shown in Figure 4. The compounds the students screened came from the ZINC database (http://zinc.docking.org/pdbqt/) [12]. However, the conformation with the best binding affinity is the one most likely to occur, so the students were supposed to remove the data for the other conformations of the same compound from the data. Using Excel, the students were able to remove the extra conformations for each compound and sort the remaining data to quickly identify the compounds with the best binding affinity. The hits are subsequently visualized in PyMOL. An example is shown in Figure 5.

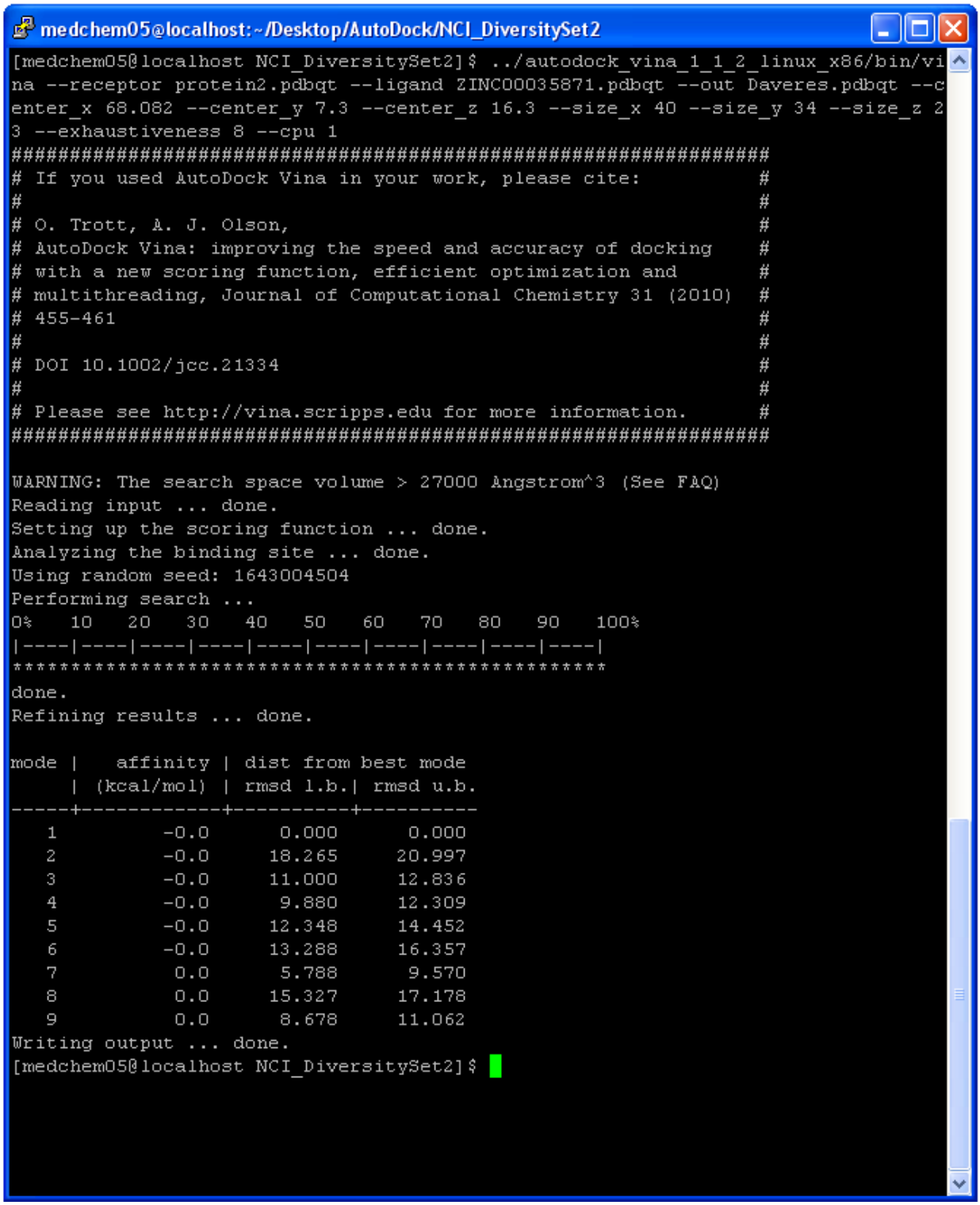

Figure 4: Portion of the laboratory manual showing the output of AutoDock Vina after screening a molecule.

A handout with commands for PyMOL is included in the supplementary materials with this paper (PyMOL_Commands.doc). Visualization of the binding conformations with the protein is important since a compound may have a high binding affinity for a protein, but may not inhibit its activity. When visually inspecting the hits with PyMOL, students were instructed to verify that the compounds bind to the active site or to a known

allosteric site of the protein. When ranking the hits and trying to identify a few lead compounds to pursue, if specific details about the protein are known, such as, if a residue is essential to the protein's function, then any compounds interacting with these residues should be given extra consideration.

During the evaluation of the hits in PyMOL, students had to critically evaluate how the potential inhibitor was predicted to bind to the targeted protein. Two main aspects were focused on during the evaluations of binding: orientation and the interactions between the compounds and the protein. First, did it bind in a manner that would inhibit enzymatic activity? Typically this could be determined by it binding to either the active site or a known allosteric site. Also compounds predicted to interact with residues previously shown to be important were especially noteworthy. Secondly, students examined what interactions the compounds have with the targeted protein, such as hydrogen bonding, ionic interactions, and hydrophobic interactions, as these interactions determine its binding affinity. The students learned the relationship between how the thermodynamics of the calculated values relate to how the compounds interact with the protein. By being able to visualize the predicted binding conformations in PyMOL, students were able to see the interactions that lead to the predicted binding affinity. Compounds displaying a large number of favorable interactions displayed the greatest binding affinity. Lastly, many students fail to see the importance of understanding thermodynamics and this project allows the students to see a real world application of thermodynamics. During the lab section, students were given a brief explanation on how the docking program measures the energy between the compounds and the protein.

For this course, Medicinal Chemistry, thirteen students consisting of biology, chemistry, and biochemistry majors participated in the activity. Each group of students had to create a written report of their finding where they had to give some background about the disease, explain the function and importance of the selected target, and demonstrate that they had identified a potential inhibitor for the targeted protein. The groups were also required to present their work to the rest of the class with a PowerPoint presentation.

4. STUDENT REACTIONS

The students in the course were given an anonymous survey at the end of the semester. The survey is included in the supplementary materials (survey.pdf). The survey responses indicated a significantly increased awareness of the availability of supercomputing resources. The surveys also showed that the students learned how to use the software for the project, including AutoDock Vina and PyMOL, and that the students learned new techniques in Excel. The surveys demonstrated that the students became more comfortable using the command prompt and they also learned some simple UNIX commands. Because there is a significant amount of scientific software that must be run from the command prompt, increasing the students' comfort with the command prompt is very important as we try to prepare them for their future careers. The students

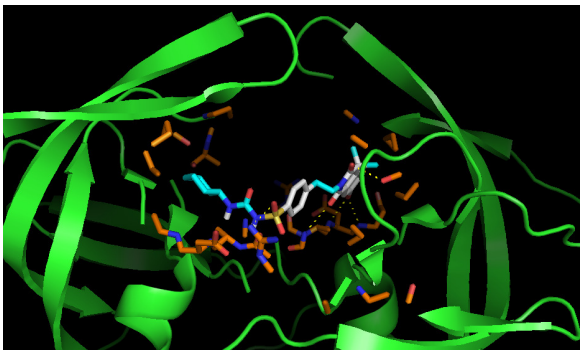

Figure 5: An image of one of the top binding inhibitors as calculated by AutoDock Vina. This is an example of an image the students will generate during the project using PyMOL.

also learned the importance of computation in science, as an alternative method of solving problems, so they understand that science can be done outside of a wet laboratory. They also understood that supercomputing could be applied to problems in other domains and would recommend its use for other projects. In addition to learning computational science techniques, students also demonstrated an increased understanding of fundamental chemistry concepts.

The students in the class reacted very favorably to the laboratory. All of the students felt that the laboratory manual was easy to follow. At the end of the course, over one third of the students expressed a desire to continue working on the projects and in particular, work more on the computational aspect of the project and conduct virtual screens of more compounds to try to find more potent inhibitors. These students, who were completing their Junior year, will be working on directed research projects related to the course projects in the upcoming year. A large percentage of the graduating Seniors also expressed that they would have continued working on the projects if they were not graduating, and several said they might be willing to come back over the summer to continue until they had found jobs. One student commented that she thought the computing aspect of the project was "extremely interesting and educational." Because this student was not very comfortable with using computers for science to begin with, we found that this was very encouraging.

5. INSTRUCTOR REACTIONS AND LESSONS LEARNED

It took 2.5 hours spread over two days for students to complete the laboratory activities other than the data analysis portion. The students were able to complete the

Phase one activities in 1 hour, with each student working on their own computer. During the first day, although the students were given the laboratory manuals, the instructor showed the students how to do the tasks on a computer where the screen was projected at the front of the classroom. This was done to try to get the students more comfortable with some of the tasks that they might be less familiar with. On the second day, students worked in their project groups, with one group per computer and the students were told to follow the laboratory manual's instructions but to feel free to ask questions whenever they had any trouble. Because of the detail of the laboratory manual, which included numerous screen captures, this worked well. The instructors felt that the second day went smoother than the first day and believed that forcing the students to follow the laboratory manual rather than having one instructor demonstrating the tasks at the front of the room worked very well. However, it may have been important to help the students get comfortable with the tasks during the first day of the laboratory by having them watch the instructor rather than having them simply follow the instructions on the laboratory manual.

During the process of the students downloading the ssh client and the scp client, we learned that although the laboratory manual was very detailed, the students tended to have difficulty entering URLs correctly. As we progressed through the laboratory, we discovered that the same idea held for places where the students needed to type commands. Instructors using this lab should be aware of the difficulties that the students had with entering URLs and commands so they are prepared for the inevitable questions about why something does not work.

One issue that we did not anticipate was the amount of time that it took the server the students used in Phase one and Phase two of the laboratory to run the virtual screens. The server used was a dual-core desktop computer that was 4-5 years old and while it did the processing quickly enough during our testing of the laboratory materials before giving the laboratory to the students, multiplying the tasks the computer needed to do by 13 proved to be too much for the computer to handle gracefully. While it completed all the tasks, it was slow enough in Phase one of the laboratory that we put the students into their project groups for Phase two and Phase three. We recommend that others use a computer better equipped to handle the computational demands of the number of students.

There are a few suggestions that we have for other instructors who will use this module when teaching their courses. The number of students in each group in our class ranged from 2 to 5 students. The groups were allowed to divide the work up as they wanted. In the smaller groups, it appeared that all the student were extremely active. In the larger groups, the amount of work varied vastly between students. Thus, one suggestion we have is limiting the size of the student groups to 2-3 students. The instructor may also want to load the protein files that students will use and the shell script onto the server before the laboratory, if they want to shift the focus more towards the chemistry aspect and minimize the computer science portion of the project.

The instructor may want to mention before the laboratory that not every command that the students enter will result in significant visual output in the command window, as this confused several of our students.

6. CONCLUSIONS

We were able to develop a hands-on laboratory project that allows students to gain valuable experience in computational science with real-world applications. We have been able to present the material in a manner that engaged the students and stimulated interest in computational science research. Because all of the software is open source and all the required resources beyond what exist in any college computer lab are freely available through grants, this project can be done at schools of any size at no cost. The project can be run as a pre-packaged standalone laboratory assignment just to introduce students to virtual screening and computational chemistry or as a large semester-long project. If the project is run as a full course project, it could be used to prepare students for directed research projects in drug discovery and senior thesis work. Since computational science has become a more substantial part of a number of scientific disciplines, this project could be used as a model to develop other computational science lab projects.

7. FUTURE WORK

In continuing the development of this project, we will extend it to a full semester project. In the full semester project, students would be required to use the information gathered from evaluating the hits in virtual screen to generate a second generation of inhibitors. Thus, when students are evaluating the hits in PyMOL they will have to identify any additional interaction that can be utilized by an inhibitor. This could be done by adding and additional hydrogen bonding acceptor or donor, creating a hydrophobic group to utilize a hydrophobic pocket, or removing a group that is creating an unfavorable steric effect. This deals with the properites a compounds should posses to be a more likely drug candidate. Once the students have designed a set of compounds they will virtually construct them using Jmol (http://jmol.sourceforge.net/) and Open Babel (http://openbabel.sourceforge.net/), which are both open source software. The compounds will then be re-screened to indentify which chemical modications had the largest effect on the binding affinity. Lastly students would be asked to propose a synthesis for their top three hits.

In future work, we will create an electronic lab kit, containing a step-by-step laboratory manual and either all the files or links to all the files, depending on licensing restrictions, that instructors would need to recreate the laboratory at their own institutions.

8. ACKNOWLEDGEMENTS

This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575.

We wish to thank the Extreme Science and Engineering Discovery Environment (XSEDE) program, which supported this work by providing the supercomputer time through grant TG-MCB120071. We also wish to thank the Texas Advanced Computing Center (TACC), which provided the supercomputer we used for this work. We would also like to acknowledge the students in the course for their participation in this project.

9. REFERENCES

[1] Silverman, Richard B. The Organic Chemistry of Drug Design and Drug Action. 2nd ed. Amsterdam ; Boston: Elsevier Academic Press, 2004.

[2] RCSB Protein Data Bank http://www.rcsb.org/pdb/home/home.do.

[3] Trott, O.; Olson, A. J. *Journal of Computational Chemistry* **2010,** *31*, 455-461.

[4] The PyMOL Molecular Graphics System, Version 1.2r3pre, Schrödinger, LLC. http://www.pymol.org/

[5] Baudry, J.; Hergenrother, P. J. *Journal of Chemical Education* **2005,** *82*, 890.

[6] Artavanis-Tsakonas, K.; Weihofen, W. A.; Antos, J. M.; Coleman, B. I.; Comeaux, C. A.; Duraisingh, M. T.; Gaudet, R.; Ploegh, H. L. *The Journal of biological chemistry* **2010,** *285*, 6857-66.

[7] Sotomayor, M.; Weihofen, W. A.; Gaudet, R.; Corey, D. P. *Neuron* **2010,** *66*, 85-100.

[8] Sutch, B. T.; Romero, R. M.; Neamati, N.; Haworth, I. S. *Journal of Chemical Education* **2012**, 89 (1), pp 45–51.

[9] Yuriev, E.; Chalmers, D.; Capuano, B. *Journal of Chemical Education* **2009**, 86 (4), p 477.

[10] Manallack, D. T.; Chalmers, D. K.; Yuriev, E. *Journal of Chemical Education* **2010**, 87 (6), pp 625–627.

[11] AutoDock Vina – molecular docking and virtual screening program http://vina.scripps.edu/tutorial.html.

[12] Irwin, J. J.; Shoichet, B. K. *Journal of Chemical Information and Modeling* **2004,** *45*, 177-182.