JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

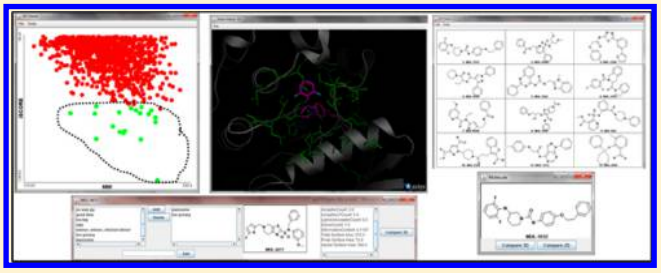
pubs.acs.org/jcim

VSViewer3D: A Tool for Interactive Data Mining of Three-Dimensional Virtual Screening Data

Kyle I. Diller[†] and David J. Diller*[,†,‡]

[†]Data2Discovery Consulting, East Windsor, New Jersey 08520, United States

ABSTRACT: The VSviewer3D is a simple Java tool for visual exploration of three-dimensional (3D) virtual screening data. The VSviewer3D brings together the ability to explore numerical data, such as calculated properties and virtual screening scores, structure depiction, interactive topological and 3D similarity searching, and 3D visualization. By doing so the user is better able to quickly identify outliers, assess tractability of large numbers of compounds, visualize hits of interest, annotate hits, and mix and match interesting scaffolds. We demonstrate the utility of the VSviewer3D by describing a use case in a docking based virtual screen.



INTRODUCTION

Virtual screening is an increasingly important tool for lead discovery in many drug discovery programs. The most straightforward application of virtual screening is searching through an in house collection of molecules. In this case, it suffices for the computational expert to run the virtual screen either on a structure-based or ligand-based 3D query, select the top hits, and pass the list to the biology group to test. Only after some of the virtual hits are found to be truly biologically active does the full drug discovery team get involved.

The decision making during a virtual screening workflow can, however, be much more complex. For example, if the molecules being screened are from a commercially available database a number of factors beyond a match to the 3D query are usually discussed prior to ordering the virtual hits. These almost always include issues that are fundamentally chemical in nature such as physical−chemical desirability, novelty, and chemical tractability. Many of these issues are difficult to quantify computationally and indeed may be dependent on the experience of the drug discovery team members.[1,2] To address these issues, the optimal approach is to include members of the drug discovery team beyond the computational experts.

In both the above-mentioned applications, virtual screening is used primarily in the early stages of drug discovery as a hit finding tool. At the other extreme, virtual screening could be used as an integral tool for lead optimization. In this application, virtual screening is used to continually scan large areas of chemical space much of which represents molecules that have never been synthesized internally or commercially. The virtual hits could be presented to the team as candidates for synthesis or more likely as a way to generate ideas and ultimately stimulate the creativity of the team. Here, many of the objectives of the search are implicit and perhaps not even known to the individual team members until they see a hit and its fit in a binding site or its alignment to a query molecule. Experimental scientists might not have the expertise to fully understand virtual screening data but often have a wealth of implicit information regarding chemical synthesis, project structure activity relationships or any of the myriad of the other aspects of drug discovery. Without active engagement of the experimentalists this type of information is often lost or applied too late to affect the project.

It is this last example of virtual screening that we address here. In particular, we describe an application, the Virtual Screening Viewer (VSviewer3D), which allows users to interactively search through 3D virtual screening data. The VSviewer3D offers views of the data including a 2D scatter plot, individual 2D molecule structures, a 2D molecular grid view, and 3D visualization. It is intended to encourage the user to explore the data rather than just simply select the molecules with the best score. By exploring the data we feel the user gains a better understanding of how molecules fit to the virtual 3D query and can then better decide if a good score is in fact reasonable. In addition, it has functionality to add and search user comments for each molecule making groups of molecules easier to recall and discuss. Additionally, it has functionality to view individual or entire sets of molecule structures. Finally, it has fast methods for 2D and 3D comparison allowing the user for example to search for molecules that are fit into a binding site or align to a 3D query molecule in a similar manner but have different chemical structures.

IMPLEMENTATION

The VSviewer3D is not a virtual screening engine. Rather, it is intended as a tool for interactive mining of 3D virtual screening data. The virtual screen could take the form of docking into a binding site, alignment to a 3D query, or a pharmacophore search. Throughout this manuscript we will use the generic

Received: September 2, 2014
Published: November 25, 2014

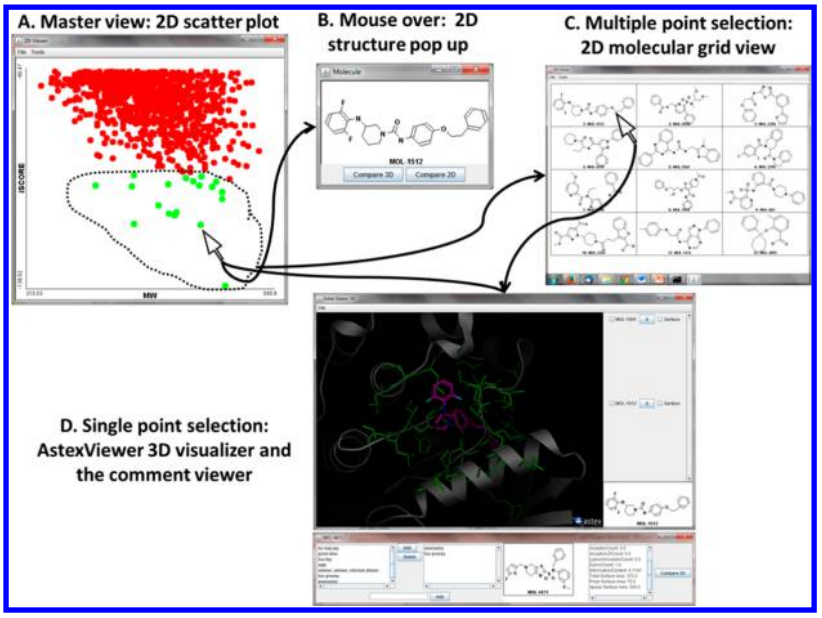


Figure 1. Views available in the Virtual Screening Viewer. The initial view is the 2D scatter plot (A) where each point represents a single molecule. Placing the mouse over a single point will show the 2D structure of the underlying molecule in the molecule pop up viewer (B). Selection of multiple points will bring up the 2D molecular grid view (C) and the data table (not shown). Selection of a single point in the scatter plot or a single structure in the structure grid view will bring up both the Astex Viewer and the comment viewer (D).

term "3D query" to indicate the binding site, the 3D alignment query or the pharmacophore model to which the molecules are being docked, aligned, or fit.

The VSviewer3D consists of a number of views: a 2D scatterplot, a 2D molecular grid view, a data table, a chemistry pop-up window, a comment viewer, and a 3D structure viewer. See Figure 1 for the overall organization of the views. The main entry point of the program is the 2D scatter plot. The main purpose of this view is to create a simple way for users to search through data. A particular strength of a simple scatter plot for examining virtual screening data is that a scatter plot efficiently enables the user to identify outliers. Since truly active molecules are rare, the outliers are typically of greatest interest. The user can interactively explore the data by changing the x and y axes simply by clicking on the axes labels or by selecting a different value from the axis menus. The user can add additional dimensionality to the data by coloring and sizing by the value of any of the descriptors. Hovering over points shows the corresponding 2D molecular structure in the pop-up window allowing the user to see the underlying chemical structure represented by the point. The remainder of the views can be accessed by either selecting a subset of the data or clicking a single point.

By selecting a subset of points, the user is shown two new views: the data table and the 2D molecular grid view. The data table shows all the data values of each molecule including whether the point is selected or not. It displays the average and the standard deviation of the selected and the nonselected data separately. This allows the user to see if the selected points are different from the nonselected with respect to any of the descriptors. The second view that is presented upon multiple selection is the 2D molecular grid view. This window displays the 2D structures of all the molecules selected in a grid style. The structures in the 2D molecular grid view can be sorted by any of their underlying properties, virtual screening scores, etc. This view allows the user to quickly scan the selected hits and focus on the molecules of greatest interest.

By clicking on a data point in the 2D scatter plot or a molecule in the 2D molecular grid view, the user is shown two additional views: the 3D structure viewer and the comment viewer. The comment viewer displays the data associated with the selected molecule, any associated comments, and the 2D structure of the selected molecule. In this view, the user can view the comments for the selected molecule. The user can also add comments from a master list, i.e., the list of all comments from every molecule in the project, or add a new comment. The value of allowing the user to manage comments associated with each molecule is that later the user can search for and select all the molecules with a given comment thus making it easy to recall molecules of interest. Finally, from the comment viewer the user can do similarity searches of all the hits either by 2D chemical similarity or by 3D similarity. These similarities are then available within the scatter plot for comparison and further analysis.

The 3D viewer is the second view shown to the user when clicking on a data point in the 2D scatter plot or a molecule in the 2D molecular grid view. The 3D viewer is an extension of the powerful 3D molecular viewer: the Astex Viewer.[3] The Astex Viewer is fully utilized to display the 3D structure. The view builds on the Astex Viewer by adding a side panel that allows the user to perform standard operations in a single mouse click including hiding/showing a molecule, showing a dot surface, or deleting the molecule from the 3D viewer. The 3D query, whether it is a protein binding site or an aligned set of actives molecules, can be read into the Astex Viewer to allow the user to examine each molecule in the context of the 3D query.

As input, 2D and 3D coordinates are necessary for each molecule. In addition, any number of data fields including calculated descriptors and scores from the virtual screens are needed for input. Finally, user comments must be stored and reread as the data is shared between users. Three avenues to handling the diverse information were considered. The first possibility was to use multiple files, such as one SD file for the

dx.doi.org/10.1021/ci500532j | J. Chem. Inf. Model. 2014, 54, 3446−3452

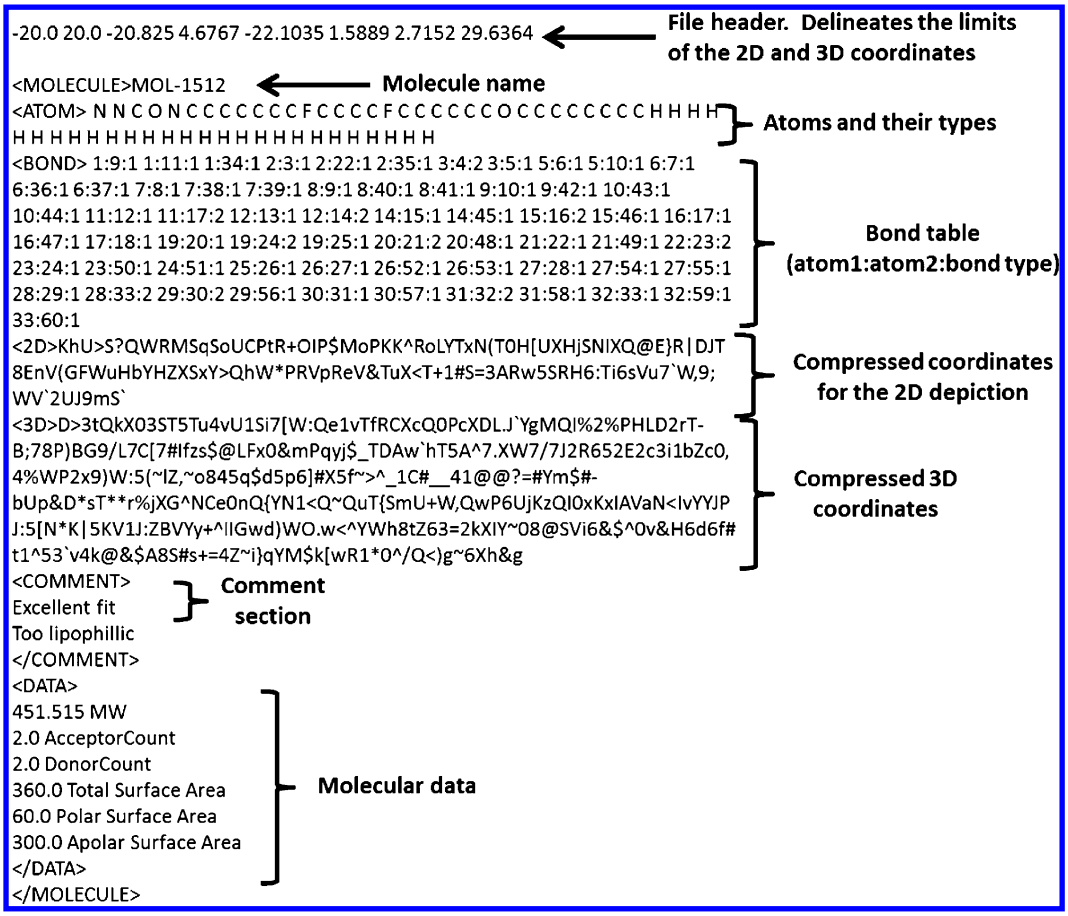


Figure 2. Single molecule record from an example VSV file. The first line is the head information. The first two lines indicate limits of the 2D depictions for all the molecules. The last six numbers indicate the limits of the 3D coordinates for all the molecules. The 3D limits might for example come from the docking box. The remainder shows the format for a single molecule record in the file.

3D coordinates, one SD file for the 2D coordinates, and a CSV file for the data. We felt that this solution was too clumsy and prone to corruption and incongruities between the various files. The second option was to develop a database scheme to handle all the information. We felt that this would be too cumbersome to justify given the ephemeral nature of the data being generated. Rather than handling all this data through multiple files (such as SDF/CSV etc.), we developed a single file format that maintains all this information albeit in a much more compressed format. Figure 2 shows a single molecule record in a VSV file.

The file achieves significant compression of the 2D and 3D coordinates through a simple scheme which allows the file to remain a true text file. As a comparison, a VSV file with 2D and 3D coordinates of approximately 20 000 molecules requires 31 Mb of disk space compared to 115 Mb for the 3D SD file and 68 Mb for the 2D SD file—approximately a 6-fold savings in disk space. The first line in a VSV file is the limits for the 2D and the 3D coordinates. The limits for the 3D coordinates are usually established by the 3D query. These could be a box from a docking calculation or a box around a query molecule.

MOLECULAR COMPARISONS

Upon finding an interesting molecule from a virtual screen, one often asks the question whether there are other molecules that are docked or aligned in a similar fashion. Further, one often wants to know whether there are chemically dissimilar molecules that nevertheless dock or align in a similar fashion. To address these questions we implemented a simple 2D fingerprint which gives a literal assessment as to whether two molecules are chemically similar and a 3D fingerprint which gives an assessment as to how similar two molecules have been fit to the 3D query.

The 3D fingerprint takes advantage of the fact that the 3D coordinates for all the molecules have a relative alignment due to the fact that they are all fit/aligned to the same 3D query. Thus, their coordinates can be directly compared without further alignment. The challenge, however, is that the comparison must be extremely fast, on the order of tens of thousands of comparisons per second, because the user can potentially request many of these comparisons in real time. In order to make the 3D comparisons fast enough the VSviewer3D precomputes a 3D fingerprint in the following manner. The 3D search box in the header information on the VSV file is split into a grid with cell size of 0.5 Å. The 3D fingerprint for a molecule is simply a bit string for which each bit points to a single cell in the grid. A 1 would indicate the molecule fills at least half the given cell and 0 indicates the molecule fills less than half the given cell. While the computation of the fingerprint takes a noticeable amount of time while the file is read—approximately 6 s to load 20 000 molecules. The 3D comparisons are subsequently nearly instantaneous.

dx.doi.org/10.1021/ci500532j | J. Chem. Inf. Model. 2014, 54, 3446−3452

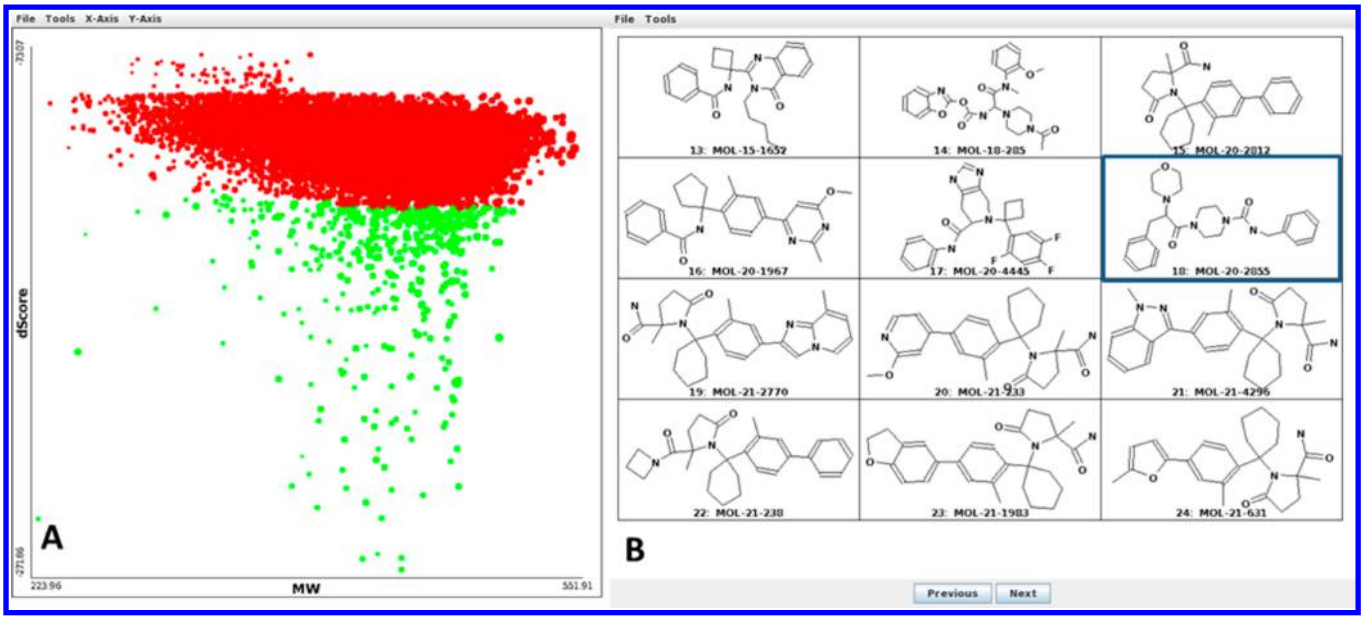


Figure 3. Best docked/scored poses. The scatter plot (A) shows the docking score (*y*-axis) versus the molecular weight (*x*-axis) of all the hits. This depiction clearly shows that a small fraction of the molecules have docking scores that separate them from the vast majority of the virtual screening hits. In this case, these standout molecules have been selected which is indicated by the green coloring. This opens up the 2D molecular grid view (B) which allows the users to easily scan the top hits for interesting scaffolds. In this case MOL-20-2855 is highlighted. Selection of this box would lead to the 3D structure of this molecule being sent to the Astex Viewer (see Figure 4) where it can be examined relative to the 3D query and any other hits that have previously been selected. Any selected structure would also be put in the comment viewer where the user can save any comments, mark the molecule as high interest, and do further analysis such as 2D and 3D comparisons to the rest of the hits.

EXAMPLE APPLICATION

Here we demonstrate the potential of visual data mining of 3D virtual screening data using the functionality of the VSviewer3D through a simple example: small molecule docking of a large number of virtual molecules into the Jak2 pseudokinase domain using the crystal structure 4fvr.[4] Rather than focusing on the ATP binding site, we docked into a deep pocket adjacent to the substrate binding site. To create a demonstration data set, we docked approximately 100 000 virtual molecules and kept the top 20 000, computed several scoring parameters and physical properties such as molecular weight, logP etc. using the open source chemical informatics toolkit RDkit.[5] In this case the molecules were docked using a proprietary docking program. The only aspect of the docking program that is relevant is that ultimately the results, i.e., the docked poses, scores, etc., could be saved as an SD file. A simple utility program, which relies on the CDK toolkit,[6] is then used to convert the 3D SD file into the VSV file needed by the VSviewer3D. As virtually every available docking program and 3D alignment program can output results as an SD file this allows the VSviewer3D to access the results of nearly any 3D virtual screen.

The initial view is the scatter plot. This view allows the user to change the axes through either the drop down menus or by right/left clicking on the axis labels. Each point can be colored or sized by the value of any of the calculated descriptors. As a first step Figure 3a shows a plot of the docking scores (dScore/*Y*-axis) versus molecular weight (MW/*X*-axis) with the points sized by their polar surface area. First this plot allows the user to examine the trade-off between docking score and molecular size. Since docking scores invariably loosely correlate with the size of the molecule this is often a limitation of simply selecting the molecules with the best score. By examining the plot the user can adjust to this effect and not be biased towards large molecules. Second and more importantly, from this plot it is apparent that a relatively small number of molecules have docking scores that stand out from the majority of molecules. The visualization makes it clear that there are molecules that standout without employing a preconceived score cutoff or predetermined number of top hits. In this case, approximately 125 of the molecules stand out from the 20 000. The structures of these molecules can then be easily examined by selecting them in the scatter plot. Selection of these standout molecules leads to the 2D molecular grid view opening with the selected molecules—see Figure 3b. This allows the user to quickly scan the structures and examine in detail those with promising scaffolds.

The molecules with the most promising scaffolds can then be viewed in the context of the 3D query simply by selecting the structure. Upon selection, the Astex Viewer is opened displaying the selected molecule in the context of the 3D query, in this case the Jak2 site. Figure 4 shows molecule MOL-20-2855 in the Jak2 site. This allows the user to quickly assess the strengths and weaknesses of the selected molecule.

As a final example, we demonstrate the utility of structure comparisons to further analyze and exploit 3D virtual screening data. Here we begin with the molecule shown in the comment viewer in Figure 4c, MOL-20-2855. The comment viewer has buttons to perform the 2D and 3D similarity calculations as described above. Once calculated, these similarities are then made available for analysis in the scatter plot. Figure 5a shows the 3D similarity to MOL-20-2855 (*y*-axis) versus the 2D similarity (*x*-axis). Molecules in the lower right are those that are chemically similar but docked to Jak2 in a different way. Molecules in the upper right are those that are chemically similar and similarly docked. Finally, molecules in the upper left are those that are chemically different but docked to Jak2 in a manner similar to the docking mode for MOL-20-2855. These molecules in the upper left are potentially the most interesting.

dx.doi.org/10.1021/ci500532j | J. Chem. Inf. Model. 2014, 54, 3446−3452

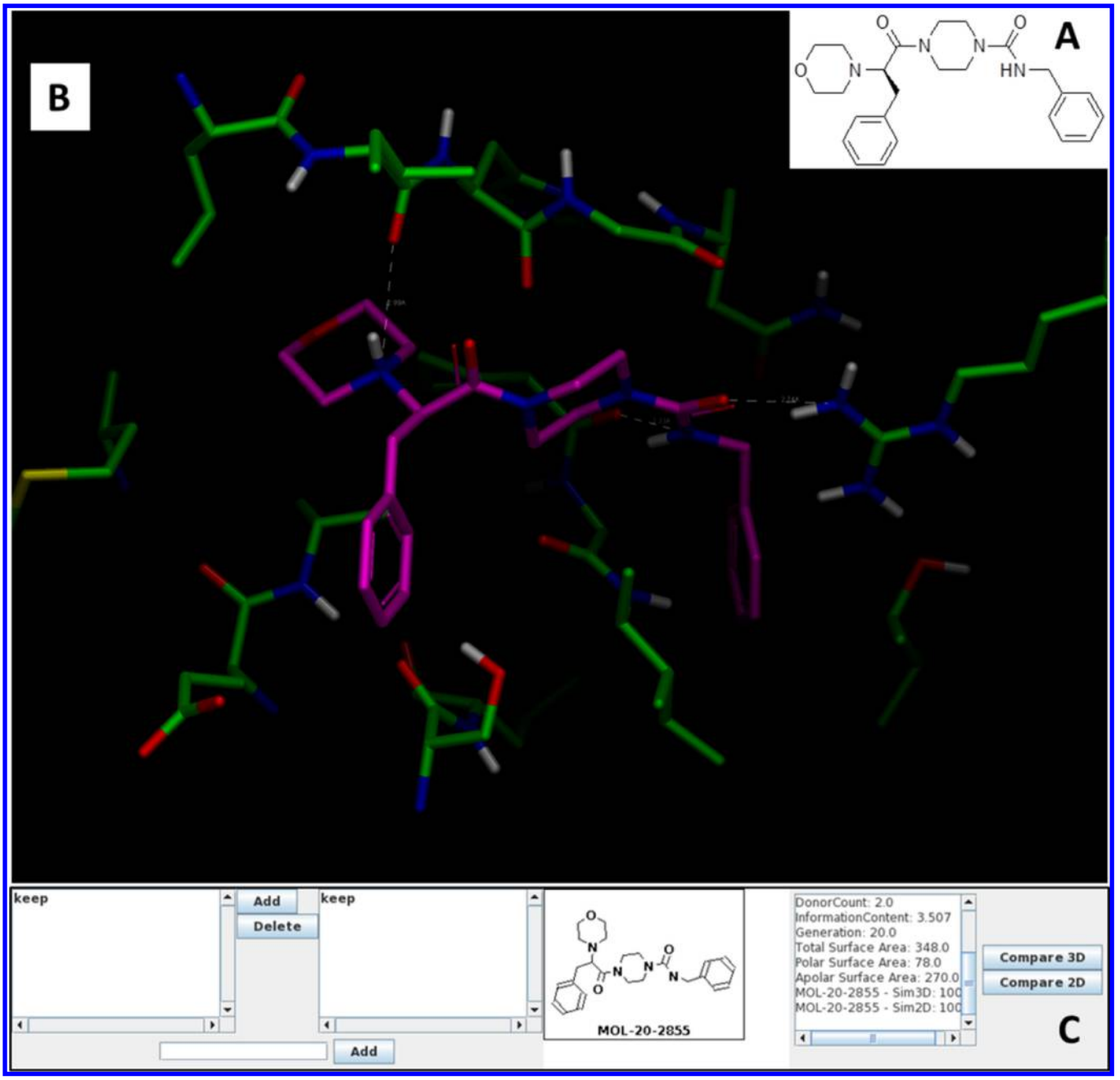


Figure 4. MOL-20-2855 in the Jak2 binding site. Selection of a structure from the 2D molecular grid view leads to the 3D structure appearing in the Astex Viewer. (A) Structure of MOL-20-2855. (B) Image generated directly from the Astex Viewer. (C) Comment viewer.

By selecting these, Figure 5b, the 2D molecular grid view is opened and the user can quickly scan for interesting new scaffolds. Selection of the most interesting of these molecules puts them into the Astex Viewer where they can be compared in 3D to the original hit of interest: MOL-20-2855. Figure 6 shows just the resulting 3D overlay of a handful of chemically distinct molecules that dock to Jak2 similarly to MOL-20-2855. One potential outcome of such an analysis is that a drug discovery team can mix and match various hits to improve the fit to the target protein or the chemical tractability or novelty of a scaffold of interest.

■ FINAL REMARKS

Virtual screening is a powerful tool for finding hits for drug discovery projects. Given quality starting information, hit rates approaching 10% are common. With this success in mind, virtual screening has the potential to play an important role in the lead optimization stage. In particular, we imagine a scenario where large areas of virtual chemical space are constantly being virtually screened against complicated 3D queries and property based filters with the best results being presented to the team every few weeks. In order for this to be successful the entire drug discovery team must be engaged so that the results can be pushed toward areas of chemistry space that not only match the 3D query but also fit the needs of the team including chemical tractability, novelty, appropriate physical characteristics etc. Ultimately, by engaging the full team the virtual screening results can suggest new molecules and more importantly spur the creativity of the team.

dx.doi.org/10.1021/ci500532j | J. Chem. Inf. Model. 2014, 54, 3446−3452

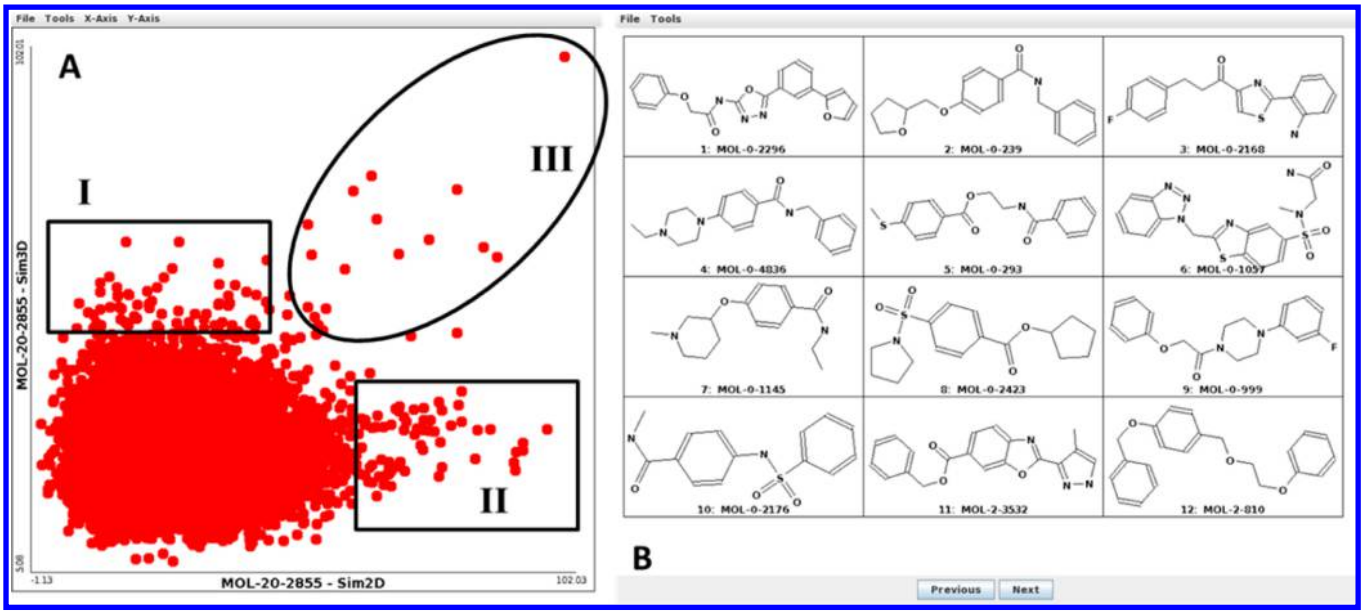


Figure 5. Results of a similarity comparison to the Mol-20-2855 hit. Once a hit of interest is selected it appears in the comment viewer. From the comment viewer the user can calculate both 2D and 3D similarities to the chosen hit. This data is then available within the scatter plot as shown. (A) Scatter plot of the 3D (*y*-axis) similarity versus the 2D (*x*-axis) similarity to Mol-20-2855. Here we have highlighted three regions. Region I is those molecules with similar docked poses but different chemical structures. Region II is those molecules with similar chemical structures but different docked poses. Region III is those molecules with similar chemical structures and similar docked poses. In the search for new scaffolds region I is the most interesting. (B) Region I is selected, and the molecules then appear in the 2D molecular grid view where they can be sorted for example from least chemically similar to the most chemically similar. This allows the user to select structures that are chemically distinct from the original hit but docked in a similar fashion. See Figure 6 for examples of the 3D alignments of chemically distinct but similarity docked molecules.

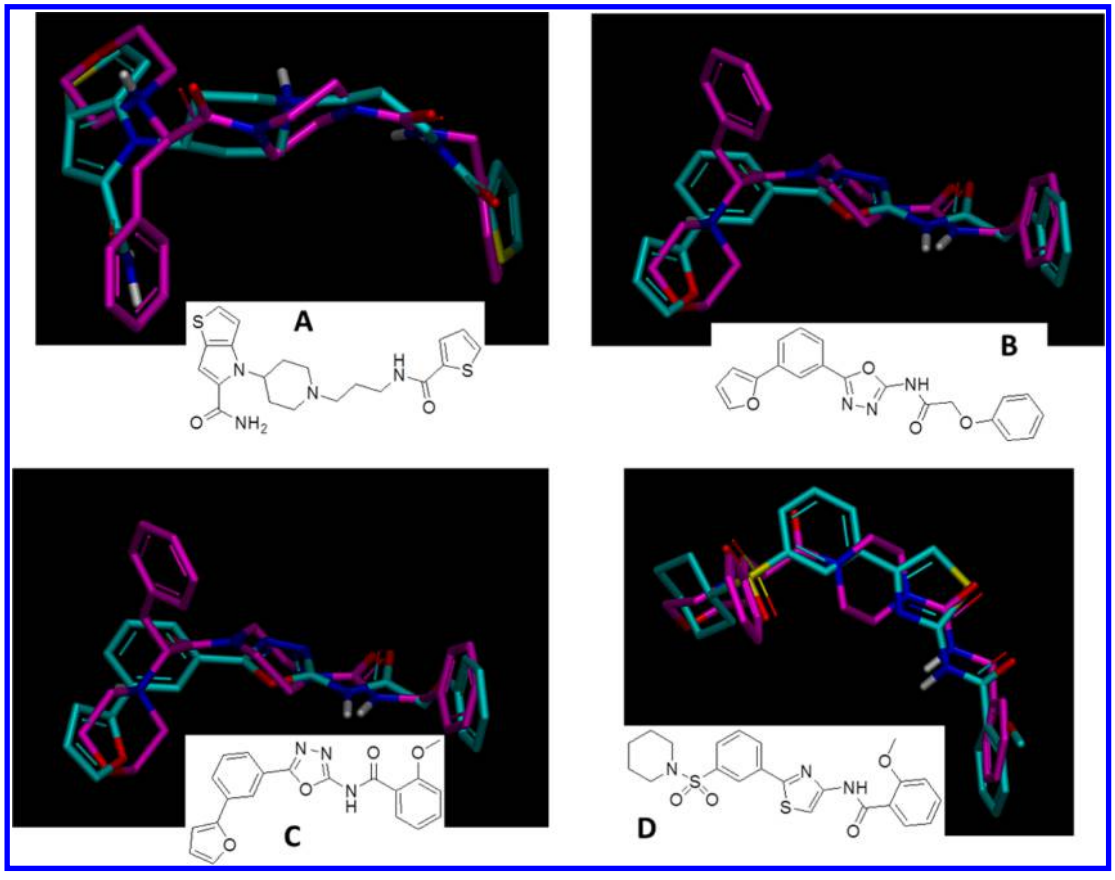


Figure 6. Examples of chemically distinct hits but similar docked poses to Mol-20-2855. Being able to examine different structures with similar docked poses allows the user the opportunity to mix and match molecules to potentially achieve a superior fit or a more chemically tractable scaffold. All images are saved directly from the Astex Viewer.

dx.doi.org/10.1021/ci500532j | *J. Chem. Inf. Model.* 2014, 54, 3446−3452

To address this need, we have developed and here we describe a lightweight application for interactively viewing 3D virtual screening results. The application is written in java and uses the powerful 3D Astex Viewer making it platform independent. Our hope is that this application is a step toward integrating virtual screening into the main stream of lead optimization. To this end, the VSviewer3D has been made available under the BSD license and is freely accessible to anyone interested. The code is available via source forge either by searching for VSViewer3D or via the link https://sourceforge.net/projects/vsviewer3d. This includes the java source for the VSViewer3D and a utility program that relies on the CDK toolkit[6] to convert 3D SD files from virtual screens to VSV files. The VSViewer3D is capable of reading VSV files for exploring 3D virtual screening data, 2D SD files for exploring structure/property relationships, and CSV files for exploring data in the absence of structure. The VSviewer3D has been tested on computers running either Windows or Linux machines with at least Java 6.0 installed.

AUTHOR INFORMATION

Corresponding Author

*E-mail: djrdiller@gmail.com. Phone: 609-216-4311.

Present Address

‡D.J.D.: CMDBioscience Inc., 5 Science Park, New Haven CT 06511

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

We gratefully acknowledge Robert Kirk DeLisle both for testing versions of the program and for many helpful suggestions in preparing the manuscript.

REFERENCES

(1) Kutchukian, P. S.; Vasilyeva, N. Y.; Xu, J.; Lindvall, M. K.; Dillon, M. P.; Glick, M.; Coley, J. D.; Brooijmans, N. Inside the mind of a medicinal chemist: the role of human bias in compound prioritization during drug discovery. *PloS one* **2012**, *7*, e48476.

(2) Lajiness, M. S.; Maggiora, G. M.; Shanmugasundaram, V. Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *Journal of medicinal chemistry* **2004**, *47*, 4891−6.

(3) Hartshorn, M. J. AstexViewer: a visualisation aid for structure-based drug design. *Journal of computer-aided molecular design* **2002**, *16*, 871−81.

(4) Bandaranayake, R. M.; Ungureanu, D.; Shan, Y.; Shaw, D. E.; Silvennoinen, O.; Hubbard, S. R. Crystal structures of the JAK2 pseudokinase domain and the pathogenic mutant V617F. *Nature structural & molecular biology* **2012**, *19*, 754−9.

(5) Landrum, G. RDKit: Cheminformatics and Machine Learning Software. http://www.rdkit.org/ (accessed August 7, 2014).

(6) Steinbeck, C.; Han, Y.; Kuhn, S.; Horlacher, O.; Luttmann, E.; Willighagen, E. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences* **2003**, *43*, 493−500.