
PYTHON FOR STRUCTURAL BIOINFORMATICS

Tutorial presented by
Sophie COON & Michel SANNER

The Molecular Graphics Laboratory
The Scripps Research Institute, La Jolla, CA

After a brief overview of our software development strategy, the attendees will learn how to use and extend one of our tools: the Python-based Molecular Visualization environment (PMV).

The goals of this tutorial include: getting the user comfortable using PMV, presenting some of the features that set PMV aside from other molecular viewing programs and demonstrating using the Python programming language to extend the application.

Although prior knowledge of the Python programming language will be helpful, it is not a strict requirement for the tutorial.

I. INTRODUCTION

During the first part of the tutorial, we will present the scientific interests of our laboratory and the software engineering challenge that we are facing. We will then describe the component-based software development strategy that we have developed in response to this challenge.

We will give a short introduction of the Python programming language that will cover the features used later on in the tutorial.

II. PMV DEMO

During the second part of the tutorial, we will demonstrate the Python-based Molecular Visualization environment (PMV). This, general purpose, molecular viewer has been developed in our laboratory. This section will be based on a live demo based on the tutorial provided in appendix A of this handout.

III. PMV BUILDING BLOCS

In this section we will describe some of the PMV building blocs: The MolKit package which deals with the representation and manipulation of molecules, the 3-Dimensional visualization package DejaVu and the ViewerFramework package. We will conclude this section by showing how those building blocs have been combined to create PMV.

IV. EXTENDING PMV

In the last part of the tutorial we will describe extending PMV using various approaches.

We will:

- use the log strings generated by PMV commands to create resource files
- use the python programming language to write a script allowing the user to color the molecular surface of the HIV protease by the shortest distance to all the atoms of a ligand
- transform this script into a PMV macro, which is a simple command
- point out the steps necessary to turn this macro into a full fledged PMV command

Finally, we will briefly introduce the visual-programming environment ViPEr, and demonstrate how it can be used to create a computational network corresponding to the distance calculation script described above.

CONCLUSION

We will conclude the tutorial by discussing the validity of our approach, the choice of the Python programming language.

There will be time for questions after the end of each section.

DOWNLOAD AND INSTALL THE MGLTOOLS ON YOUR MACHINE

PMV is part of larger collection of tools called MGLTOOLS which also contains the visual programming environment ViPER among other things. We strongly that you download a complete release with the Python interpreter (even if you already have a python interpreter installed). This will approach makes it trivial to get PMV and ViPER running.

To install the latest version PMV and ViPER do the following:

- 1- Go to the MGL [download site](http://www.scripps.edu/~sanner/python/packager.html) (<http://www.scripps.edu/~sanner/python/packager.html>)
- 2- Login in our the distribution site (you will need to register if this is your first download)
- 3- Download and install the completeRelease of the MGLTools for the platform of your choice (sgi, windows, sun, pc linux and macOSX). This distribution includes the precompiled binary of the python interpreter version 1.5 for the unix platform and 2.0 for windows and mac. Follow the installation instruction located:

(http://www.scripps.edu/~sanner/python/documentation/completeRelease_instruction.html)

FOR ALL PLATFORMS EXCEPT MACOSX

- 4- Download and install the latest update of the MGLTools.

FOR THE MACOSX

- 4- The distribution for MACOSX already contains the latest update.

FOR ALL PLATFORMS

Follow the update instruction located:

http://www.scripps.edu/~sanner/python/documentation/updateMGLTools_instruction.html

If you need help installing these tools, please contact Sophie Coon (sophieco@scripps.edu).

When PMV is started the window shown in Figure 1 appears on the screen.

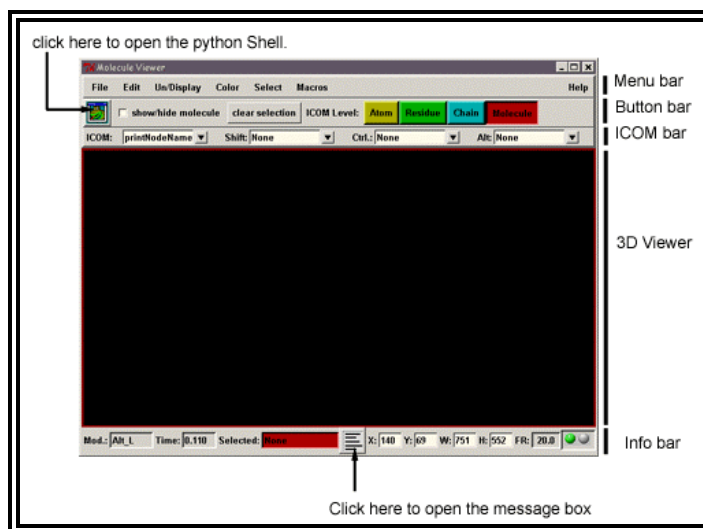


Fig 1: Pmv at start up.

APPENDIX A: PMV TUTORIAL

I- BASIC COMMANDS:	3
EX 1: READING THE FIRST MOLECULE:	3
EX 2: BASIC INTERACTIONS WITH THE VIEWER:	3
EX 3: BASIC GRAPHICAL REPRESENTATIONS OF A MOLECULE IN PMV:	4
1- COLOR COMMANDS	4
2- DISPLAY COMMANDS:	5
II- PMV FUNDAMENTALS	6
EX 1: THE SELECTION IN PMV	6
EX 2: UNDOING LAST COMMANDS	8
EX 3: LOADING NEW COMMANDS AND MODULES IN PMV:	8
EX 4: PICKING COMMANDS :	9
III- ADVANCED MANIPULATIONS	9
EX 1: SECONDARY STRUCTURE REPRESENTATION:	9
EX 2: REPRESENTING THE PROTEASE BY A MOLECULAR SURFACE USING MSMS	10
EX 3: COLOR MSMS BY TEMPERATURE FACTOR (COLORBYPROPERTY)	11
EX 4: LOADING A NEW MOLECULE	12
EX 5: ADDING HYDROGEN AND COMPUTE THE GASTEIGER CHARGES ON THE LIGAND:	12
EX 6: HIDING THE PROTEASE:	13
EX 7: CHANGING THE BACKGROUND COLOR:	14
EX 8: SAVING AN IMAGE:	14

(a)

V. BASIC COMMANDS:

When PMV is started, a number of menus appear on the menu bar which indicated that commands have automatically been loaded at start up:

- File Commands under the File menu
- Color Commands under the Color menu
- Display Commands under the Display menu
- etc...

The user can choose which modules and commands are to be loaded at start up, this will be presented in the section "Extending PMV".

A. Reading the first molecule:

PMV has an extensible set of molecular data files parsers. It currently supports reading:

- PDB from the Protein Data Bank.
- PDBQ and PDBQS AutoDock formats.
- PQR Mead format.
- Mol2 Tripos format.

step 1: Read the "protease.pdb" file describing the HIV protease.

- ⌘ File -> Read Molecule
- ⌘ select **protease.pdb** in the file browser
- ⌘ click on the Open button to read the file.

The red LED in the bottom right corner of the application turns red while PMV is reading the file. This LED turns red when the program is busy, after it completes a task the LED turns back to green. The time it took to PMV to execute the command is printed in the entry box labeled "Time" on the info bar.

The HIV protease is now loaded in PMV and displayed as lines (Fig 1.).

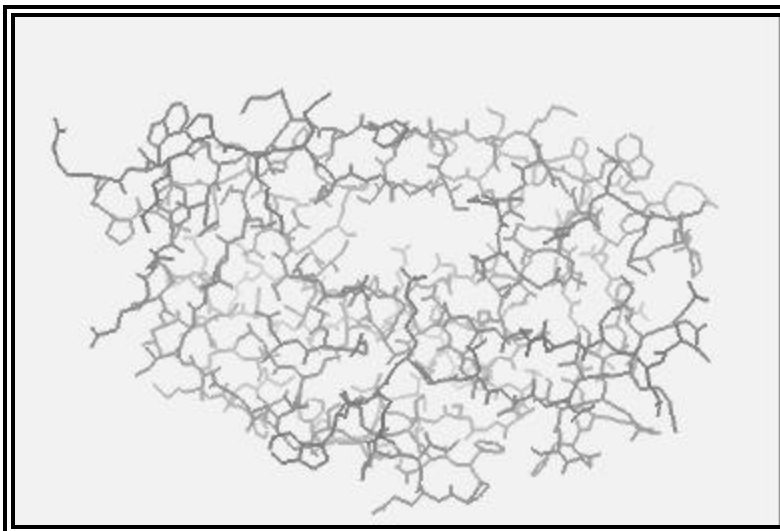


Fig 2: HIV protease (protease.pdb) displayed as lines in PMV.

B. Basic interactions with the viewer:

Once a molecule is loaded in the application, the user can interact with it using the mouse buttons. The table below lists the default mouse bindings for a 3-buttons mouse. On a laptop

with a 2-buttons mouse pad, clicking on the two buttons simultaneously usually emulates the middle button of the 3-buttons mouse.

Action	Mouse Button	Modifier
Rotation	Middle	-
Scaling	Middle	SHIFT
XY Translation	Right	-
Z Translation	Right	SHIFT
Picking	Left	-

Those bindings can be modified by using the "bind action to mouse button" command that is located under the "File" menu.

If the molecule moves out of the viewport, the user can always do the following to get it back:

- ✍ Use the **R** or **r** key to reset all the transformation performed so far on the molecule.
- ✍ Use the **N** or **n** key of your keyboard to normalize the scene.
- ✍ Use the **C** or **c** key of your keyboard to set the center of rotation at the center of the molecule. to the center of your molecule.
- ✍ The **D** or **d** key of your keyboard allows you to turn on or off the depth cueing.

C. Basic graphical representations of a molecule in PMV:

1. Color commands

One way to alter the representation of a molecule is to change its color. PMV provides several coloring schemes available in the "colorCommands" module. Some of these commands have been loaded by default in PMV and are available under the "Color" menu.

step 2: *Color the protease by atom type*

This coloring scheme gives information on the protein atomic composition.

- ✍ Color -> By Atom type
- ✍ Select the check box labeled **lines** to color the lines geometry representing the molecule. This is the only choice because right now the protease is only represented by lines
- ✍ Click on the OK button to carry out the color command.

Now all the Carbon atoms of the protease are colored in gray, the Oxygen atoms in red, the Nitrogen atoms in blue and the Sulfur atoms in yellow.

step 3: Color the protease by residue type

Coloring the protease by residue type using the **Rasmol** (Fig 2A) or the **Shapely** (Fig 2B) coloring schemes will give the user some information on the residue composition of the protein. In both schemes the polar residues, which can be found on the surface of the protein, have a bright color whereas the non-polar residues have a darker color.

ALA	■	ARG	■	ASN	■	ASP	■	ALA	■	ARG	■	ASN	■	ASP	■
CYS	■	GLN	■	GLU	■	GLY	■	CYS	■	GLN	■	GLU	■	GLY	■
HIS	■	ILE	■	LEU	■	LYS	■	HIS	■	ILE	■	LEU	■	LYS	■
MET	■	PHE	■	PRO	■	SER	■	MET	■	PHE	■	PRO	■	SER	■
THR	■	TRP	■	TYR	■	VAL	■	THR	■	TRP	■	TYR	■	VAL	■
RASMOL COLORING SCHEME.							SHAPELY COLORING SCHEME.								

Fig 3: (A) Rasmol Coloring scheme, (B) Shapely Coloring scheme

step 4: Color the protease by Chain or by Molecule

Coloring a molecule by chain allows the user to distinguish the various parts of a multimeric structure whereas coloring by molecules will distinguish the different molecules loaded so far in the application.

Both the **By Chain** and **By Molecule** coloring schemes use a predefined palette of 20 different colors.

2. Display Commands:

The user can also alter the representation of a molecule by changing the geometry used to display it. PMV provides a set of display commands available in the displayCommands module. A subset of these commands has been loaded at start up and is located under the Un/Display menu. In PMV, a molecule can be represented by several geometries at the same time and each geometry can be colored using a different color scheme.

step 5: Display the protease by CPK.

- ✍ Un/Display -> CPK
- ✍ In the display panel: select "display", set the scale factor to 1, and the quality to 10.
- ✍ Click on the OK button to carry out the command with the chosen parameters.

Each atom is, now represented by a sphere. This representation is also known as "Space filling" representation.

step 6: Color this new geometry using the "Shapely" coloring scheme

- ✍ Color -> By Residues -> Shapely
- ✍ In the Choose Geometry panel select the "CPK" geometry
- ✍ Click on the OK button to carry out the coloring command.

step 7: Display the protease by sticks and balls

When representing a molecule by "sticks and balls" each bonds is displayed as a cylinder and each atom as a sphere. The following parameters can be set by the user:

- the size of the spheres
- the quality of the spheres (number of triangles used to represent the sphere)

- the size of the cylinder.

step 8: *Undisplay CPK and sticks and balls*

VI. PMV FUNDAMENTALS

A. The selection in PMV

In PMV, commands are always applied to the **current selection**. However for convenience when nothing is selected, commands are applied to all the molecules in the viewer. This behavior can be modified through the setUserPreference command located under the File menu. In this section the user will learn how to identify what the current selection is and how to create one.

In PMV, the molecules are represented by a 4 levels hierarchical tree, which mirrors the inner hierarchy of a protein. These four levels are Molecule, Chain, Residue and Atom.

On the right hand side of the button bar there are 4 buttons labeled: "Molecule", "Chain", "Residue" and "Atom". These buttons allow the user to choose at which level of the molecule a command must be carried out.

A set of selection commands has been loaded by default at start-up and is available under the Select menu.

step 9: *Select the chain B of the protease using the Select From String commands*

- ✍ Select -> From String
- In the "Select From String" panel (Fig 4)
- ✍ click on **Chain List...** in the dropdown menu select the **'A'** checkbox. The type-in entry labeled "Molecule" now contains "protease" string and the type-in entry labeled "Chain" the word "B".
 - ✍ Click on the **Select** button to perform the selection
 - ✍ In the panel that appears click **OK** to change the level of the selection from Molecule to Chain.

Several things happened:

The Chain Button level is now selected

The background color of the entry labeled selection is now Cyan and the selection holds 1 Chain.

Yellow crosses are now displayed on each atom of the current selection.

step 10: *Display the chain B by CPK and color by Residues using the Shapely coloring scheme*

- ✍ Display -> CPK
- ✍ Color -> By Residues -> Shapely

As expected only the chain B of the protease is now displayed by CPK and the CPK geometry is colored by residue type.

The current selection stays active until the user clears it using the clearSelection command that can be found on the button bar.

step 11: *Clear the selection.*

- ✍ Button bar -> Clear Selection

The "SelectFromString" command is a powerful command and as mentioned earlier the user can type a string in each entry that will be matched against the name of the nodes (Fig 3). The '*' character replaces any character, ',' allows the user to specify a list of nodes and '-' a range of nodes.

step 12: Select all the carbons all the oxygen atoms of all the isoleucine residues of the chain A of the protease.

- ✂ Select -> Select From String
- ✂ Click on the Clear Form and Clear Selection.
- ✂ In the type-in entry labeled Atom type: O*,C*
- ✂ In the type-in entry labeled Residue : ILE*
- ✂ In the type-in entry labeled Chain : A
- ✂ In the type-in entry labeled Molecule: protease
- ✂ Click on Select to perform the selection.
- ✂ Click on the OK button to change the level of the selection from Chain to Atom.

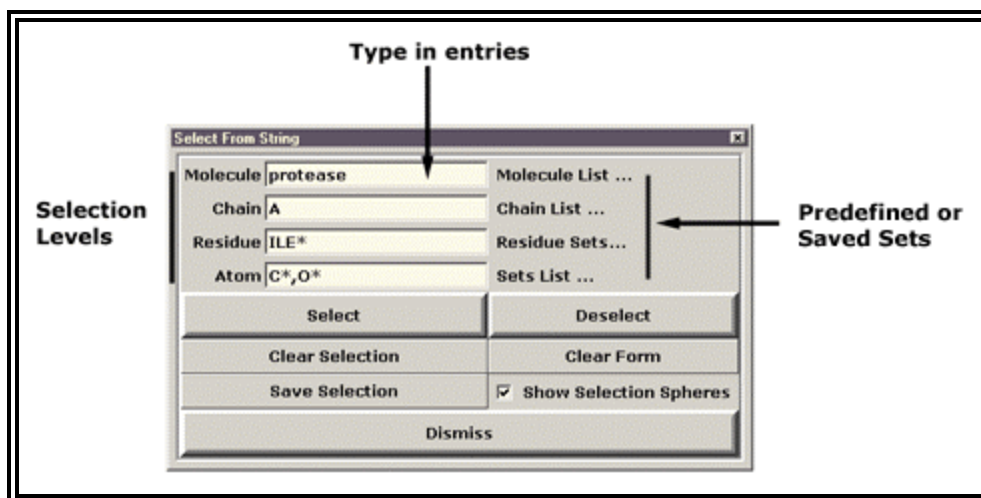


Fig 4: Select From String GUI.

- **type-in entries:** corresponding to the 4 levels of selection to allow the user to enter a string that will be matched against the name of the corresponding node.
 - Molecule : name of the file without the extension
 - Chain : chain ID of the pdb file
 - Residue : 3 letter AA name concatenated to the residue number and the residue seq ID
 - Atom : element type and the letter code and number for its position in the residue (CA for carbon alpha, O1).
- List of sets for each level is available by clicking on the label on the right hand side of the type in entry.
- **Select** button :to add to the current selection
- **Deselect**: to remove from the current selection,
- **Clear Selection**: to clear the current selection
- **Save Selection**: to save the current selection as a set
- **Clear Form**: to clear the form
- **Show Selection Spheres**: toggle on or off the selection spheres which are the yellow crosses.
- **Dismiss** button to withdraw the form.

Now the selection level is the "Atom" level and the current selection holds 26 atoms. If the user hadn't cleared the previous selection the new selected atoms would have been added to the previous selection.

In PMV the current selection is homogeneous which means that two residues and three atoms cannot be selected at the same time. If the user changes the selection level from Atom to Residue, the current selection will be expanded to all the residues having at least one atom in the current selection. When going back to the Atom level, all the atoms of the selected residues will now be part of the current selection.

B. Undoing last commands

Most of the PMV commands are undoable using the undoCommands available under the "Edit" menu.

step 13: To undo the two "setlcom Level" commands

- ✍ Edit -> Undo Setlcom Level (to undo the command that changed the selection level from Residue to Atom)
- ✍ Edit -> Undo Setlcom Level (to undo the command that changed the selection level from Atom to Residue.)

PMV keeps a stack of the last undoable commands that have been executed. By default the stack can hold up to 100 commands. This number can be modified through the setUserPreference command. The stack is reset when a molecule is deleted.

C. Loading new commands and modules in PMV:

There are many more commands available in PMV than the subset used so far,. New commands can be loaded dynamically by the user in the application at any time. Commands are grouped into modules, which are in turn grouped into packages.

Either a whole module can be loaded at once in the application using the loadModule command or individual commands using the loadCommand command.

step 14: Load the secondaryStructureCommands module into PMV:

- ✍ File -> loadModule
- ✍ Select **PMV** in the "Package" combobox, the secondaryStructureCommands being a PMV module.
- ✍ Select **secondaryStructureCommands** in the "Choose a module" listChooser.
- ✍ click on the **loadModule** button to load the module or double click on the **secondaryStructureCommands entry** itself to load the module.
- ✍ click on the **DISMISS** button to withdraw the panel.

A new menu "Compute" appears on the menu bar with a cascade called "secondary structure". The menus "Color" and "Un/display" have also two new entries "By Secondary Structure Type" and "Secondary structure".

step 15: Load the computeMsmsMol and displayMsmsMol, undisplayMsmsMol commands in PMV:

- ✍ File -> loadCommand
- ✍ Select **PMV** in the package combobox msmsModule being a PMV module.
- ✍ Select **msmsCommands** in the Choose a module ListChooser, the two commands we want to load are both part of the msmsCommands module.
- ✍ Select **msmsMol**, **displayMsms** and **undisplayMsms** in the Choose Command ListChooser

☞ Click on the **loadCommand** button to load the commands in PMV and the **DISMISS** button to withdraw the panel.

A new cascade named "msms" under the "Compute" menu has been created and a new entry named "Msms Mol" under the "Un/display" menu has been created as well.

D. Picking commands:

So far all the commands were applied to a selection, which was created using one of the select commands available in PMV.

Another way to carry out a command is to use the mouse. PMV maintains the relationship between the data structures representing a molecule and the various graphical representations of that molecule (called geometries). This relationship makes it possible to indicate molecular fragments by picking or dragging a box around their graphical representation. Such commands are called picking commands (PCOM). Any command that has only one required argument that describes a molecular fragment, can become a picking command. The printNodeName command is bound to the picking event by default. At the Molecule level when picking on the lines representing the molecule the name of the molecule will be printed in the message box. To open the message box click on the "message box" icon at the bottom of pmv (Fig1). Changing the PCOM level from Molecule to Atom will print the full names of the underlying atoms. Other command can be bound to the mouse and a modifier using the PCOM combobox.

step 16: Bind the select command to the mouse:

☞ ICOM -> None combobox
☞ Choose "select" in the first PCOM combobox of the button bar labeled None.

The PCOM level is Atom so if the user picks on the lines, the underlying atoms will be added to the current selection. In fact, the number of atoms in the current selection increases.

Picking commands can be also bound the picking event combined to keyboard modifiers such as SHIFT, CTRL, and ALT.

VII. ADVANCED MANIPULATIONS

A. Secondary Structure representation:

In the next step the user will represent the protease by a traditional ribbon diagram. The secondaryStructure commands module has been loaded previously.

step 17 Select the protease using the interactive command "select" as described above.

step 18 Represent the current selection by a traditional ribbon diagram

☞ Compute -> secondary structure -> Ribbon
☞ Un/Display -> Lines -> undisplay

The protease is now represented by a traditional ribbon diagram where the:

- beta strands are represented by an extruded arrow
- helices by a extruded rectangle
- coils and Turns by a extruded circle.

In this representation, the two flaps of the HIV protease are really visible (Fig 4.)

The Secondary structure geometries, like any other geometry in PMV, can be colored using one of the available coloring commands.

step 19 Color the protease by secondary structure type

☞ Color -> Secondary Structure Type -> secondary structure -> OK

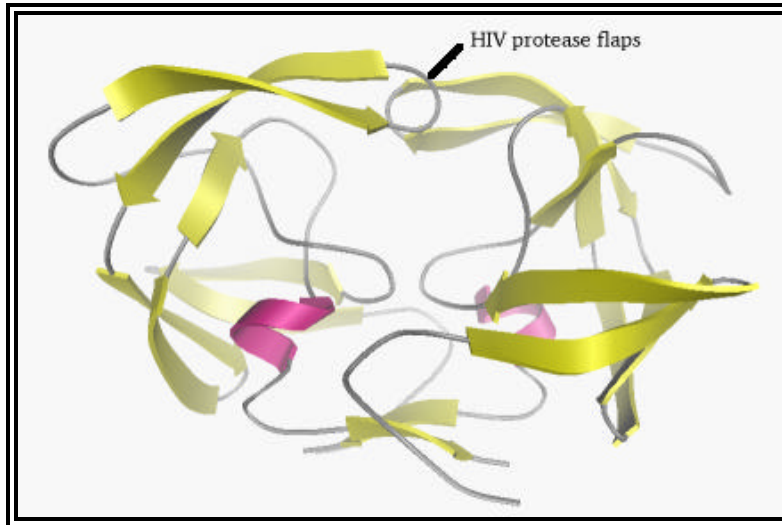


Fig 5: HIV protease represented as a traditional ribbon diagram colored by secondary structure. (beta strand: yellow, alpha helices: pink, coil: gray, turn: blue)

The "Ribbon" command is actually a shortcut to display the current selection using a traditional ribbon diagram. The "Ribbon" command executes first the "computeSecondaryStructure" command on the current selection then the "extrudeSecondaryStructure" command. The secondary structure can be represented by other geometries.

step 20: Represent the helices by an extruded ellipse:

- ✍ Select -> Select a set
- ✍ Select all the helices entry in the Choose an item using the Shift and Ctrl key to select several item at a time then click on the OK button to carry out the selection.
- ✍ Compute -> Secondary Structure -> Extrude Secondary Structure
- ✍ Select ellipse in the Choose a shape listchooser then click on the OK button
- ✍ keep or set the size of the demi-grand axis
- ✍ keep or set the size of the demi-small axis
- ✍ Add or not a front and /or a end cap
- ✍ Click on the **OK** button to extrude the chosen shape.

Now the helices of the protease are represented by extruded ellipses. Although, each secondary structure elements can use its own shape for the extrusion, only one shape is allowed by secondary structure element.

B. Representing the protease by a Molecular surface using MSMS

To represent the protease using a molecular surface the user will need the MSMS command loaded earlier.

step 21: Select the protease using the "Direct Select" command:

- ✍ Select -> Direct Select
- In the "Direct Select" panel that appears
- ✍ select **protease** in the Molecule List drop down check box
 - ✍ click on the **Dismiss** button to withdraw the form.

step 22: Compute the solvent excluded surface of the molecule in the current selection

- ✍ Compute -> Msms Mol
- ✍ In the panel that appears set the following parameters.
 - 1.5 for the probe radius
 - 3.0 for the density of triangles used to represent the surface.
- ✍ Click on the **OK** button to compute the surface of the molecule in the current selection.

The protease has now an additional geometry representing the molecular surface. This geometry can be colored using any coloring command available in PMV.

C. Color msms by temperature factor (ColorByProperty)

The introspection capability of Python allows to find out at runtime what attributes of molecular fragments can be used for coloring a given geometry. The ColorByProperty command interactively builds a panel that lets the user choose the property to be used for coloring. The list of properties depends on the selection level (Atoms have different attributes that Residues).

step 23: Select the protease using one of the selection command

step 24: Load the colorByProperty command using the load command command.

- ✍ File -> LoadCommand -> ColorCommands -> ColorByProperty & ColorByExpression

step 25: To color the msms surface by temperature factor

- ✍ Color -> ColorByProperty -> msms
- In the "Color by properties" panel that appears:
 - ✍ Select the Atom level to display the atoms properties in the "Property listchooser"
 - ✍ Select temperatureFactor in that list
 - ✍ Select the "Edit" radiobutton to edit the colormap
 - ✍ Click on the OK button to carry out the color command.

The msms is now colored by temperature factor using a RGB ramp. The user can modify this color map using the colormap editor.

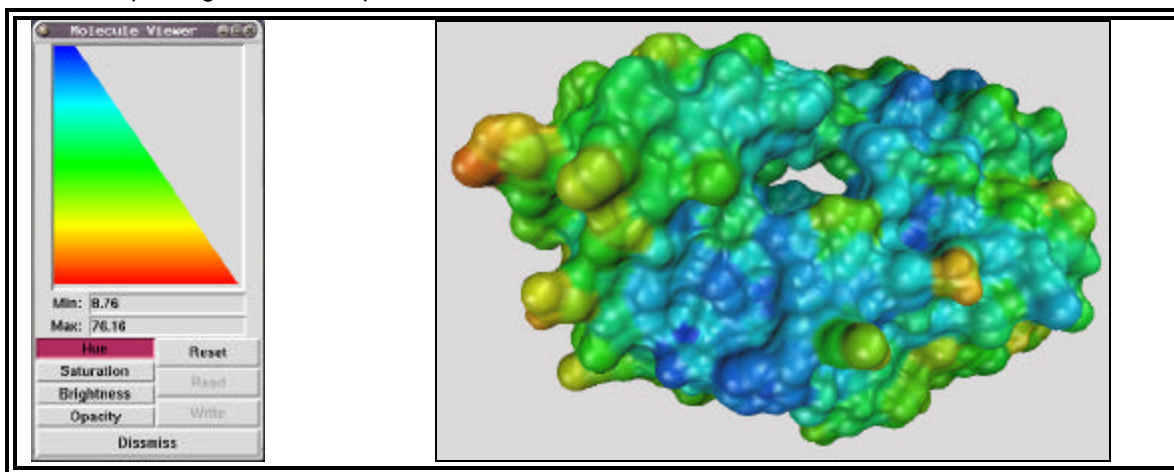


Fig 6: HIV protease (protease.pdb) represented by a msms surface colored by the atom property temperature factor using a RGB ramp.

Using this editor the user can for example decrease the opacity of the low values of the color map therefore the blue regions of the surface will then be transparent.

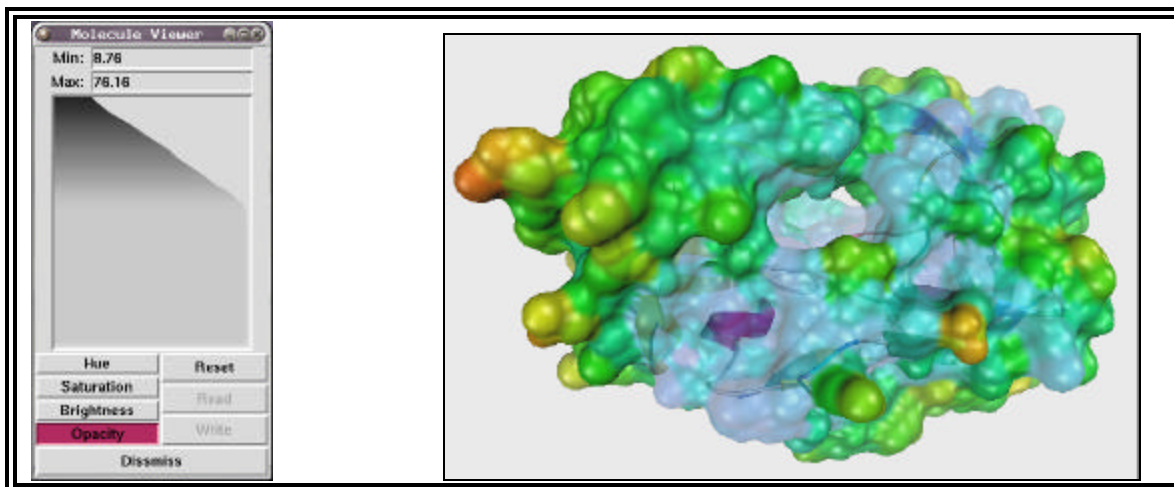


Fig 7: HIV protease (protease.pdb) represented by a msms surface and secondary structure, colored by the atom property temperatureFactor using a RGB ramp where the low values are made more transparent.

D. Loading a new molecule

The molecular surface allows nicely shows the tunnel forming the active site of the protease. We will now load one of the HIV protease inhibitors (indinavir) and to display it using a sticks and balls representation.

- ✂ File -> readMolecule -> indinavir.pdb
- ✂ Select -> Direct Select -> Molecule List ... -> indinavir
- ✂ Un/Display -> Lines -> undisplay
- ✂ Un/Display -> Sticks and Balls
- ✂ Set the cylinder radius at 0.20
- ✂ Set the balls radius at 0.40
- ✂ Set the balls quality at 20

E. Adding hydrogen and compute the gasteiger charges on the ligand:

step 26: Select the ligand "indinavir" using one of your favorite selection commands.

step 27: Load the new commands Add Hydrogens and Compute Gasteiger.

These two commands are in the edit commands along with all the commands implementing functionality to edit a molecule.

- ✂ File->loadModule->editCommands

step 28: Add Hydrogens to the ligand indinavir:

- ✂ Edit -> hydrogens -> add
- ✂ In the "Add hydrogens" panel that appears on the screen select:

- ✍ All hydrogens
- ✍ noBondOrder because indinavir was read from a PDB file
- ✍ yes to renumber atoms to include new hydrogens

step 29: *Compute Gasteiger charges.*

- ✍ Edit -> Charges -> Compute Gasteiger charges.

step 30: *Color the ligand by charges.*

The charge property of an atom is stored into a dictionary therefore we cannot use the color by property command. However use the colorByExpression command and access this property using a python expression.

- ✍ LoadCommand -> Color -> By Expression

The color by expression commands allows you to type in a python expression that will be evaluated at the chosen level. The result of the evaluation needs to be a list of numeric values to create a color map to color the chosen geometry

- ✍ Color -> By Expression
- ✍ Select Atom level, the charges property being an atom property
- ✍ In the text widget uncomment the line lambda ... by deleting the # character
- ✍ Click on the "eval expression" to evaluate the given expression
- ✍ Choose the RedWhite ramp to color the indinavir by gasteiger charges
- ✍ Click OK to carry out the color command.

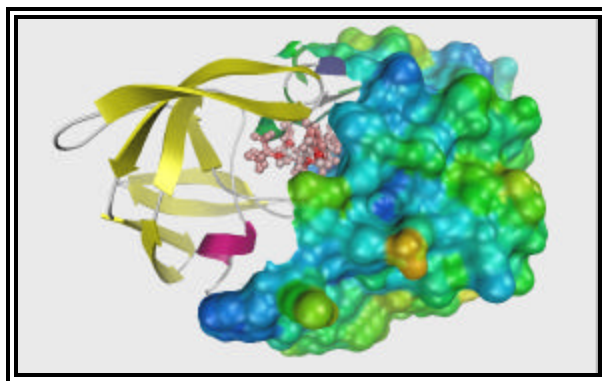


Fig 8: Indinavir represented by sticks and balls and colored by Gasteiger charges.

The chain A of the protease is represented by traditional ribbon diagram colored by secondary structure elements and the chain B is represented by a MSMS surface colored by temperature Factor.

F. Hiding the protease:

When a molecule is represented by several geometries, using the show/hide molecule command will undisplay / display all the geometries at the same time, effectively hiding/showing the molecule.

- ✍ Menu bar -> Show/Hide Molecule
- ✍ protease

G. Changing the background color:

step 31: Open the DejaVu GUI

☞ Macro -> DejaVu -> Show/Hide DejaVu GUI

step 32: Change the background color

☞ Property-> Camera-> Background color

H. Saving an image:

A scene created in PMV can be saved as an image file. Various format are available such as tiff, jpeg etc..

savelImage command.

☞ File-> loadCommand -> fileCommands -> SavelImage

☞ File -> save -> image

Type a file name and choose a file format to save your image. Don't forget to add the extension to the file name.