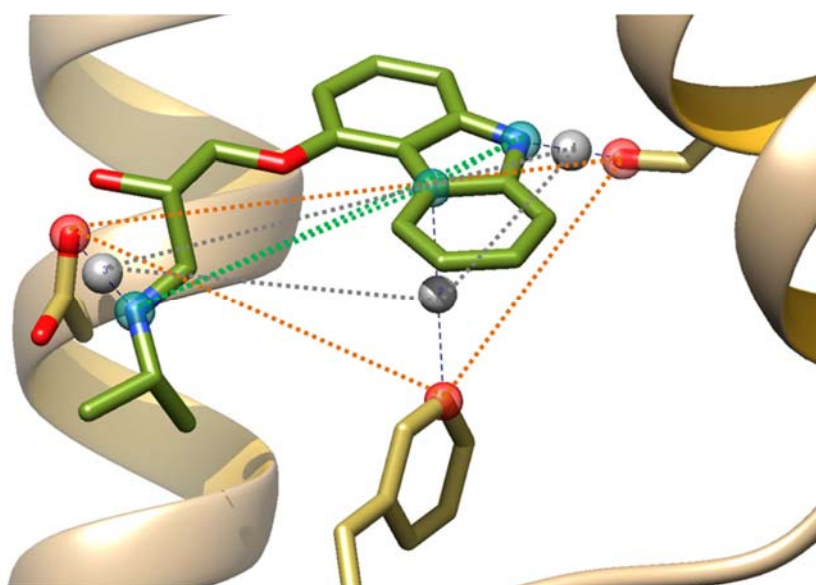


# IChem: A Toolkit for detecting, comparing and predicting protein-ligand interactions



**Jérémy DESAPHY, Franck DA SILVA, and Didier ROGNAN**

*Laboratoire d'Innovation Thérapeutique, UMR 7200 CNRS-Université de Strasbourg, F-67400 ILLKIRCH*

Email: rognan@unistra.fr



## Literature Corner

Please have a look at these articles for detailed information on the basic principles and concepts underlying IChem usage

- Marcou, G. and Rognan, D. (2007) Optimizing fragment and scaffold docking by use of molecular interaction fingerprints. *J. Chem. Inf. Model.*, **47**, 195-207



- Desaphy, J., Azdimousa, K., and Rognan, D. (2012) Comparison and druggability prediction of protein-ligand binding pockets from pharmacophore-annotated shapes. *J. Chem. Inf. Model.*, **52**, 2287-2299



- Desaphy, J., Ducrot, P., Raimbaud, E. and Rognan, D. (2013) Encoding protein-ligand interaction patterns in fingerprints and graphs. *J. Chem. Inf. Model.*, **53**, 623-637



- Desaphy, J. and Rognan, D. (2014) scPDBFrag: a database of protein-ligand interaction patterns for bioisosteric replacements. *J. Chem. Inf. Model.*, **54**, 1908-1918



- Gabel, J., Desaphy, J. and Rognan, D. (2014) Beware of machine learning-based scoring functions - On the danger of developing black boxes. *J. Chem. Inf. Model.*, **54**, 2807-2815



- Da Silva, F., Desaphy, J., Bret, G. and Rognan, D. (2015) IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein-Protein Interfaces. *J. Chem. Inf. Model.*, **55**, 2005-2014



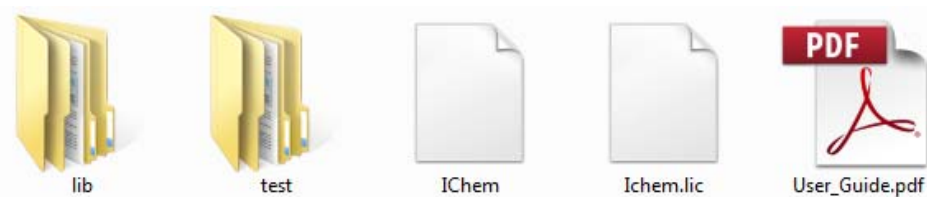
- Slynko, I. Da Silva, F., Bret, G. and Rognan, D. (2016) Docking pose selection by interaction pattern graph similarity: application to the D3R grand challenge 2015. *J. Comput.-Aided Mol. Des.*, **30**, 669-683.





## Installation

IChem is provided as a zipped archive file (**IChem.tgz**) containing the following material:



- **lib**: a directory containing necessary template files
- **test**: a directory containing some test input/output files
- **Ichem.lic** a license file
- **IChem**: A 64-bit Linux executable (CentOS 7.1)
- **User\_Guide.pdf**: this manual

Source some environment variables and untar the IChem distribution in the IChem root directory

For csh users: *setenv ICHEM\_DIR /your\_IChem\_root\_directory*  
*cp IChem.tgz \$ICHEM\_DIR*  
*setenv ICHEM\_LIC \$ICHEM\_DIR/Ichem.lic*  
*cd \$ICHEM\_DIR; tar -cvfz IChem.tgz*

for bash users: *ICHEM\_DIR="/your\_IChem\_root\_directory"; export ICHEM\_DIR*  
*cp IChem.tgz \${ICHEM\_DIR}*  
*ICHEM\_LIC=\${ICHEM\_DIR}/Ichem.lic; export ICHEM\_LIC*  
*cd \${ICHEM\_DIR}; tar -cvfz IChem.tgz*

### Note to users:

IChem commands with options/tool/arguments will be displayed in *italic* characters after the ">" prompt

Input/output filenames will be displayed in **bold violet** characters

Terminal output will be displayed with a **gray background**

A copy of all test input/output files is given in the test directory of the IChem distribution. Before using IChem, please take the time to read the description of the technology. Articles to read will be mentioned by the following icon:





## A short introduction to IChem

IChem is a multi-task program for detecting, analyzing and comparing protein-ligand interactions. It is composed of several tools:

Tool	Purpose
<b>Binding mode tools</b>	
<b>IFP</b>	Detect and save interactions as fingerprints (binding-site dependent)
<b>ints</b>	Detect and save interaction patterns( binding-site independent)
<b>grim</b>	Detect and save interactions as graphs
<b>Generic tools</b>	
<b>genkey</b>	License file generator (privileged usage)
<b>license</b>	Output license details
<b>pdbconv</b>	Convert PDB into MOL2 files
<b>realign</b>	Rotation/translation of atomic coordinates
<b>sims</b>	Fingerprint comparison
<b>utils</b>	Miscellaneous (buried surface area, ligand fragmentation)
<b>Cavity detection</b>	
<b>Volsite</b>	Cavity detection and druggability/ligandability prediction
<b>Protein-Protein interfaces</b>	
<b>DetectPPI</b>	Cavity detection and druggability/ligandability prediction



## IChem Usage

The correct syntax for using IChem is:

---

```
> IChem options tool arguments
```

---

Alternatively, options, tool and arguments can be stored in a parameter file (any name) that can be called from IChem with the `-F` option:

---

```
> IChem -F parameter_file
```

---

Example of a parameter\_file content

```
options_1 tool_1 arguments_1
options_2 tool_2 arguments_2
...
options_n tool_n arguments_n
```

Multiple IChem commands can be run from a single parameter file, just by adding for every line a novel list of options/tool/arguments

### Verbose and debugging mode:

IChem has been written in order to simplify standard output. If you want more details than that saved in the regular output, please use the `--verb` option:

---

```
> IChem --verb options tool arguments
```

---

Moreover, a debugging mode is available in IChem by calling the `--debug` option with one or more of the following values, separated by a comma:

```
READ  MOL2 file parser
DINT  Binding mode detection
GRID  Three-dimensional grid
MOLD  Molecular analysis
```

---

```
> IChem -debug READ,DINT,MOLD options tool arguments
```

---



## IChem menu

Just typing **\$ICHEM\_DIR/IChem** (or just **IChem** if \$ICHEM\_DIR is sourced in your path) gives you access to the full IChem menu

---

**> IChem**

---

### KEY Generator

**genKey** *year month day allowed\_tools*

year :                    Expiring year (4 digits : 2014)  
month:                    Expiring month (2 digits : 09)  
day :                      Expiring day (2 digits : 23)  
allowed\_tools: each value is separated by space and must have a value of either 1 or 0. Must follow this order  
1st:                        Licence Key generator  
2nd:                        Molecular realignment  
3rd:                        BSA Calculation  
4th:                        IFP generator  
5th:                        Interaction detection  
6th:                        Graph Interaction Matching  
7th:                        VolSite  
8th:                        PDB to MOL2 conversion  
9th:                        Patching MOL2 conversion  
10th:                       Utils  
11th:                       Sims  
12th:                       Scoring  
13th:                       Detection and analysis of PPI

Example :

```
genKey 2014 1 28 0 0 1 0 0 0 0 1 0 1 1 1
      |<-DATE->|<-----TOOLS----->
```

## realign - Molecular alignment

**realign** *rigidM mobilM applied1 applied2*

||-> rigidM : reference molecule to apply alignment to  
 ||-> mobilM : comparison molecule to apply alignment from  
 ||-> applied: molecule to apply rotation/translation to

### [General options]

-gmatch N (NAME) Use graph matching to align  
 NAME Atom Name matching  
 ATMN Atomic Name matching  
 MOL2 MOL2 Type matching  
 CALP CAlpha Atom matching (protein only)  
 --wMob Also outputs the aligned mobilM  
 -rule R

By default, the program will perform an atom by atom match, without taking care of what kind of atom it match. If you want to perform a match by regarding only some atoms, this index\_string is here to do so

ex : -i '2-3|1-6|23-160' Will match the second atom from the reference with the third from the comparison, the first with the sixth ...

## IFP - Interaction FingerPrint

**IFP** *protein ligand*

**IFP** *protein ligand ligand\_ref*

### [General options]

-name N (LIG) Name of the fingerprint -Default: Name of the ligand  
 --polar Detect and output only polar interactions  
 --metal Detect and output only metal interactions  
 --extended Include within the fingerprint:  
 |--> Metal/Acceptor interaction  
 |--> Weak Hydrogen bonds  
 |--> PI-Cation interactions

### [testing options]

-d\_Hb N (3.5) Hbond length (Angstroem)  
 -d\_Hyd N (4.5) Hydrophobic length (Angstroem)  
 -d\_Io N (4.0) Ionic length (Angstroem)  
 -d\_Me N (2.8) Metal/Acceptor length (Angstroem)  
 -d\_Ar N (4.0) Aromatic interaction length (Angstroem)  
 -a\_H N (Pi) HBond angle (rad.)  
 -at\_H N (Pi/3)) HBond tolerance angle (rad.)  
 -a\_ArFF N (Pi) Aromatic Face to Face interaction angle (rad.)  
 -at\_ArFF N (Pi/6) Aromatic Face to Face tolerance angle (rad.)  
 -a\_ArEF N (Pi/2) Aromatic Edge to Face interaction angle (rad.)  
 -at\_ArEF N (Pi/3) Aromatic Edge to Face tolerance angle (rad.)  
 --ligD print all ligand possible interactions

Please note that ligand file can be a multi-mol2 file

## INTERACTION GENERATOR

### *ints prot lig out*

-type (CENT)	Alter positioning output, multiple values are allowed, separated by space
PROT	InterPROT positioning
LIG	InterLIG positioning
CENT	Centered positioning
MERG	Merged all 3 above
-fgps	Fingerprint format
STD	Standard (1 0 21 0 0 3)
SVM	SVM format (1:1 3:21 6:3)
CMP	Compressed (1 [1 21 [2 3)
--small	Compressed fingerprint

### [General options]

-name (prot)	Name of molecule in out file
-logf	Name of log file
-d_Hb N (3.5)	Hbond length (cut-off in Å)
-d_Hyd N (4.5)	Hydrophobic length (cut-off in Å)
-d_Io N (4.0)	Ionic length (cut-off in Å)
-d_Me N (2.8)	Metal/Acceptor length (cut-off in Å)
-d_Ar N (4.0)	Aromatic interaction length (cut-off in Å)
-a_H N (180)	HBond angle (cut-off in rad.)
-at_H N (60)	HBond tolerance angle (cut-off in rad.)
-a_ArFF N (180)	Aromatic Face to Face interaction angle (cut-off in rad.)
-at_ArFF N (30)	Aromatic Face to Face tolerance angle (cut-off in rad.)
-a_ArEF N (90)	Aromatic Edge to Face interaction angle (cut-off in rad.)
-at_ArEF N (60)	Aromatic Edge to Face tolerance angle (cut-off in rad.)
--noMerge	Do not merge hydrophobic interactions
--newH	Less permissive definitions of hydrogen bonds

## GRIM - GRaph Interaction Matching :

<i>grim refProt refLig CompProt CompLig</i>	(1)
<i>grim refInts complInts</i>	(2)
<i>grim refProt refFile dockFile</i>	(3)

### [Note]

- (1) use --multim2 to use multimol2
- (3) refFile & dockFile can be multimol2 files

### [General options]

NOTE : INTERACTION GENERATION General Options also accessible

-rn N (Ref)	Reference name
-cn N (Comp)	Comparison name
--values	Only output score and not alignment
-sim N (0)	Boolean telling whether the pair is similar or not
-outInt (MERG)	Output only one kind of interaction positioning
MERG	All aligned interactions are outputted
LIG	InterLIG positioning
CENT	Centered positioning
PROT	InterPROT positioning
NOTE : outInt useless when used with --values	
-match N (MERG)	Align only with a specific position
MERG	Align with ALL interaction points
LIG	Align only with ligand interaction points



PROT	Align only with protein interaction points
CENT	Align only with centered interaction points
-max N (1)	Maximal number of outputted cliques.
-size N (3)	Minimal size of a clique.
--all_cliques	Detect all cliques and not only maximal one
-score N (FCT)	Scoring method function
STD	Scored by decreasing SumCl and increasing RMSD
FCT	Scored with scoring function
--newH	Less permissive definitions of hydrogen bonds

## VOLSITE - Cavity detection in a mol2 file

**volsite prot lig** (1)  
**volsite prot** (2)

### [General options]

-step N (1.5)	Edge length of each box (Å)
-boxS N (20)	Edge length of the main box (Å)
-b N (55)	Minimal threshold for buriedness
-name N	PDB Name for output cavity names
-n N (5)	Minimal neighbours for buried cavity boxes
-nPTS N (35)	Minimal number of cubes to consider it a cavity
--dna	Consider DNA as part of the protein
--cofactor	Consider cofactor as part of the protein
--solvent	Consider solvent as part of the protein
--hydrogen	consider hydrogens
--desc	Write a descriptor file name descriptor.txt
--svm	Build a svm property file
-drog N	Observed druggability
--pharm	Generate a pharmacophore (.chm) from cavity
--outExclu	Output exclusion sphere in pharmacophore file

## PDB Process

**pdbconv protein[.pdb].mol2] output\_dir pdb\_id**

--wMOL2	Use MOL2 File as Input. PDB Options are not available
--wUnDrug	Output undruggable cavities
--noLig	PDB with no Ligand

By default all the following options are included. All chains will be kept

### [PDB Options]

--HARMSIZE	Harmonize size line to 80 characters
--MSEMET	Change MSE to MET
--CSECYS	Change CSE to CYS
--MOVHET	move HETATM to the end of file
--ALTATM	select alternative atoms
--NUMATM	renumerotate atoms
--UPDMAS	update the MASTER line
--TOMOL2	convert to a molecular representation (instead of flat file)

if you use one of the option below, you MUST use also --TOMOL2 option or use --wMOL2 option

### [MOL2 Options]

--RESTYP	apply Residue Class (cofactor/STD_AA/MOD_AA/Ligand ...)
--BONDSE	create bonds
--CLNUNW	clean unwanted residues
--MOL2TY	apply MOL2 types according to templates



Z5&amp;24x3/V



## IChem Tutorial

### 1. License check and generation

The **IChem license** command enables you to see details of your license

---

#### > **IChem license**

---

The terminal output should look like this ...

```
Licence key generator : YES
Molecular realignment : YES
  BSA Calculation : YES
    IFP generator : YES
      Interaction detection : YES
        Graph Interaction Matching : YES
          VolSite : YES
            PDB to MOL2 conversion : YES
              Patching MOL2 conversion : YES
                Utils : YES
                  Sims : YES
                    Scoring : YES
                      Detection, analysis of PPI : YES
Your licence expires on (Y/M/D): 2018/12/31
```

For each module, a YES/NO flag indicate whether you have rights to use the corresponding utility. The last line gives the expiration date of your current license.

If your license scheme allows you to generate license keys, the **IChem** *genkey* command permits you to define any possible license file

**genKey** *year month day allowed\_tools*

year : Expiring year (4 digits : 2014)  
 month: Expiring month (2 digits : 09)  
 day : Expiring day (2 digits : 23)

allowed\_tools: each value is separated by space and must have a value of either 1 or 0. Must follow this order

First : Licence Key generator  
 Second : Realigning molecules  
 Third : BSA Calculation  
 Fourth : IFP generator  
 Fifth : Interaction detection  
 Sixth : Graph Interaction Matching  
 Seventh: VolSite  
 Eighth : PDB to MOL2 conversion  
 Ninth : Patching MOL2 conversion  
 Tenth : Utils  
 Eleventh: sims  
 Twelfth: scoring  
 Thirteenth: Detection, analysis of PPI

---

> **IChem** *genKey* **2016 1 28 0 0 1 0 0 0 0 1 0 1 1**

---

grants you access until Jan.28<sup>th</sup> 2016 to 5 modules: BSA calculation, Patching MOL2 conversion, sims and scoring.

## 2. Protein-Ligand Interaction Fingerprints (IFPs)

The **IChem IFP** command registers in a bit string the interactions between a protein (active site) and a ligand.

**IFP protein ligand**

**IFP protein ligand ligand\_ref**

[General options]

-name N (LIG) Name of the fingerprint -Default: Name of the ligand  
 --polar Detect and output only polar interactions  
 --metal Detect and output only metal interactions  
 --extended Include within the fingerprint:  
     |--> Metal/Acceptor interaction  
     |--> Weak Hydrogen bonds  
     |--> PI-Cation interactions

[testing options]

-d\_Hb N (3.5) Hbond length (Angstroem)  
 -d\_Hyd N (4.5) Hydrophobic length (Angstroem)  
 -d\_Io N (4.0) Ionic length (Angstroem)  
 -d\_Me N (2.8) Metal/Acceptor length (Angstroem)  
 -d\_Ar N (4.0) Aromatic interaction length (Angstroem)  
 -a\_H N (Pi) HBond angle (rad.)  
 -at\_H N (Pi/3) HBond tolerance angle (rad.)  
 -a\_ArFF N (Pi) Aromatic Face to Face interaction angle (rad.)  
 -at\_ArFF N (Pi/6) Aromatic Face to Face tolerance angle (rad.)  
 -a\_ArEF N (Pi/2) Aromatic Edge to Face interaction angle (rad.)  
 -at\_ArEF N (Pi/3) Aromatic Edge to Face tolerance angle (rad.)  
 --ligD print all ligand possible interactions

Please note that ligand file can be a multi-mol2 file

For more information, see:



Marcou G, Rognan D.

J Chem Inf Model. 2007 Jan-Feb;47(1):195-207.

[Optimizing fragment and scaffold docking by use of molecular interaction fingerprints.](#)

In a first mode, interactions between an active site (mol2 file format) and one/several ligands (single or multimol2 file) are outputted in a table along with the IFP bit string. In a second mode, similarity of the IFP(s) to that of one (several) references is outputted in addition.

**-1st mode: computes an IFP between a protein active site and a ligand**

**Input directory:** \$ICHEM\_DIR/test/IFP

**Input files:** site.mol2, ligand.mol2

**Output file:** ligand.ifp

> IChem *IFP* *site.mol2* *ligand.mol2* > *ligand.ifp*

The following information is contained in the output file (e.g. ligand.ifp) or at the terminal if no file redirection (... >...) is given:

HBond LIG	OD1	78	ASP 113-A	O17	1	CAU	0-XX	2.60795	134.67
HBond PROT	ND2	421	ASN 312-A	O17	1	CAU	0-XX	2.76961	159.236
Hydrophobic	CZ3	327	TRP 286-A	C16	2	CAU	0-XX	4.05558	/
Hydrophobic	CZ	349	PHE 289-A	C16	2	CAU	0-XX	3.74597	/
Ionic LIG	OD2	79	ASP 113-A	N19	4	CAU	0-XX	2.94042	/
Ionic LIG	OD1	78	ASP 113-A	N19	4	CAU	0-XX	3.60381	/
Hydrophobic	CZ3	46	TRP 109-A	C21	6	CAU	0-XX	4.43435	/
Hydrophobic	CG2	64	THR 110-A	C21	6	CAU	0-XX	4.42961	/
Hydrophobic	CB	163	PHE 193-A	C21	6	CAU	0-XX	4.07199	/
Hydrophobic	CH2	47	TRP 109-A	C22	7	CAU	0-XX	3.72321	/
Hydrophobic	CZ	438	TYR 316-A	C22	7	CAU	0-XX	4.46257	/
Hydrophobic	CG2	125	VAL 117-A	C15	8	CAU	0-XX	4.0788	/
Hydrophobic	CZ3	327	TRP 286-A	C15	8	CAU	0-XX	4.06159	/
Hydrophobic	CE2	348	PHE 289-A	C15	8	CAU	0-XX	4.0318	/
Hydrophobic	CG2	90	VAL 114-A	C13	10	CAU	0-XX	3.81147	/
Hydrophobic	CG2	90	VAL 114-A	C12	11	CAU	0-XX	3.95913	/
Hydrophobic	CG1	124	VAL 117-A	C12	11	CAU	0-XX	3.98025	/
Hydrophobic	CG1	89	VAL 114-A	C11	12	CAU	0-XX	3.83646	/
Hydrophobic	CG1	124	VAL 117-A	C11	12	CAU	0-XX	4.29915	/
Hydrophobic	CB	288	SER 207-A	C11	12	CAU	0-XX	4.10275	/
Hydrophobic	CG1	89	VAL 114-A	C10	13	CAU	0-XX	4.0407	/
Hydrophobic	CB	288	SER 207-A	C10	13	CAU	0-XX	3.68825	/
Hydrophobic	CG2	90	VAL 114-A	C9	15	CAU	0-XX	4.09865	/
HBond LIG	OG	248	SER 203-A	N7	16	CAU	0-XX	3.31671	127.779
Hydrophobic	CG2	185	THR 195-A	C1	20	CAU	0-XX	4.32061	/
Hydrophobic	CE2	168	PHE 193-A	C6	21	CAU	0-XX	3.49585	/
Hydrophobic	CE2	168	PHE 193-A	C6	21	CAU	0-XX	3.49585	/

```

|A M82|A V86|A W109|A T110|A D113|A V114|A L115|A V117|A T118|A C191|A F193|A T195
|A Y199|A A200|A I201|A S203|A S204|A I205|A S207|A F208|A W286|A F289|A F290|A N293
|A Y308|A N312|A Y316
000000000000001000000100000000010110000000000001000000000000000010000001000000
0000000000000000000000000000000100000000000000010000001000000000000000000000
0000000000010001000000

```

In the first section, all intermolecular interactions between active site and ligand are tabulated

- 1<sup>st</sup> column: type of interaction
- 2<sup>nd</sup> column: interacting protein atom name
- 3<sup>rd</sup> column: interacting protein atom number
- 4<sup>th</sup> column: interacting protein residue name & number
- 5<sup>th</sup> column: interacting ligand atom name
- 6<sup>th</sup> column: interacting ligand atom number
- 7<sup>th</sup> column: interacting ligand residue name & number
- 8<sup>th</sup> column: interaction distance
- 9<sup>th</sup> column: interaction angle (for H-bonds only)

In the second section, the interaction fingerprint is displayed with 7 bits/residue in normal mode.

The first line describes the chain number and residue number

The second line is the bit string (1: interaction, 0: no interaction). The order is the following:

- 1<sup>st</sup> position: hydrophobic
- 2<sup>nd</sup> position: aromatic (face-to-face)
- 3<sup>rd</sup> position: aromatic (edge-to-face)
- 4<sup>th</sup> position: h-bond (protein donor)
- 5<sup>th</sup> position: h-bond (ligand donor)
- 6<sup>th</sup> position: ionic (protein charged +)
- 7<sup>th</sup> position: ionic (ligand charged +)

**-2nd mode: compute the IFP similarity (TC coefficient) between docked ligand and reference**

```
> IChem IFP site.mol2 docked.mol2 ligand.mol2 >docked Tc.ifp
```

--newH Less permissive definitions of hydrogen bonds

Interactions are detected based on default topological rules that can be modified whenever necessary:

---

> **IChem** -d\_Hb DHB -d\_Hyd DHYD -d\_Io DIO -d\_Me DME -d\_Ar DAR -a\_H AH -at\_H ATH -a\_ArFF AARFF -at\_ArFF ATARFF -a\_ArEF -AAREF -at\_AREF ATAREF -nomerge IFP  
arguments

---

Default values are described in the following table:

Parameter	Description	Default Value
DHB	Donor-acceptor distance threshold for H-bond interaction	3.5 Å
DHYD	Donor-acceptor distance threshold for hydrophobic interaction	4.5 Å
DIO	Anion-cation distance threshold for ionic interaction	4.0 Å
DME	Metal-acceptor distance threshold for metal chelation	2.8 Å
DAR	Distance threshold between two aromatic ring centers	4.0 Å
AH	Angle threshold between donor-H-acceptor for H-bond	$\pi$
ATH	Tolerance angle for H-bond	$\pi/3$
AARFF	Angle threshold between the plane of two aromatic cycles for face to face aromatic interactions	$\pi$
ATARFF	Tolerance angle for face to face aromatic interactions	$\pi/6$
AAREF	Angle threshold between the plane of two aromatic cycles for edge to face aromatic interactions	$\pi/2$
ATAREF	Tolerance angle for edge to face aromatic interactions	$\pi/6$
nomerge	After detecting all hydrophobic contacts, a filtering process is engaged to avoid too many hydrophobic pseudoatoms. Using this flag avoids the filtering process	no

The interaction rules options can be used for the following IChem tools: **IFP**, **ints**, **grim**



**Some notes:**

- Beware: the mol2 file should comply with the standard mol2 file format from TRIPOS (<http://www.tripos.com/data/support/mol2.pdf>)
- If you do not have SYBYL, you can generate correct mol2 files with UCSF chimera (<http://www.cgl.ucsf.edu/chimera>)
- IFPs are active-site dependent (nbits/residue) and cannot be compared across active sites of different compositions (e.g. different number of residues).
- Please rather use the active site and not the protein as input, to restrict the size of the IFP and prevent generating bit strings with mostly "0" values.
- Compare your output with that given in \$ICHEM\_DIR/test/IFP/output



### 3. Interaction Fingerprint Triplet (TIFPs)

TIFPs are coordinate-frame invariant interaction fingerprints presenting the advantage to compare ligand binding to completely different active sites, whatever their composition.

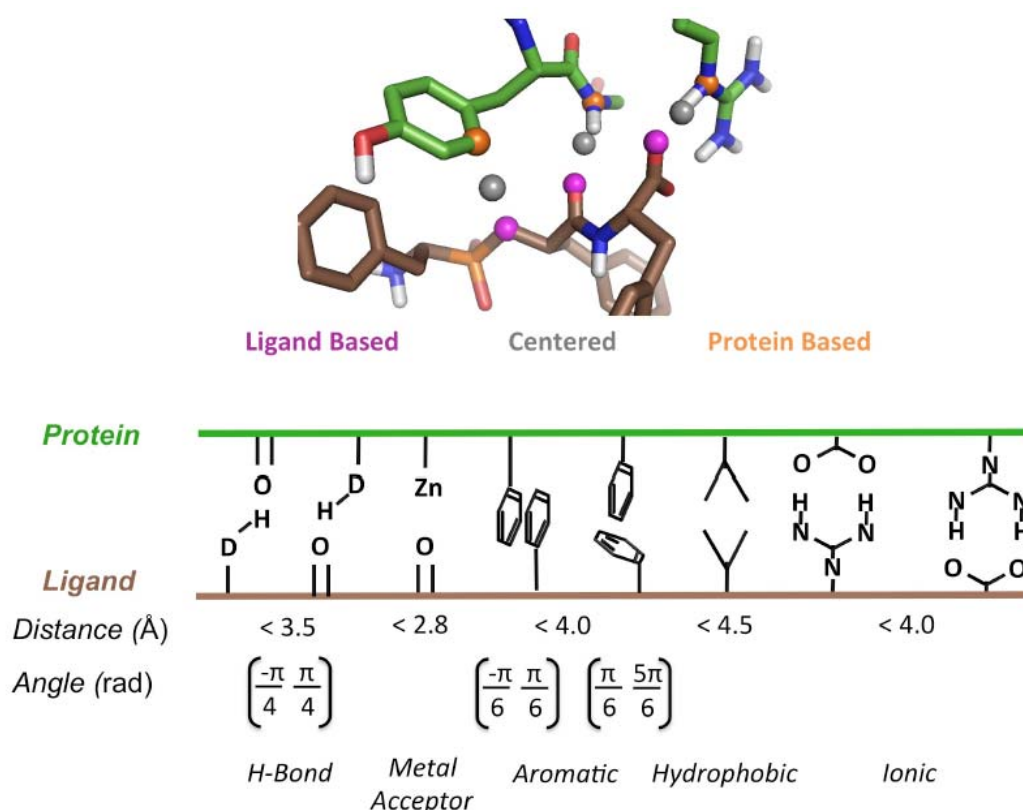
For more information, see:



Desaphy J, Raimbaud E, Ducrot P, Rognan D.  
J Chem Inf Model. 2013 Mar 25;53(3):623-37

[Encoding protein-ligand interaction patterns in fingerprints and graphs.](#)

Interactions are encoded by pseudoatoms in a MOL2 file format with the following properties: interaction type, atomic coordinates (**Figure 1**).



**Figure 1.** Encoding protein-ligand interaction by pseudoatoms. Each interaction is detected on the fly from standard topological rules and is represented by a pseudoatom with 3 possible atomic coordinates: the interacting ligand atom (LIG mode, violet balls), the interacting protein atom (PROT mode, orange balls), the barycenter of interacting protein and ligand atoms (CENT mode), the three aboved-described atoms (MERG mode).

**ints prot lig out**

-type (CENT)	Alter positionning output, multiple values are allowed, separated by space
PROT	InterPROT positionning
LIG	InterLIG positionning
CENT	Centered positionning
MERG	Merged all 3 above
-fgps (STD)	Fingerprint format
STD	Standard (1 0 21 0 0 3)
SVM	SVM format (1:1 3:21 6:3)
CMP	Compressed (1 [1 21 [2 3)
--small	Compressed fingerprint

[General options]

-name (prot)	Name of molecule in out file
-logf	Name of log file
-d_Hb N (3.5)	Hbond length (cut-off in Å)
-d_Hyd N (4.5)	Hydrophobic length (cut-off in Å)
-d_Io N (4.0)	Ionic length (cut-off in Å)
-d_Me N (2.8)	Metal/Acceptor length (cut-off in Å)
-d_Ar N (4.0)	Aromatic interaction length (cut-off in Å)
-a_H N (Pi)	HBond angle (cut-off in rad.)
-at_H N (Pi/3)	HBond tolerance angle (cut-off in rad.)
-a_ArFF N (Pi)	Aromatic Face to Face interaction angle (cut-off in rad.)
-at_ArFF N (Pi/6)	Aromatic Face to Face tolerance angle (cut-off in rad.)
-a_ArEF N (Pi/2)	Aromatic Edge to Face interaction angle (cut-off in rad.)
-at_ArEF N (Pi/3)	Aromatic Edge to Face tolerance angle (cut-off in rad.)
--noMerge	Do not merge hydrophobic interactions
--newH	Less permissive definitions of hydrogen bonds

**-Listing interactions and outputting interaction pseudoatoms**

Input directory: \$ICHEM\_DIR/test/TIFP

Input files: site.mol2, ligand.mol2

Output files: 2rh1\_ints.txt, 2rh1\_site\_INTS\_C.mol2

---

> **IChem** -logf 2rh1\_ints.txt -type CENT ints site.mol2 ligand.mol2

---

The 2rh1\_ints.txt file contains the same information as the table outputted by the **IChem IFP** command (recall pages 13, 14)

The 2rh1\_site\_INTS\_X.mol2 file describes the properties and atomic coordinates of the interaction pseudoatoms. Depending on the -type option (CENT, PROT, LIG, MERG), pseudoatoms are mapped onto ligand-interacting atoms (-type CENT), protein-interacting atoms (-type PROT), barycenter of ligand and protein-interacting atoms (-type CENT, default option), or on all above-described atoms (-type MERG)

Filename	-type mode
Header_INTS_C.mol2	CENT mode
Header_INTS_L.mol2	LIG mode
Header_INTS_P.mol2	PROT mode
Header_INTS_M.mol2	MERG mode

#### Atomic description of the interaction pseudoatoms:

Atom Name	Described interaction
CA	hydrophobic
CZ	aromatic
O	Hydrogen-bond (interacting protein atom is acceptor)
OG	Hydrogen-bond (interacting protein atom is both acceptor and donor)
N	Hydrogen-bond (interacting protein atom is donor)
OD1	Ionic (interacting protein atom is negatively charged)
NZ	Ionic (Interacting protein atom is positively charged)
ZN	Metal coordination

#### -Generating interaction fingerprints

With respect to *IChem IFP* , *IChem ints* outputs generic and binding site-independent interaction fingerprints consisting in all possible combinations of triplets of interaction pseudoatoms

Input directory: `$ICHEM_DIR/test/TIFP`

Input files: `site.mol2`, `ligand.mol2`

Output files: `2rh1.fgp`

---

> *IChem* `-fgps STD ints site.mol2 ligand.mol2 2rh1_full.fgp`

---

The fgp output file can appended after every novel *IChem ints* command if the same outputfile name is given. By default, a standard fingerprint format is outputted, but the `-fgps` option enables you to control the fingerprint format:

Format mode	<code>-fgps option</code>	Example	Description
Standard	STD	0 1 0 0 132 0 1 0 0 0	Each value is separated by a space Inefficient for long and sparse fingerprints (many null values)
SVM	SVM	2:1: 5:132 7:1	Only non-null values are outputted and defined by <i>POSITION:VALUE</i> separated by a space Very efficient for long sparse fingerprints Not efficient for long and dense fingerprints (few null values)
Compressed	CMP	[1 1 [2 132 [1 1 [3	Non-null values are explicitly outputted. Null values are encoded by a "[" sign followed by the number of consecutive null values Efficient for long and sparse fingerprints

By default, the standard fingerprint is made of 12510 integers, most of which encoding triplets of pseudoatoms describing hydrophobic contacts. The full fingerprint can be pruned to remove rare triplets (based on the analysis of ca. 10 000 protein-ligand PDB Structures) and lead to a simpler fingerprint of 210 integers (--small option)

---

> **IChem** --small -fgps STD ints site.mol2 ligand.mol2 2rh1\_small.fgp

---

#### 4. Graph Matching (GRIM)

GRIM is a tool to match protein-ligand complexes using a graph matching algorithm focusing on protein-ligand interaction pseudoatoms. It thereby enables an interaction-based alignment of different protein-ligand complexes that can be quantified by an empirical scoring function (GrScore) and used to post-process docking poses by similarity to known protein-ligand interaction patterns.

In a recent international contest (D3R Docking Challenge 2015), **Grim** was ranked **2<sup>nd</sup>** out of **44** scoring functions to predict the binding mode of 36 inhibitors prior to the release of protein-bound X-ray coordinates ([Gathiaka et al. J Comput.-Aided Mol. Des, 2016](#))

For more information, see:



Desaphy J, Raimbaud E, Ducrot P, Rognan D., J Chem Inf Model. 2013; 53: 623-637:  
[Encoding protein-ligand interaction patterns in fingerprints and graphs.](#)

Synko I, Da Silva F, Bret G, Rognan D. J Comput Aided Mol Des. 2016, 30: 669-683.  
[Docking pose selection by interaction pattern graph similarity: application to the D3R grand challenge 2015.](#)

<b>grim</b> refProt refLig CompProt CompLig	(1)
<b>grim</b> refInts complnts	(2)
<b>grim</b> refProt refFile dockFile	(3)

[Note]

- (1) use --multim2 to use multimol2
- (3) refFile & dockFile can be multimol2 files

[General options]

-rn	N (Ref)	Reference name
-cn	N (Comp)	Comparison name
--values		Only output score and not alignment
-sim	N (0)	Boolean telling whether the pair is similar or not
-outInt	(MERG)	Output only one kind of interaction positioning
	MERG	All aligned interactions are outputted
	LIG	InterLIG positioning
	CENT	Centered positioning
	PROT	InterPROT positioning
NOTE : outInt useless when used with --values		
-match	N (MERG)	Align only with a specific position
	MERG	Align with ALL interaction points (recommended)
	LIG	Align only with ligand interaction points
	PROT	Align only with protein interaction points
	CENT	Align only with centered interaction points
-max	N (1)	Maximal number of outputted cliques.
-size	N (3)	Minimal size of a clique.
--all_cliques		Detect all cliques and not only maximal one
-score	N (FCT)	Scoring method function
	STD	Scored by decreasing SumCl and increasing RMSD
	FCT	Scored with scoring function
--newH		Less permissive definition of hydrogen bonds

**-1<sup>st</sup> mode: Graph matching from structures**

**Input directory:** \$ICHEM\_DIR/test/GRIM

**Input files:** 2rh1\_prot.mol2, 2rh1\_lig.mol2, 4amj\_prot.mol2, 4amj\_lig.mol2

**Output files:** Complnts.mol2, Grifp\_res.csv, GRIM\_ints.mol2, GRIM\_lig.mol2, GRIM.log, GRIM\_prot.mol2, Reflnts.mol2

---

```
> IChem -sim 1 -rn 2rh1 -cn 4amj grim 2rh1_prot.mol2 2rh1_lig.mol2 4amj_prot.mol2 4amj_lig.mol2
```

---

**-2<sup>nd</sup> mode: Graph matching from interaction pseudoatoms**

**Input directory:** \$ICHEM\_DIR/test/GRIM

**Input files:** 2rh1\_INTS\_M.mol2, 4amj\_INTS\_M.mol2

**Output files:** Complnts.mol2, Grifp\_res.csv, GRIM\_ints.mol2, GRIM\_lig.mol2, GRIM.log, GRIM\_prot.mol2, Reflnts.mol2

---

```
> IChem -sim 1 -rn 2rh1 -cn 4amj grim 2rh1_INTS_M.mol2 4amj_INTS_M.mol2
```

---

Whatever the mode, 7 files are outputted:

**Complnts.mol2:** interaction pseudoatoms (merged mode) of the complex to fit

**Grifp\_res.csv:** summary of the GRIM alignment

**GRIM\_ints.mol2:** aligned interaction pseudoatoms of the complex to fit

**GRIM\_lig.mol2:** coordinates of the fitted ligand, aligned to the reference

**GRIM.log:** output of the GRIM alignment

**GRIM\_prot.mol2:** coordinates of the fitted protein, aligned to the reference

**Reflnts.mol2:** interaction pseudoatoms (merged mode) of the reference complex

The **Grifp\_res.csv** file (appended at every novel IChem grim command) is a table that looks like this:

NCli	Ref	Comp	Simil	LIG	CENTER	PROT	SumCl	RMSD	RLig	RCent	RProt	CLig	CCent	CProt	GrSc	RMSDAI	NPol
1	2rh1	4amj	1	9	6	7	0.0640	0.2763	14	20	20	15	21	21	0.7535	0.5589	10

NCli: Id of the clique (starts at 0)

Ref: Name of the reference (protein name if **-rn** is not used)

Comp: Name of the comparison (protein name if **-cn** is not used)

Simil: Similarity flag (empty is **-sim** is not used)

LIG: Number of matched LIG pseudoatoms

CENTER: Number of mgrimatched CENTERED pseudoatoms

PROT: Number of matched PROT pseudoatoms

SumCl: SumCl score (clique-based)

RMSD: root-mean square deviation (in Å) of the clique

RLig: Number of reference LIG pseudoatoms

RCent: Number of reference CENTERED pseudoatoms

RProt: Number of reference PROT pseudoatoms

CLig: Number of comparison LIG pseudoatoms

CCent: Number of comparison CENTERED pseudoatoms

CProt: Number of comparison PROT pseudoatoms

GrSC: Grim score (empirical score)

RMSDAI: root-mean square deviation (in Å) of the clique according to GrScore

NPol: Number of matched polar interaction pseudoatoms

### -3<sup>rd</sup> mode: Post-processing docking results by interaction graph matching

Input directory: \$ICHEM\_DIR/test/GRIM/screen

Input files: docked.mol2, ligand.mol2, site.mol2

Output files: Grimscreen\_res.tsv, GrScreen.csv

---

> IChem -sim 1 grim site.mol2 ligand.mol2 docked.mol2

---

The file docked.txt summarized the docking results obtained with Surflex-Dock: pose nr, docking score (pkd), rmsd to the true X-ray pose

Pose	pkd	rmsd
ligand_000	8.63	1.14
ligand_001	8.35	1.08
ligand_002	8.12	1.00
ligand_003	8.07	1.14
ligand_004	7.54	1.19
ligand_005	7.10	1.28
ligand_006	7.04	1.41
ligand_007	6.90	1.08
ligand_008	6.72	1.17
ligand_009	6.67	1.34
ligand_010	6.46	1.23
ligand_011	6.34	1.47
ligand_012	6.31	2.02
ligand_013	6.30	3.67
ligand_014	6.28	1.50
ligand_015	6.28	1.27
ligand_016	6.20	2.09
ligand_017	6.16	1.18
ligand_018	6.07	3.07
ligand_019	6.04	3.54

All docking poses (docked.mol2) are matched to the X-ray pose (ligand.mol2) for rescoring based on the interaction pattern graph similarity score (GrScore). The full output is stored in the Grimscreen\_res.tsv output file.

NCl	Ref	Comp	Simil	LIG	CENTER	PROT	SumCl	RMSD	RLig	RCent	RProt	CLig	CCent	CProt	GrSc	RMSDAL	NPol
1	2rhl_site	REF	1	11	13	11	0.0575	0.2035	15	26	20	16	27	18	0.8721	0.0051	8
1	2rhl_site	REF	1	11	6	12	0.0413	0.1863	15	26	20	17	28	18	0.8570	0.0030	5
1	2rhl_site	REF	1	13	12	11	0.0485	0.1851	15	26	20	18	31	22	0.8979	0.0034	7
1	2rhl_site	REF	1	12	13	13	0.0591	0.2229	15	26	20	18	30	21	0.9148	0.0033	10
1	2rhl_site	REF	1	12	13	12	0.0556	0.1612	15	26	20	17	27	17	0.9042	0.0032	7
1	2rhl_site	REF	1	8	12	10	0.0413	0.2225	15	26	20	14	28	16	0.8038	0.0067	4
1	2rhl_site	REF	1	10	4	3	0.0225	0.1845	15	26	20	15	30	17	0.7015	0.0117	2
1	2rhl_site	REF	1	9	15	16	0.0775	0.2100	15	26	20	15	24	18	0.9250	0.0044	10
1	2rhl_site	REF	1	13	13	13	0.0611	0.3393	15	26	20	18	34	23	0.9219	0.0060	12
1	2rhl_site	REF	1	4	6	8	0.0181	0.3472	15	26	20	17	29	19	0.6776	0.0048	0
1	2rhl_site	REF	1	7	9	7	0.0369	0.3185	15	26	20	15	30	18	0.7263	0.0084	6
1	2rhl_site	REF	1	9	5	2	0.0252	0.3542	15	26	20	16	26	16	0.6644	0.0103	3
1	2rhl_site	REF	1	4	2	1	0.0085	0.2867	15	26	20	14	25	15	0.5642	0.1907	0
1	2rhl_site	REF	1	5	4	4	0.0186	0.7238	15	26	20	15	28	17	0.6007	0.0242	2
1	2rhl_site	REF	1	3	5	9	0.0201	0.2705	15	26	20	17	29	16	0.6789	0.0025	1
1	2rhl_site	REF	1	3	9	12	0.0383	0.2189	15	26	20	18	24	17	0.7449	0.0013	4
1	2rhl_site	REF	1	3	2	4	0.0093	0.3322	15	26	20	17	29	19	0.5889	0.0811	0
1	2rhl_site	REF	1	9	12	14	0.0567	0.2337	15	26	20	17	26	18	0.8786	0.0034	7
1	2rhl_site	REF	1	2	3	3	0.0079	0.4525	15	26	20	16	31	16	0.5545	0.1352	0
1	2rhl_site	REF	1	3	7	7	0.0196	0.4476	15	26	20	17	29	19	0.6452	0.0105	1

The pose with the highest Grimscore is not the top-ranked one (ligand\_000) but pose 007 (highlighted in red). Interestingly, it is the pose with the second smallest rmsd to the true X-ray pose.

## - GRIM-based alignment of protein-ligand complexes

*realign rigidM mobilM applied1 applied2*

||-> rigidM : reference molecule to apply alignment to  
 ||-> mobilM : comparison molecule to apply alignment from  
 ||-> applied: molecule to apply rotation/translation to

[General options]

-gmatch N (NAME) Use graph matching to align  
 NAME Atom Name matching  
 ATMN Atomic Name matching  
 MOL2 MOL2 Type matching  
 CALP CAlpha Atom matching (protein only)  
 --wMob Also outputs the aligned mobilM  
 -rule R

By default, the program will perform an atom by atom match, without taking care of what kind of atom it match. If you want to perform a match by regarding only some atoms, this index\_string is here to do so

ex : -i '2-3|1-6|23-160' Will match the second atom from the reference with the third from the comparison, the first with the sixth ...

Input directory: **\$ICHEM\_DIR/test/REALIGN**

Input files: **4amj\_INTS\_M.mol2, GRIM\_ints.mol2, 4amj\_prot.mol2 4amj\_lig.mol2**

Output files: **rot\_4amj\_INTS\_M.mol2, rot\_4amj\_lig.mol2, rot\_4amj\_prot.mol2**

---

```
> IChem --wMob realign GRIM_ints.mol2 4amj_INTS_M.mol2 4amj_lig.mol2
4amj_prot.mol2
```

---

**IChem *realign*** will apply to the native ligand and/or protein input files (4amj\_lig.mol2, 4amj\_prot.mol2) a set of rotation/translations (deduced from the previous step) to align them to the reference GRIM input files (2rh1\_lig.mol2, 2rh1\_prot.mol2). The move mimics the transformation of interaction pseudoatoms previously saved: **4amj\_INTS\_M.mol2 → GRIM\_ints.mol2**

The aligned coordinates of the complex to fit are in the rot\_xxx.mol2 files.

## 5. Cavity detection and druggability prediction (VolSite)

**IChem VolSite** is a structure-based tool to automatically detect cavities at the surface of a target protein, and predict their ligandability (structural druggability)

For more information, see:



Desaphy J, Azdimousa K, Kellenberger E, Rognan D.

J Chem Inf Model. 2012 Aug 27;52(8):2287-99.

[Comparison and druggability prediction of protein-ligand binding sites from pharmacophore-annotated cavity shapes.](#)

**IChem Volsite** can be run in two modes depending on whether coordinates of a bound ligand are given (ligand-restricted mode) or not (unrestricted mode).

```

volsite prot lig                (1)
volsite prot                   (2)
    
```

### [General options]

```

-step N (1.5)    Edge length of each box (Å)
-boxS N (20)     Edge length of the main box (Å)
-b N (55)        Minimal threshold for buriedness
-name N          PDB Name for output cavity names
-n N (5)         Minimal neighbours for buried cavity boxes
-nPTS N (35)     Minimal number of cubes to consider it a cavity
--dna            Consider DNA as part of the protein
--cofactor       Consider cofactor as part of the protein
--solvent        Consider solvent as part of the protein
--hydrogen       consider hydrogens
--desc           Write a descriptor file name descriptor.txt
--svm            Build a svm property file
-drog N          Observed druggability
--pharm          Generate a pharmacophore (.chm) from cavity
--outExclu       Output exclusion sphere in pharmacophore file
    
```

### -Detection of all possible cavities (unrestricted mode)

Input directory: **\$ICHEM\_DIR/test/Volsite/unrestricted**

Input files: **protein.mol2**

Output files: **CAVITY\_Nx\_ALL.mol2** (x=1-21), **VolSite\_stat.csv**

---

> **IChem Volsite protein.mol2**

---

VolSite produces a summary (**VolSite\_stat.csv**) and a mol2 file (**CAVITY\_Nx\_ALL.mol2; x=1-21**) for each detected cavity. For the test example (**protein.mol2**), 21 cavities are detected at the surface of the protein and numbered (1 to 21) from the largest to the smallest.



The **VolSite\_stat.csv** summary recapitulates the properties of all detected cavities as follows:

Name	NCav	Size	NPts	Volume	CA	CZ	O	OG	OD1	N	NZ	DU	Drugg
/	1	ALL	326	1100.25	92	76	15	27	3	55	33	25	1.18203
/	2	ALL	134	452.25	57	10	3	13	9	10	9	23	0.693659
/	3	ALL	90	303.75	68	8	0	0	0	2	0	12	1.28227
/	4	ALL	74	249.75	30	11	5	4	1	8	0	15	0.329918
/	5	ALL	71	239.625	40	0	0	1	0	11	7	12	0.194943
/	6	ALL	69	232.875	14	4	3	0	14	7	19	8	-2.00612
/	7	ALL	64	216	16	3	0	4	0	16	19	6	-0.957034
/	8	ALL	55	185.625	27	8	1	3	0	7	0	9	0.610959
/	9	ALL	52	175.5	18	4	3	0	0	13	3	11	-0.279857
/	10	ALL	50	168.75	16	0	4	0	11	13	0	6	-1.34395
/	11	ALL	50	168.75	28	5	5	0	0	3	0	9	-0.0135895
/	12	ALL	49	165.375	36	3	1	0	0	1	1	7	0.501487
/	13	ALL	47	158.625	6	3	2	9	1	19	6	1	-1.00259
/	14	ALL	42	141.75	22	9	0	0	0	7	0	4	0.249771
/	15	ALL	41	138.375	25	6	0	0	0	7	0	3	0.162486
/	16	ALL	40	135	15	0	7	0	1	11	0	6	-1.01662
/	17	ALL	40	135	21	6	2	1	0	4	0	6	0.220955
/	18	ALL	39	131.625	17	2	0	2	7	4	0	7	-0.914936
/	19	ALL	38	128.25	19	1	2	5	0	4	0	7	0.078545
/	20	ALL	35	118.125	5	6	5	2	0	13	0	4	-0.854578
/	21	ALL	31	104.625	16	7	1	1	3	3	0	0	-0.119008
END													

1<sup>st</sup> column: name of the protein (by default, mol2 file protein header)

2<sup>nd</sup> column: Cavity number

3<sup>rd</sup> column: Size (ALL: no distance cut-off to the cavity center is applied to define cavity points)

4<sup>th</sup> column: Number of cavity points

5<sup>th</sup> column: Cavity volume in Å<sup>3</sup>

6<sup>th</sup> column: number of hydrophobic cavity points (CA)

7<sup>th</sup> column: number of aromatic cavity points (CZ)

8<sup>th</sup> column: number of hydrophogen-bond acceptor cavity points (O)

9<sup>th</sup> column: number of hydrogen-bond acceptor/donor cavity points (OG)

10<sup>th</sup> column: number of negatively ionizable cavity points (OD1)

11<sup>th</sup> column: number of hydrogen bond donor cavity points (N)

12<sup>th</sup> column: number of positively ionized cavity points (NZ)

13<sup>th</sup> column: number of dummy cavity points (DU): no protein atom < 4.5 Å

14<sup>th</sup> column: estimated druggability (Drugg): druggable if Drugg > 0; undruggable if Drugg < 0

#### -Detection of a cavity around a particular ligand (ligand-restricted mode)

Input directory: \$ICHEM\_DIR/test/Volsite/ligand

Input files: protein.mol2, ligand.mol2

Output files: CAVITY\_Nx\_y.mol2 (x = 1-3; y = 4, 6, 8, 12, ALL), VolSite\_stat.csv

---

> IChem **volsite** protein.mol2 ligand.mol2

---

VolSite produces a summary (**VolSite\_stat.csv**) and a mol2 file (**CAVITY\_Nx\_y.mol2**; x = 1-3; y = 4, 6, 8, 12, ALL) for each detected cavity. For the test example (**protein.mol2**, **ligand.mol2**), 3 cavities are detected at the surface of the protein and numbered (x= 1 to 3) from the largest to the smallest. y indicates the maximal distance (in Å) between cavity points and any input ligand heavy atom. For each cavity, 5 truncation modes are then applied to define binding sites of increasing size (4, 6, 8, 10, 12 Å) around the input ligand. If y = ALL, no truncation is applied.

The **VolSite\_stat.csv** summary recapitulates the properties of all detected cavities as follows:

Name	NCav	Size	NPts	Volume	CA	CZ	O	OG	OD1	N	NZ	DU	Recovery	Drugg
2rh1_CAU 1	1	4	122	411.75	30	35	3	17	0	23	8	6	45.082	/
2rh1_CAU 1	1	6	174	587.25	49	46	5	23	0	32	10	9	31.6092	/
2rh1_CAU 1	1	8	219	739.125	63	58	8	24	0	38	14	14	25.1142	/
2rh1_CAU 1	1	12	310	1046.25	90	74	12	27	3	52	28	24	17.7419	/
2rh1_CAU 1	1	ALL	326	1100.25	92	76	15	27	3	55	33	25	16.8712	1.18203
2rh1_CAU 2	2	4	2	6.75	0	0	0	1	0	1	0	0	0	/
2rh1_CAU 2	2	6	6	20.25	1	0	0	3	0	2	0	0	0	/
2rh1_CAU 2	2	8	12	40.5	2	0	0	6	0	4	0	0	0	/
2rh1_CAU 2	2	12	32	108	3	3	2	9	0	13	1	1	0	/
2rh1_CAU 2	2	ALL	47	158.625	6	3	2	9	1	19	6	1	0	1.00259
2rh1_CAU 3	3	4	0	0	0	0	0	0	0	0	0	0	-nan	/
2rh1_CAU 3	3	6	2	6.75	0	0	0	2	0	0	0	0	0	/
2rh1_CAU 3	3	8	13	43.875	7	1	1	3	0	1	0	0	0	/
2rh1_CAU 3	3	12	54	182.25	27	8	1	3	0	6	0	9	0	/
2rh1_CAU 3	3	ALL	55	185.625	27	8	1	3	0	7	0	9	0	0.61095

- 1<sup>st</sup> column: name of the ligand (by default, mol2 file ligand header)
- 2<sup>nd</sup> column: Cavity number
- 3<sup>rd</sup> column: Size (truncation mode)
- 4<sup>th</sup> column: Number of cavity points
- 5<sup>th</sup> column: Cavity volume in Å<sup>3</sup>
- 6<sup>th</sup> column: number of hydrophobic cavity points (CA)
- 7<sup>th</sup> column: number of aromatic cavity points (CZ)
- 8<sup>th</sup> column: number of hydrophogen-bond acceptor cavity points (O)
- 9<sup>th</sup> column: number of hydrogen-bond acceptor/donor cavity points (OG)
- 10<sup>th</sup> column: number of negatively ionizable cavity points (OD1)
- 11<sup>th</sup> column: number of hydrogen bond donor cavity points (N)
- 12<sup>th</sup> column: number of positively ionized cavity points (NZ)
- 13<sup>th</sup> column: number of dummy cavity points (DU): no protein atom < 4.5 Å
- 14<sup>th</sup> column: % of the binding site enclosing the ligand
- 15<sup>th</sup> column: estimated druggability (Drugg): druggable if Drugg > 0; undruggable if Drugg < 0

### - Outputting cavity properties

The **IChem --desc** command outputs cavity properties used by our SVM druggability predictor.

**Input directory:** \$ICHEM\_DIR/test/Volsite/unrestricted

**Input files:** protein.mol2

**Output files:** CAVITY\_Nx\_ALL.mol2 (x=1-21), VolSite\_stat.csv, 2rh1\_protein\_descriptor.txt

---

> **IChem --desc volsite protein.mol2**

---

The 2rh1\_protein\_descriptor.txt file outputs for every cavity a vector of 89 reals as follows:

#	Descriptor
1	Volume
2	% of aromatic cavity points (CZ)
3	% of hydrophobic points (CA)
4	% of h-bond acceptor points (O)
5	% of negatively ionizable points (OD1)

6	% of h-bond acceptor & donor points (OG)
7	% of h-bond donor points (N)
8	% of positively ionizable points (NZ)
9	% of dummy points (DU)
10	Percent of CZ points with a projection value below 40
11	Percent of CZ points with a projection value between 40 and 50
12	Percent of CZ points with a projection value between 50 and 60
13	Percent of CZ points with a projection value between 60 and 70
14	Percent of CZ points with a projection value between 70 and 80
15	Percent of CZ points with a projection value between 80 and 90
16	Percent of CZ points with a projection value between 90 and 100
17	Percent of CZ points with a projection value between 100 and 110
18	Percent of CZ points with a projection value between 110 and 120
19	Percent of CZ points with a projection value of 120
20	Percent of CA points with a projection value below 40
21	Percent of CA points with a projection value between 40 and 50
22	Percent of CA points with a projection value between 50 and 60
23	Percent of CA points with a projection value between 60 and 70
24	Percent of CA points with a projection value between 70 and 80
25	Percent of CA points with a projection value between 80 and 90
26	Percent of CA points with a projection value between 90 and 100
27	Percent of CA points with a projection value between 100 and 110
28	Percent of CA points with a projection value between 110 and 120
29	Percent of CA points with a projection value of 120
30	Percent of O points with a projection value below 40
31	Percent of O points with a projection value between 40 and 50
32	Percent of O points with a projection value between 50 and 60
33	Percent of O points with a projection value between 60 and 70
34	Percent of O points with a projection value between 70 and 80
35	Percent of O points with a projection value between 80 and 90
36	Percent of O points with a projection value between 90 and 100
37	Percent of O points with a projection value between 100 and 110
38	Percent of O points with a projection value between 110 and 120
39	Percent of O points with a projection value of 120
40	Percent of OD1 points with a projection value below 40
41	Percent of OD1 points with a projection value between 40 and 50
42	Percent of OD1 points with a projection value between 50 and 60
43	Percent of OD1 points with a projection value between 60 and 70
44	Percent of OD1 points with a projection value between 70 and 80
45	Percent of OD1 points with a projection value between 80 and 90
46	Percent of OD1 points with a projection value between 90 and 100
47	Percent of OD1 points with a projection value between 100 and 110
48	Percent of OD1 points with a projection value between 110 and 120
49	Percent of OD1 points with a projection value of 120
50	Percent of OG points with a projection value below 40
51	Percent of OG points with a projection value between 40 and 50
52	Percent of OG points with a projection value between 50 and 60
53	Percent of OG points with a projection value between 60 and 70
54	Percent of OG points with a projection value between 70 and 80
55	Percent of OG points with a projection value between 80 and 90
56	Percent of OG points with a projection value between 90 and 100

57	Percent of OG points with a projection value between 100 and 110
58	Percent of OG points with a projection value between 110 and 120
59	Percent of OG points with a projection value of 120
60	Percent of N points with a projection value below 40
61	Percent of N points with a projection value between 40 and 50
62	Percent of N points with a projection value between 50 and 60
63	Percent of N points with a projection value between 60 and 70
64	Percent of N points with a projection value between 70 and 80
65	Percent of N points with a projection value between 80 and 90
66	Percent of N points with a projection value between 90 and 100
67	Percent of N points with a projection value between 100 and 110
68	Percent of N points with a projection value between 110 and 120
69	Percent of N points with a projection value of 120
70	Percent of NZ points with a projection value below 40
71	Percent of NZ points with a projection value between 40 and 50
72	Percent of NZ points with a projection value between 50 and 60
73	Percent of NZ points with a projection value between 60 and 70
74	Percent of NZ points with a projection value between 70 and 80
75	Percent of NZ points with a projection value between 80 and 90
76	Percent of NZ points with a projection value between 90 and 100
77	Percent of NZ points with a projection value between 100 and 110
78	Percent of NZ points with a projection value between 110 and 120
79	Percent of NZ points with a projection value of 120
80	Percent of DU points with a projection value below 40
81	Percent of DU points with a projection value between 40 and 50
82	Percent of DU points with a projection value between 50 and 60
83	Percent of DU points with a projection value between 60 and 70
84	Percent of DU points with a projection value between 70 and 80
85	Percent of DU points with a projection value between 80 and 90
86	Percent of DU points with a projection value between 90 and 100
87	Percent of DU points with a projection value between 100 and 110
88	Percent of DU points with a projection value between 110 and 120
89	Percent of DU points with a projection value of 120

The projection value is the number of regularly-spaced 8 Å-long vectors emitted from each cavity points intercepting the protein surface (maximum of 120)

#### Correspondence between projection value and buriedness

Projection value	Buriedness, %
< 40	<33.3
40-50	33.3 - 41.6
50-60	41.6 – 50.0
60-70	50.0 - 58.3
70-80	58.3 - 66.6
80-90	66.6 – 75.0
90-100	75 .0 - 83.3
100-110	83.3 - 91.6
110-120	91.6 - 99.9
120	100

**Some notes:**

- The ligand may be a simple atom with user-defined coordinates!
- In ligand-restricted mode, druggability is only estimated for non-truncated full cavities (3<sup>rd</sup> column = ALL)
- The druggability model is only valid for standard VolSite parameters (step=1.5, boxes=20, b = 55, n =5)! If you change these parameters, keep in mind that the predicted druggability value has no meaning
- For large and accessible cavities (e.g. protein-protein interfaces), standard settings will output small-sized cavities. The most reliable approach to treat such flat cavities is to reduce the value of the *b* parameter (e.g. b= 45) until you get a set of cavity points that fills the expected interface area.
- If you want to consider accessory molecules (co-factor, nucleic acids, water) as part of your protein, use the corresponding options (--dna, --cofactor, --solvent, --hydrogen; see all arguments page 5)
- Using `-chm` argument enables the definition of a cavity-based pharmacophore in BIOVIA chm format. The additional `-OutExclu` parameter adds exclusion spheres on a few protein atoms close to cavity site points.

## 5. PDB processing (pdbconv)

**Ichem pdbconv** is a tool to parse and process PDB files, automatically detect bound ligands (HET code) and their cavity, and estimates their druggability. It is the protocol that we currently use to set-up the sc-PDB database of druggable protein-ligand complexes (<http://bioinfo-pharma.u-strasbg.fr/scPDB>). For more information on the parsing process, please see: <http://bioinfo-pharma.u-strasbg.fr/scPDB/ABOUT>

***pdbconv*** *protein[.pdb|.mol2]* *output\_dir* *pdb\_id*

--wMOL2            Use MOL2 File as Input. PDB Options are not available  
--wUnDrug        Output undruggable cavities  
--noLig          PDB with no Ligand

By default all the following options are included. All chains will be kept

[PDB Options]

--HARMSIZE       Harmonize size line to 80 characters  
--MSEMET        Change MSE to MET  
--CSECYS        Change CSE to CYS  
--MOVHET        move HETATM to the end of file  
--ALTATM        select alternative atoms  
--NUMATM        renumber atoms  
--UPDMAS        update the MASTER line  
--TOMOL2        convert to a molecular representation (instead of flat file)

if you use one of the option below, you MUST use also --TOMOL2 option or use --wMOL2 option

[MOL2 Options]

--RESTYP        apply Residue Class (cofactor/STD\_AA/MOD\_AA/Ligand ...)  
--BONDSE        create bonds  
--CLNUNW        clean unwanted residues  
--MOL2TY        apply MOL2 types according to templates  
--SPLITM        split molecule into protein/ligand/solvent  
--SelChain N    List of chains to keep, separated by underscore  
--SELWAT        select water molecules  
--SELLIG        select ligand

**Input directory:** **\$ICHEM\_DIR/test/PDBCONV**

**Input file:** **2RH1.pdb**

**Output directory:** **./output**

---

**> IChem *pdbconv* 2RH1.pdb output 2rh1**

---

For each detected ligand, a mol2 file is given along with the corresponding ligand-free protein (mol2 file), and Volsite cavities (mol2 file) in the **output** directory (to be created before executing the command). In addition, cavity descriptors and cavity properties are stored in the **\_descriptor.txt** and **VolSite\_Stat.csv** files, respectively.

## 6. Buried Surface Area calculation

The **IChem** *utils bsa* command computes the buried surface area of the ligand

*utils bsa protein ligand*

Input directory: \$ICHEM\_DIR/test/BSA

Input files: protein.mol2, ligand.mol2

Output file: ligand.bsa

---

> **IChem** *utils bsa protein.mol2 ligand.mol2* > *ligand.bsa*

---

The output is a 3 columns table:

2rh1_CAU_1_protein	71.4338	348.375
--------------------	---------	---------

1st column: protein-ligand header

2<sup>nd</sup> column: buried surface area (%)

3<sup>rd</sup> column: ligand volume (Å<sup>3</sup>)

## 7. Protein-bound ligand fragmentation

The **IChem** *utils frag* command fragments in 3-D space a protein-bound ligand structure using a method aimed at detecting substituted ring cores. First, a ring perception algorithm is used to automatically detect aromatic and aliphatic rings. Acyclic atoms are then parsed to assign either a linker or substituent label, as whether to the corresponding bonds are connecting two rings or not. Linker atoms are left unchanged. In case of substituent atoms, single bonds involving the closest apolar carbon (in terms of bond distance) to any ring are later cleaved at the condition that the cleaved bond is at least 3 bonds away from the cyclic root atom. An anchoring atom (Z label) is then added to each of the remaining fragments to indicate the cleavage site

For more information, see:



Desaphy J, Rognan D.

J Chem Inf Model. 2014 Jul 28;54(7):1908-18

[sc-PDB-Frag: a database of protein-ligand interaction patterns for Bioisosteric replacements.](#)

*utils frag protein ligand*

Input directory: \$ICHEM\_DIR/test/Frag

Input file: protein.mol2, ligand.mol2

Output files: 2rh1\_CAU\_1\_protein\_FRAG\_1\_MOLE.mol2, 2rh1\_CAU\_1\_protein\_FRAG\_1\_INTS.mol2, xx.mol2

---

> **IChem** *utils frag protein.mol2 ligand.mol2*

---

Three mol2 files are outputted:

**2rh1\_CAU\_1\_protein\_FRAG\_1\_MOLE.mol2**: fragment from the original ligand

**2rh1\_CAU\_1\_protein\_FRAG\_1\_INTS.mol2:** interaction pseudoatoms (merged mode) for the fragment

**xxx.mol2:** interaction pseudoatoms (merged mode) for the entire ligand

#### Some notes:



- Protein and ligand 3D coordinate should be in the same coordinates frame in mol2 file format.
- Several fragments can be generated. Each fragment has an index (header\_FRAG\_index.mol2) along with the corresponding interaction pseudoatoms (header\_FRAG\_index\_INTS.mol2)

## 8. Fingerprint similarity measures

The **IChem sims** command enables you to compute similarities with various metrics between two fingerprints.

### Fingerprint Similarity

**sims ref comp** (1)

**sims file** (2)

**sims RefInt Complnt** (3)

#### [General options]

--wlnts	To use interactions instead of fingerprints (option 3 only)
--small	Use small fingerprint (option 3 only)
--binary	To add when the fingerprint is binary
-metric N (TC)	Select the metric
TC	Tanimoto metric (default)
HM	Hamming distance
RT	Ref Tversky
FT	Fit Tversky
DI	Dice
SO	Soergel

**Input directory:** \$ICHEM\_DIR/test/SIMILARITY

**Input file:** FP1.txt, FP2.txt

**Output files:** sim.txt

---

**> IChem sims FP1.txt FP2.txt >sim.txt**

---

Fingerprints of diverse nature are accepted

STD	Standard (1 0 21 0 0 3)
SVM	SVM format (1:1 3:21 6:3)
CMP	Compressed (1 [1 21 [2 3)

The output is a table listing the reference fingerprint, the target fingerprint and the similarity score:

FP1 FP2 0.428571



## 9. Detection and analysis of protein-protein interfaces (DetectPPI)

DetectPPI is a tool to detect and analyze protein-protein interfaces, and outputs cavities remote to every detected PPI

For more information, see:



Da Silva, F., Desaphy, J., Bret, G. and Rognan, D. (2015) [IChemPIC: A Random Forest Classifier of Biological and Crystallographic Protein-Protein Interfaces](#). J. Chem. Inf. Model, 55, 2005–2014.

**Input directory:** \$ICHEM\_DIR/test/PPI

**Input file:** 4NN6.pdb

**Output files:**

4NN6_interaction_A_B.ints	4NN6_prot_B_C.mol2	CAVITY_all_N2_ALL.mol2	CAVITY_B_N1_ALL.mol2
4NN6_interaction_A_C.ints	4NN6_site_A_B.mol2	CAVITY_all_N3_ALL.mol2	CAVITY_B_N2_ALL.mol2
4NN6_interaction_B_C.ints	4NN6_site_A_C.mol2	CAVITY_all_N4_ALL.mol2	CAVITY_B_N3_ALL.mol2
4NN6_ints_A_B.mol2	4NN6_site_B_C.mol2	CAVITY_all_N5_ALL.mol2	CAVITY_C_N1_ALL.mol2
4NN6_ints_A_C.mol2	CAVITY_all_N10_ALL.mol2	CAVITY_all_N6_ALL.mol2	descriptor.sre
4NN6_ints_B_C.mol2	CAVITY_all_N11_ALL.mol2	CAVITY_all_N7_ALL.mol2	VolSite_Stat.csv
4NN6_prot_A_B.mol2	CAVITY_all_N12_ALL.mol2	CAVITY_all_N8_ALL.mol2	
4NN6_prot_A_C.mol2	CAVITY_all_N1_ALL.mol2	CAVITY_all_N9_ALL.mol2	

---

> **IChem detectppi 4NN6 4NN6.pdb**

---

IChem requires a PDB file with orientated polar hydrogen atoms. You can use any program for that specific task although we strongly recommend the usage of ProToss (<http://protoss.zbh.uni-hamburg.de/>) for that purpose.

PDB\_interaction\_X\_Y.ints: interactions between protein chains X and Y

PDB\_ints\_X\_Y.mol2: interaction pseudoatoms (centered mode) between protein chains X and Y

PDB\_prot\_X\_Y.mol2: mol2file of protein chains X and Y

PDB\_site\_X\_Y.mol2: mol2file of interacting residues from chains X and Y

CAVITY\_all\_X\_ALL.mol2: Volsite cavities for the entire protein assembly

CAVITY\_X\_NY\_ALL.mol2: Volsite cavities in chain X only

descriptor.sre: set of 45 descriptors of each PPI

VolSite\_Stat.csv: Volsite output

Here is the meaning of the 45 PPI descriptors:

Name	Description
nPTS	Total number of interaction points
Hydro	% of hydrophobic interaction points
Aro	% of aromatic interaction points
Hbond	% of hydrogen-bond interaction points
Ionic	% of ionic bond interaction points
Hydro1	% of hydrophobic points (25 %<Burial <33.3%)
Hydro2	% of hydrophobic points (33.3 %<Burial <41.6%)
Hydro3	% of hydrophobic points (41.6 %<Burial <50%)
Hydro4	% of hydrophobic points (50 %<Burial <58.3%)
Hydro5	% of hydrophobic points (58.3 %<Burial <66.6%)
Hydro6	% of hydrophobic points (66.6 %<Burial <75%)

<b>Hydro7</b>	% of hydrophobic points 75 %<Burial <83.3%)
<b>Hydro8</b>	% of hydrophobic points (83.3 %<Burial <91.6%)
<b>Hydro9</b>	% of hydrophobic points (91.6%<Burial <100%)
<b>Hydro10</b>	% of hydrophobic points (Burial =100%)
<b>Aro1</b>	% of aromatic points (25 %<Burial <33.3%)
<b>Aro2</b>	% of aromatic points (33.3 %<Burial <41.6%)
<b>Aro3</b>	% of aromatic points (41.6 %<Burial <50%)
<b>Aro4</b>	% of aromatic points (50 %<Burial <58.3%)
<b>Aro5</b>	% of aromatic points (58.3 %<Burial <66.6%)
<b>Aro6</b>	% of aromatic points (66.6 %<Burial <75%)
<b>Aro7</b>	% of aromatic points 75 %<Burial <83.3%)
<b>Aro8</b>	% of aromatic points (83.3 %<Burial <91.6%)
<b>Aro9</b>	% of aromatic points (91.6%<Burial <100%)
<b>Aro10</b>	% of aromatic points (Burial =100%)
<b>Hbond1</b>	% of hydrogen bond points (25 %<Burial <33.3%)
<b>Hbond2</b>	% of hydrogen bond points (33.3 %<Burial <41.6%)
<b>Hbond3</b>	% of hydrogen bond points (41.6 %<Burial <50%)
<b>Hbond4</b>	% of hydrogen bond points (50 %<Burial <58.3%)
<b>Hbond5</b>	% of hydrogen bond points (58.3 %<Burial <66.6%)
<b>Hbond6</b>	% of hydrogen bond points (66.6 %<Burial <75%)
<b>Hbond7</b>	% of hydrogen bond points 75 %<Burial <83.3%)
<b>Hbond8</b>	% of hydrogen bond points (83.3 %<Burial <91.6%)
<b>Hbond9</b>	% of hydrogen bond points (91.6%<Burial <100%)
<b>Hbond10</b>	% of hydrogen bond points (Burial =100%)
<b>Ionic1</b>	% of ionic bond points (25 %<Burial <33.3%)
<b>Ionic2</b>	% of ionic bond points (33.3 %<Burial <41.6%)
<b>Ionic3</b>	% of ionic bond points (41.6 %<Burial <50%)
<b>Ionic4</b>	% of ionic bond points (50 %<Burial <58.3%)
<b>Ionic5</b>	% of ionic bond points (58.3 %<Burial <66.6%)
<b>Ionic6</b>	% of ionic bond points (66.6 %<Burial <75%)
<b>Ionic7</b>	% of ionic bond points 75 %<Burial <83.3%)
<b>Ionic8</b>	% of ionic bond points (83.3 %<Burial <91.6%)
<b>Ionic9</b>	% of ionic bond points (91.6%<Burial <100%)
<b>Ionic10</b>	% of ionic bond points (Burial = 100%)

Please note that we recommend the usage of our web interface (<http://bioinfo-pharma.u-strasbg.fr/IChemPIC>) to predict the relevance (biological, crystallographic) of any possible interface from a PDB file.