# IsoStar User Guide

# 1 IsoStar User Guide and Tutorials

2025.3 CSD Release

## 1.1 Conditions of Use

By using this software, you also agree to our standard licence agreement in the following link: https://www.ccdc.cam.ac.uk/licence-agreement/

The Cambridge Structural Database Portfolio (CSD Portfolio) including, but not limited to, the following: ConQuest, CSD-Editor, Decifer, Mercury, Mogul, IsoStar, CSD Conformer Generator, Hermes, GOLD, SuperStar, the CSD Python API, web accessible CSD tools and services, WebCSD, CSD sketchers, CSD data files, CSD data updates, the CSD database, sub-files derived from the foregoing data files, documentation and command procedures, test versions of any existing or new program, code, tool, data files, sub-files, documentation or command procedures which may be available from time to time (each individually a Component) encompasses database and copyright works belonging to the Cambridge Crystallographic Data Centre (CCDC) and its licensors and all rights are protected.

Any use of a Component of the CSD Portfolio, is permitted solely in accordance with a valid Licence of Access Agreement or Products Licence and Support Agreement and all Components included are proprietary. When a Component is supplied independently of the CSD Portfolio its use is subject to the conditions of the separate licence. All persons accessing the CSD Portfolio or its Components should make themselves aware of the conditions contained in the Licence of Access Agreement or Products Licence and Support Agreement or the relevant licence.

In particular:

The CSD Portfolio and its Components are licensed subject to a time limit for use by a specified organisation at a specified location.

Licences may be obtained from:

Cambridge Crystallographic Data Centre

12 Union Road

Cambridge CB2 1EZ, United Kingdom

Web: http://www.ccdc.cam.ac.uk

Telephone: +44-1223-336408

Email: admin@ccdc.cam.ac.uk

# 2 Introduction

## 2.1 Overview of IsoStar

IsoStar is a library of information about the intermolecular interactions formed by a wide variety of chemical groups. It is designed primarily for:

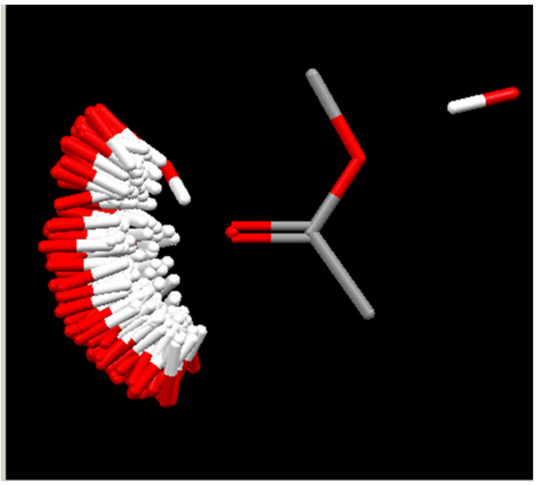
- Molecular modellers engaged in structure-based drug design.
- Medicinal chemists interested in identifying bioisosteric replacements.

- Protein crystallographers.
- Crystal engineers.

IsoStar contains data on:

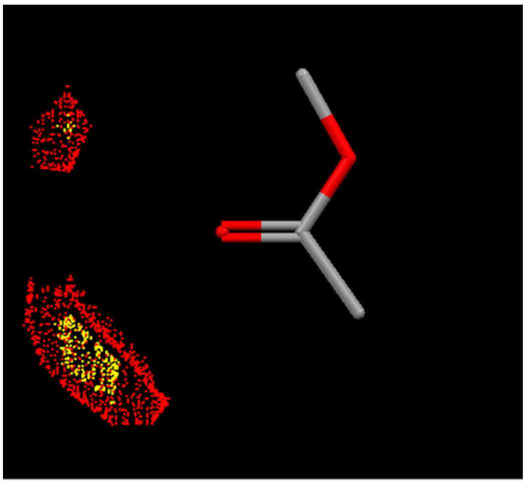
- Intermolecular interactions in small-molecule crystal structures taken from the Cambridge Structural Database (CSD).
- Protein-ligand interactions in X-ray structures from the Protein Data Bank (PDB).
- Theoretical intermolecular interaction energies calculated by the Intermolecular Perturbation Theory (IMPT) method.

Crystallographic information in IsoStar is presented as 3D scatterplots. Each scatterplot has been calculated by searching the CSD or PDB for nonbonded interactions between a pair of functional groups A and B. The A...B contacts are transformed so that the A groups are least-squared superimposed. The resulting scatterplot shows the experimentally observed distribution of B (the contact group) around A (the central group), e.g.

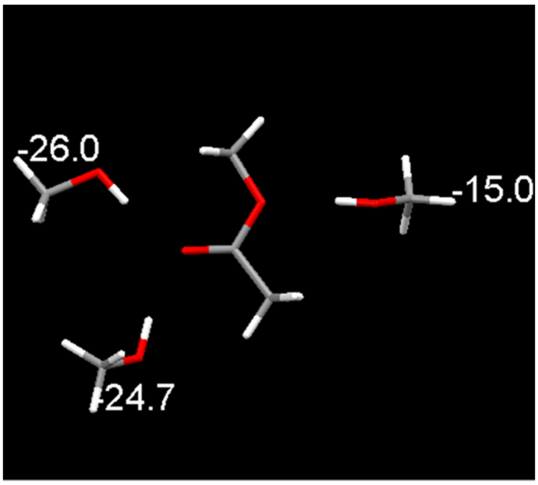


Such a scatterplot gives information about the frequencies and directionalities of intermolecular contacts. The above example shows that OH groups (the contact group in this case) strongly hydrogen-bond to the carbonyl oxygen of esters (the central group) but not to the other oxygen atom.

Each scatterplot can be converted into a contoured density surface, which is often useful for highlighting geometrical preferences, e.g.



The molecular-orbital data provide information about interaction energies and theoretical in vacuo minimum-energy geometries, e.g.



IsoStar also contains statistical data on the frequencies with which various interactions occur in crystal structures.

## 2.2 Overview of the IsoStar Interface

IsoStar uses a standard web-browser to allow selection of the interaction of interest from a pre-computed library of several hundred central groups (see Overview of IsoStar) and about fifty contact groups.

Information for each central group is presented in a table which provides links to scatterplots, tabulated MO results and statistical data.

Scatterplots can be displayed in a graphical interface, which allows:

- Customisable 3D visualisation of the scatterplot (see Using the 3D Visualiser).
- Control (by distance) over which contacts are displayed (see Displaying Contacts in a Particular Distance Range).
- Presentation of the scatterplot as a contoured density surface (see Displaying Contoured Density Surfaces).
- Hyperlinking to CSD or PDB entries that contribute to the scatterplot (see Hyperlinking from a Scatterplot to a Crystal Structure).
- Measurement of contact distances and angles (see Measuring Distances, Angles and Torsions).

Theoretical energy minima (see Theoretical Data) and crystal-statistics data (see Statistical Data) are tabulated in web-browser pages and the former can be displayed in the visualiser.

## 2.3 Calculating New Scatterplots

An ancillary program, IsoGen (see IsoGen: Quick Summary), allows you to calculate scatterplots for interactions that are not present in the standard IsoStar library.

# 3 Organisation and Navigation of IsoStar Data

## 3.1 Central Groups

Information in IsoStar is accessible via a web-browser interface and is organised on a per-group basis, e.g. there is a page that provides access to information on the nonbonded contacts formed by the group -$NO_2$, a page that gives analogous information for -CH=CH-, etc. In IsoStar, these groups are called central groups (see Overview of IsoStar).

There are over 300 central groups in the IsoStar library, ranging from single-atom substituents such as fluoro, common linking groups such as ester, ring systems such as isoxazole, amino acid side-chains, and common solvate molecules.

## 3.2 Contact Groups

For each central group (see Central Groups), information is held on its interactions with up to about fifty contact groups (see Overview of IsoStar).

The same set of contact groups is used for all central groups. They fall into the following categories:

- General groups, viz. any C,N,O,S or H atom; and any polar hydrogen (X-H, where X is N,O or S).
- Hydrophobic groups, e.g. methyl, phenyl.
- N-H hydrogen-bond donor groups, e.g. amide NH, ammonium NH.
- Various types of O-H groups, e.g. alcohol OH, water.
- Other groups containing N and/or O. Most of these are H-bond acceptor groups, e.g. cyano, carbonyl O, nitro.
- Various types of sulfur, e.g. thioether, thiocarbonyl.
- Groups containing halogen atoms, e.g. C-F, chloride.
- Groups found in amino acids, e.g. imidazole, guanidinium.

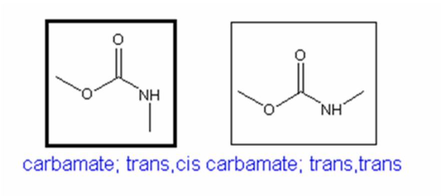
## 3.3 Ligand and Protein Classification of Central Groups

Central groups (see Central Groups) in IsoStar are divided into two main sections, Ligand and Protein. The distinction is important for the PDB part of the library and is best explained by an example. Consider the central group carbamoyl, $-CONH_2$. This appears in both the Ligand and Protein sections. However, the two sets of PDB scatterplots (see Overview of IsoStar) are not identical. This is because the plots in the Ligand section show contacts to carbamoyl groups on ligand molecules. Conversely, the plots in the Protein section show contacts to the carbamoyl groups of asparagine and glutamine side chains in proteins.

Strictly speaking, all CSD-based scatterplots should be under the Ligand section of IsoStar, since the CSD contains only small molecules, not proteins. However, CSD data are included in the Protein section as well, so that comparisons can easily be made between non-bonded contacts to protein functional groups and contacts to similar functional groups in the higher precision small-molecule CSD structures. In general, tighter chemical constraints have been used in selecting the CSD structures that contribute to the plots in the Protein section. For example, any CSD structure containing a C-$CONH_2$ group will contribute to the carbamoyl CSD-based scatterplots in the Ligand section. However, only structures containing the more precisely defined substructure -$CH_2$-$CONH_2$ will contribute to the Protein carbamoyl CSD plots.

## 3.4 Conformationally-Flexible Central Groups

When a central group (see Central Groups) adopts two or more clearly distinct conformations, each is included in IsoStar as a separate entry, e.g. carbamate trans, cis and carbamate trans, trans:

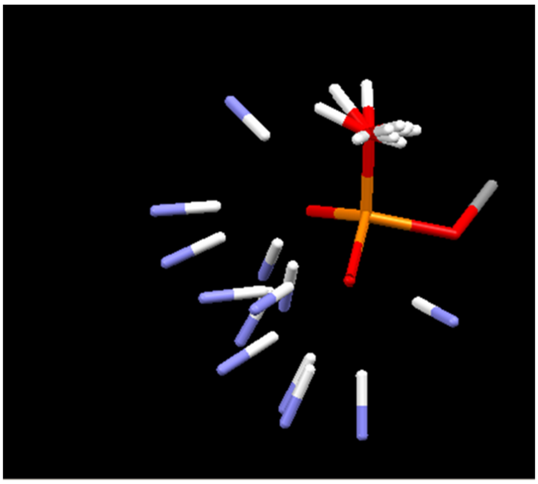


An exception may be made if one of the conformers is much more common than the other(s), in which case only the major conformer may be included.

Sometimes, the division between two conformers may not be distinct, i.e. some structures may lie half-way between two minimum-energy geometries. In this case, an arbitrary criterion is used. For example, anilines are considered to be planar if the sum of the three bond angles at nitrogen exceeds 352.5°, otherwise they are considered to be pyramidal. (The criterion of 352.5° is used throughout IsoStar to define planar and pyramidal nitrogen.)

Some central groups have ill-defined conformations. For example, the C-S-S-C torsion angle of acyclic disulfide linkages is usually about 85° but can be as low as 50° or as high as 120°. In such a case, an average geometry is calculated and individual structures are omitted if they do not overlay well enough on this average (specifically, all non-hydrogen atoms must fit to within 1 Å and the RMS deviation must be less than 0.5 Å).

In some cases, a central group adopts a well-defined conformation but contains a terminal group such as methyl or hydroxyl that shows a great deal of rotational flexibility. In this case, the H atoms may be omitted entirely or displayed as shown in the scatterplot below:

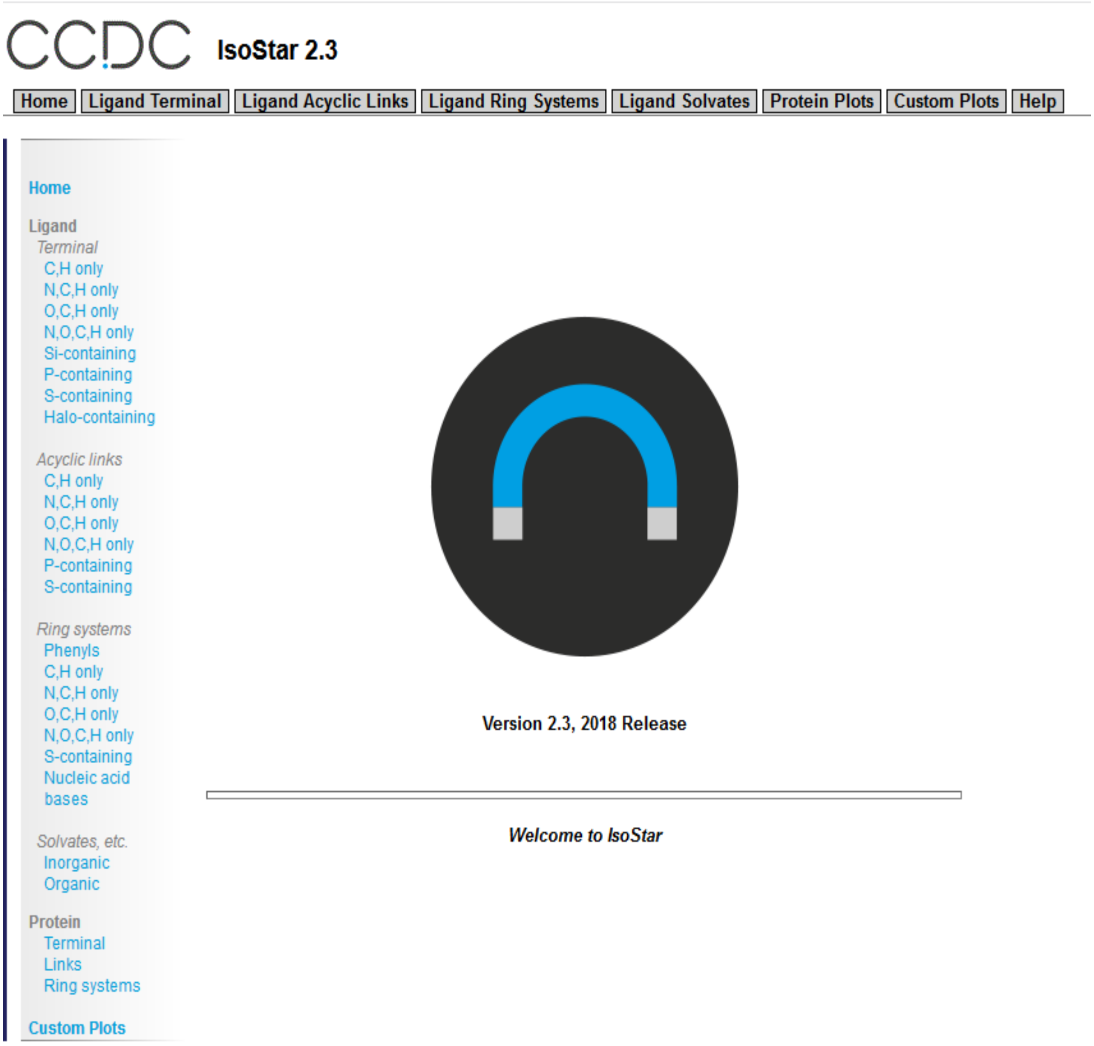


## 3.5 Central Groups with Ambiguous Protonation States

If a central group (see Central Groups) can exist in more than one conformer or more than one ionisation or tautomeric state, the various forms will be included separately, e.g. C-$OPO_3^{2-}$ and C-$OPO_3H^-$ are separate entries in the library.

## 3.6 Accessing Data for a Particular Central Group

To obtain information for a particular central group (see Central Groups), open the IsoStar home page (http://isostar.ccdc.cam.ac.uk/html/isostar.html) in a web browser. The available types of central groups are listed on the left-hand side of this page, divided firstly by whether the central group is in the Ligand or Protein section (see Ligand and Protein Classification of Central Groups); then by whether the group is a terminal substituent, linking group, ring system or solvate molecule; and then by the element(s) it contains:

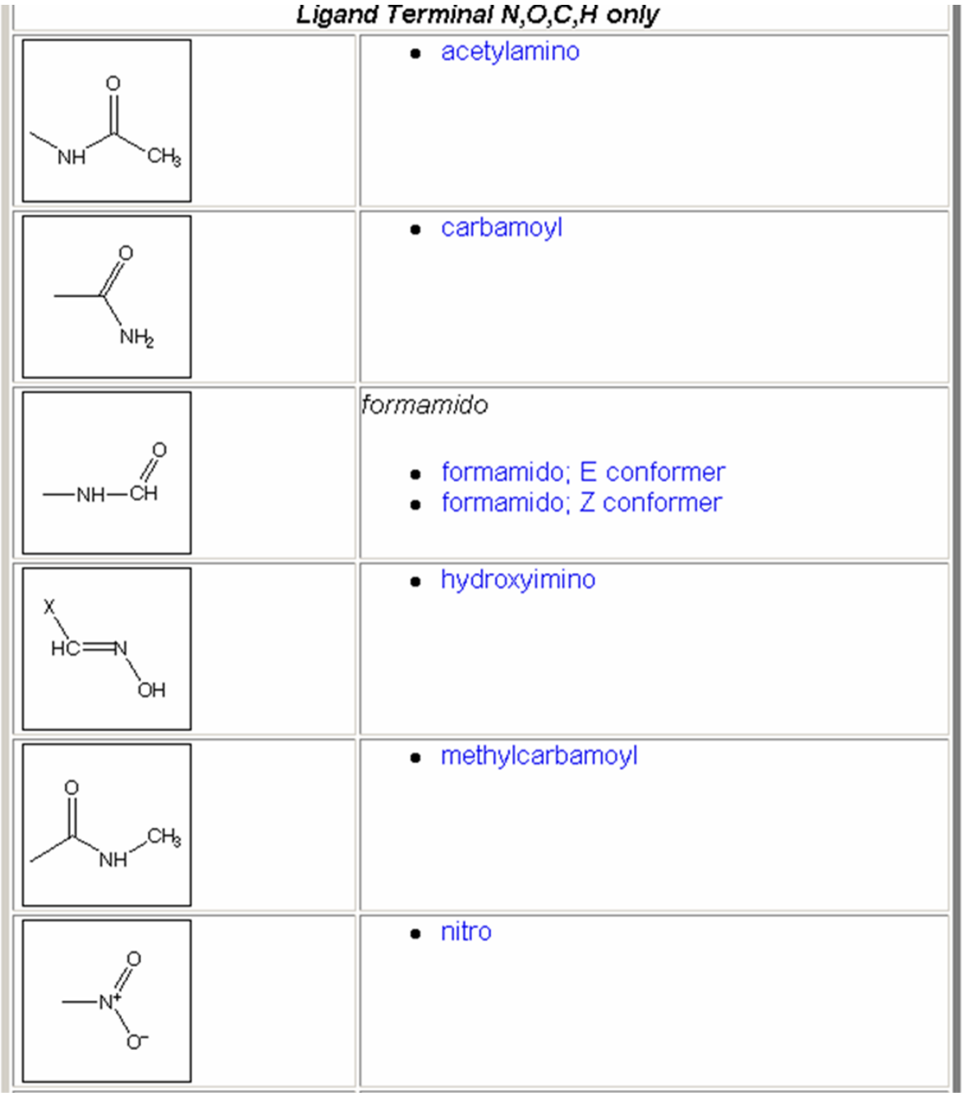


Version 2.3, 2018 Release

*Welcome to IsoStar*

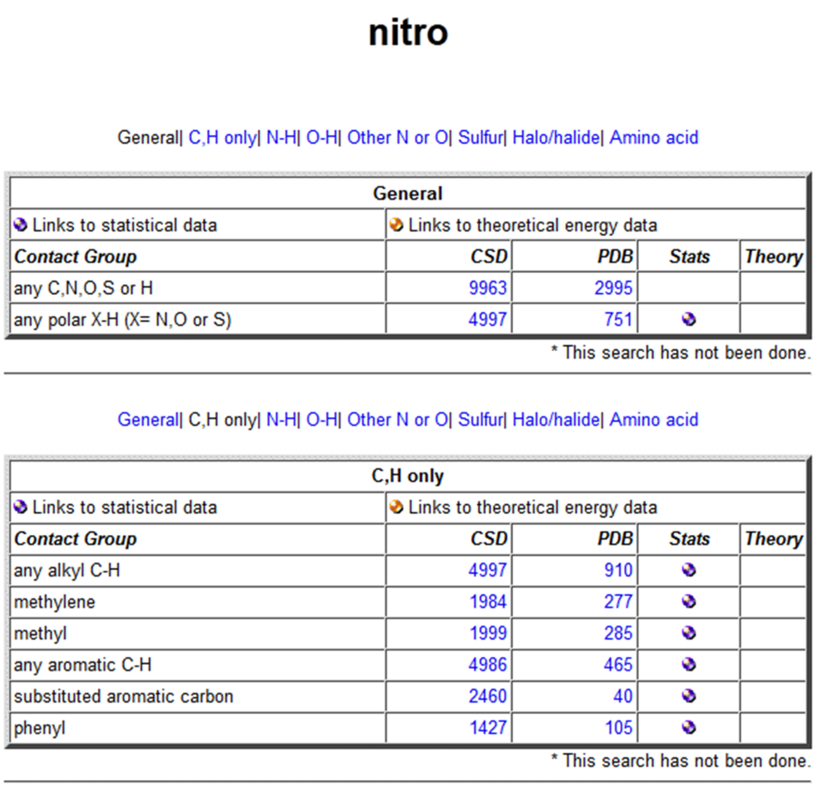
Although IsoStar contains data for many different functional groups, users may want to generate their own scatterplots for contacts not currently included in the library. Consequently, there is a directory reserved in IsoStar for Custom Plots, i.e. scatterplots created by users with the program IsoGen (see IsoGen: Quick Summary).

Clicking on one of the central-group categories produces a list of available groups, e.g.



| ***Ligand Terminal N,O,C,H only*** | |
|---|---|
| | - acetylamino |
| | - carbamoyl |
| | *formamido*<br>- formamido; E conformer<br>- formamido; Z conformer |
| | - hydroxyimino |
| | - methylcarbamoyl |
| | - nitro |

Clicking on one of the central-group names causes a table to be displayed which is the access point for all data pertaining to that group. The screenshot below shows part of this table for nitro:



nitro

General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

| General | | | | |
|---|---|---|---|---|
| ● Links to statistical data | ● Links to theoretical energy data | | | |
| *Contact Group* | *CSD* | *PDB* | *Stats* | *Theory* |
| any C,N,O,S or H | 9963 | 2995 | | |
| any polar X-H (X= N,O or S) | 4997 | 751 | ● | |

* This search has not been done.

General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

| C,H only | | | | |
|---|---|---|---|---|
| ● Links to statistical data | ● Links to theoretical energy data | | | |
| *Contact Group* | *CSD* | *PDB* | *Stats* | *Theory* |
| any alkyl C-H | 4997 | 910 | ● | |
| methylene | 1984 | 277 | ● | |
| methyl | 1999 | 285 | ● | |
| any aromatic C-H | 4986 | 465 | ● | |
| substituted aromatic carbon | 2460 | 40 | ● | |
| phenyl | 1427 | 105 | ● | |

* This search has not been done.

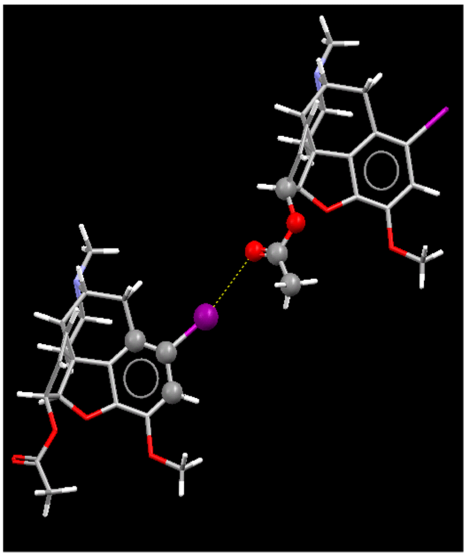
Clicking on links in this table allows you to:

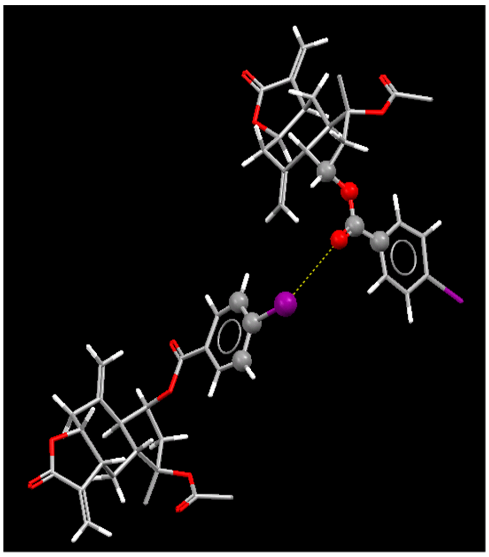
- View scatterplots based on CSD or PDB data (see Loading and Reloading Scatterplots).
- View statistical information on the frequencies of contacts (see Statistical Data).
- View theoretical results (see Theoretical Data).

# 4 Viewing and Using Scatterplots

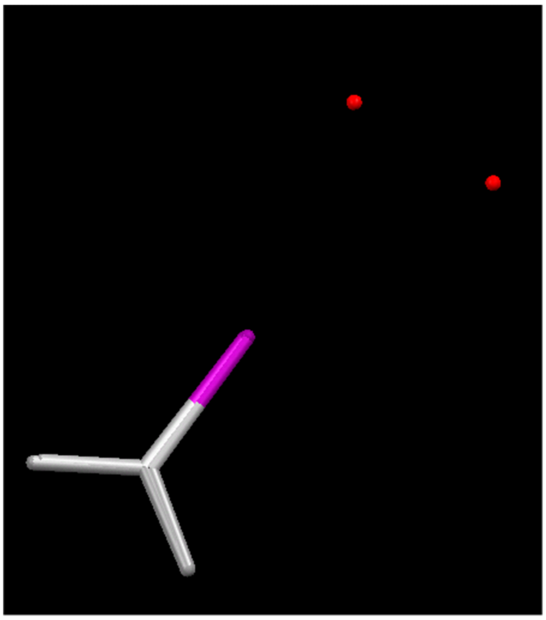
## 4.1 Introduction to Scatterplots in IsoStar

The presentation of nonbonded contacts as 3D scatterplots is best explained by an example. Suppose we are interested in the interaction of aromatic iodo substituents, C(ar)-I, with oxygen atoms. Using standard software (ConQuest), the CSD can be searched to find crystal structures containing C(ar)-I groups that form iodine...oxygen nonbonded contacts shorter than some specified distance (we use sum of van der Waals radii + 0.5Å). Two of the first hits found in this search are I...O contacts in CSD entries ACICOD and AIBFOR10:



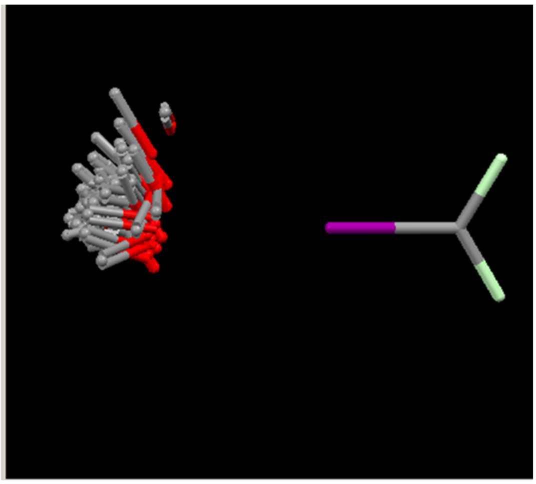


The iodine...oxygen fragments are extracted from each structure, including the first three carbon atoms of the aromatic ring. The iodine and ring atoms of one fragment are then overlaid on those from the other to give a composite picture in which the I...O contacts from the two structures are displayed simultaneously relative to the same frame of reference:



Repetition of this process for all hits found in the CSD produces a scatterplot showing how oxygens are distributed around aromatic iodo substituents in small-molecule crystal structures.

In IsoStar, this scatterplot is displayed in three dimensions, and is symmetrised (i.e. all the contacts are reflected into one quadrant, since the aromatic iodine group has two planes of mirror symmetry). A scatterplot showing all C(ar)-I...O=C contacts shorter than sum of van der Waals radii - 0.2Å, is shown below. Note that the carbonyl carbon atoms are shown as well as the oxygen atoms that are actually forming the short contacts:



It sometimes happens that a particular type of contact occurs so many times in the CSD that the resulting scatterplot is very crowded and hard to inspect. In such cases, a pseudo-random subset of contacts is taken to limit the total number of atoms in the scatterplot to about 10,000.

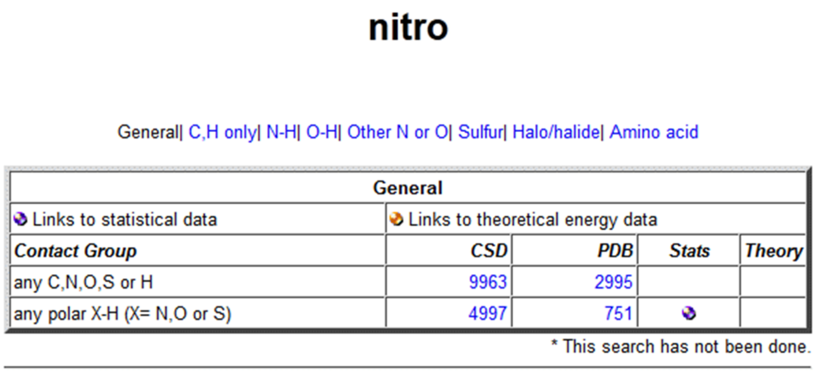
PDB-based scatterplots are produced in an analogous way, capturing information about protein-ligand, ligand-water and protein-water interactions.

# 4.2 Viewing Scatterplots

## 4.2.1 Loading and Reloading Scatterplots

Use the IsoStar web-browser interface to navigate to the table relating to the central group of interest (see Accessing Data for a Particular Central Group).

To view a scatterplot based on CSD information, click on a number in the CSD column of the table. For example, clicking on the number 4997 below will display the scatterplot of nitro...H-X contacts (X = O, N or S), as observed in small-molecule crystal structures in the CSD (4997 is the number of contacts in the scatterplot).



# nitro

General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

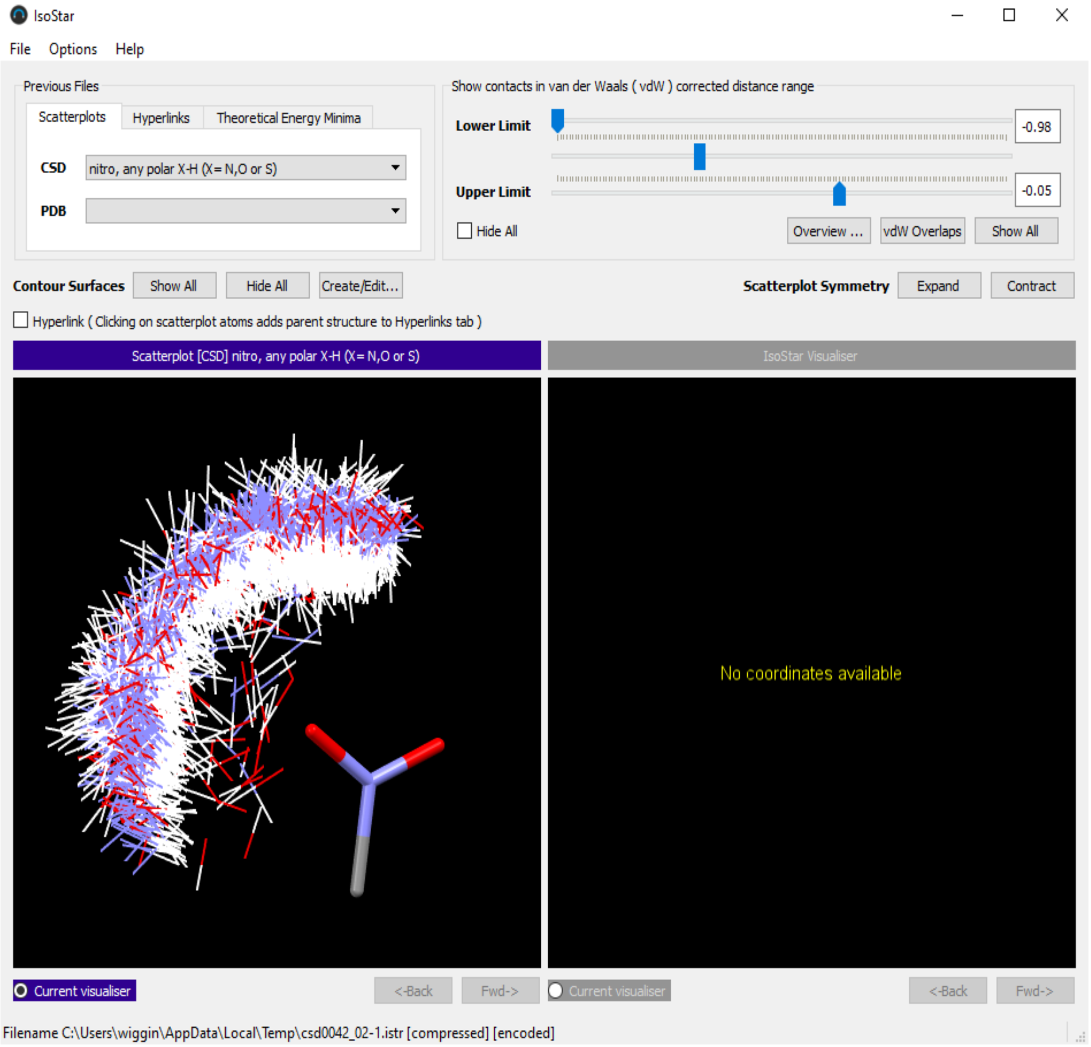
| General | | | | |
|---|---|---|---|---|
| ● Links to statistical data | ● Links to theoretical energy data | | | |
| ***Contact Group*** | ***CSD*** | ***PDB*** | ***Stats*** | ***Theory*** |
| any C,N,O,S or H | 9963 | 2995 | | |
| any polar X-H (X= N,O or S) | 4997 | 751 | ● | |

* This search has not been done.

Clicking on the adjacent number, 751, will show the scatterplot of nitro...H-X contacts observed in protein-ligand complexes from the PDB.

If a cell in the CSD or PDB columns contains an asterisk, it means that no scatterplot has been generated for that particular contact (normally because there are no experimental data).

The selected scatterplot is loaded into one of the two 3D visualisers in the IsoStar graphical interface:



CSD: nitro, any polar X-H (X= N,O or S)

PDB:

Show contacts in van der Waals ( vdW ) corrected distance range

Lower Limit: -0.98

Upper Limit: -0.05

☐ Hide All

☐ Hyperlink ( Clicking on scatterplot atoms adds parent structure to Hyperlinks tab )

Scatterplot [CSD] nitro, any polar X-H (X= N,O or S)

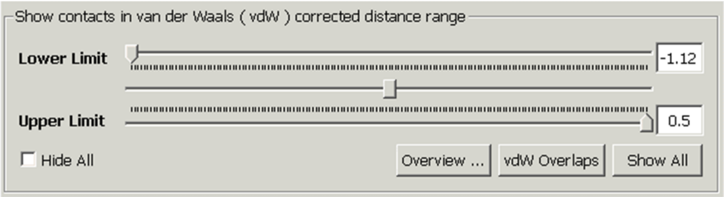
No coordinates available

Filename C:\Users\wiggin\AppData\Local\Temp\csd0042_02-1.istr [compressed] [encoded]

The graphical interface has options for controlling which visualiser is used when further scatterplots are loaded (see Controlling which Visualiser is Used), and for reloading a scatterplot that has already been viewed (see Reloading Previously Viewed Scatterplots, Structures and Minima).

### 4.2.2 Displaying Contacts in a Particular Distance Range

The top-right of the IsoStar graphical interface contains options that can be used when the current visualiser (see Controlling which Visualiser is Used) contains a scatterplot. They allow control over which contacts are displayed.



Show contacts in van der Waals ( vdW ) corrected distance range

Lower Limit -1.12

Upper Limit 0.5

□ Hide All

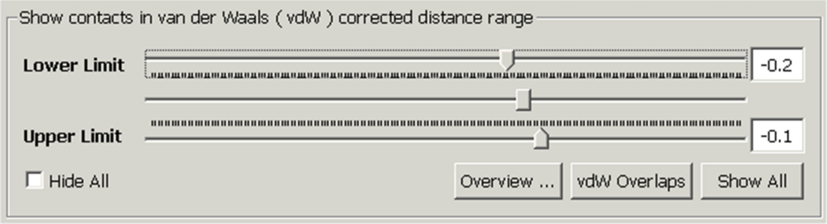
By default, the scatterplot will show all contacts shorter than (V + 0.5) Å, where V is the sum of the van der Waals radii (see Nonbonded Contact Definition; van der Waals Radii) of the atoms involved. Hitting the **vdW overlaps** button hides all contact groups except for those within van der Waals contact of any target atom (see Substructure-Search Details) in the central group; in other words, you will see only the van der Waals "clashes".

The **Overview** button lists the CSD reference codes for all the contact groups in the scatterplot, ordered from shortest contact to longest. Also available are statistics relating to the scatterplot (see Viewing Scatterplot Contact Distributions).

Hitting **Show All** will display all contacts.

Switching on the **Hide All** check box will hide all contacts; this is useful if you want to create a picture of a central group with a contoured density surface (see Displaying Contoured Density Surfaces).

The sliders allow interactive control over which contacts are displayed. If the top (Lower Limit) slider is moved to the right, short contacts will be hidden from the display. If the bottom (Upper Limit) slider is moved to the left, long contacts will be hidden. At any given slider setting, the top and bottom white boxes show the limits of the current display. For example, with the sliders in the positions shown below, all contacts in the distance range (V - 0.1) to (V + 0.2) Å are displayed, where, again, V is the sum of the van der Waals radii of the atoms involved.

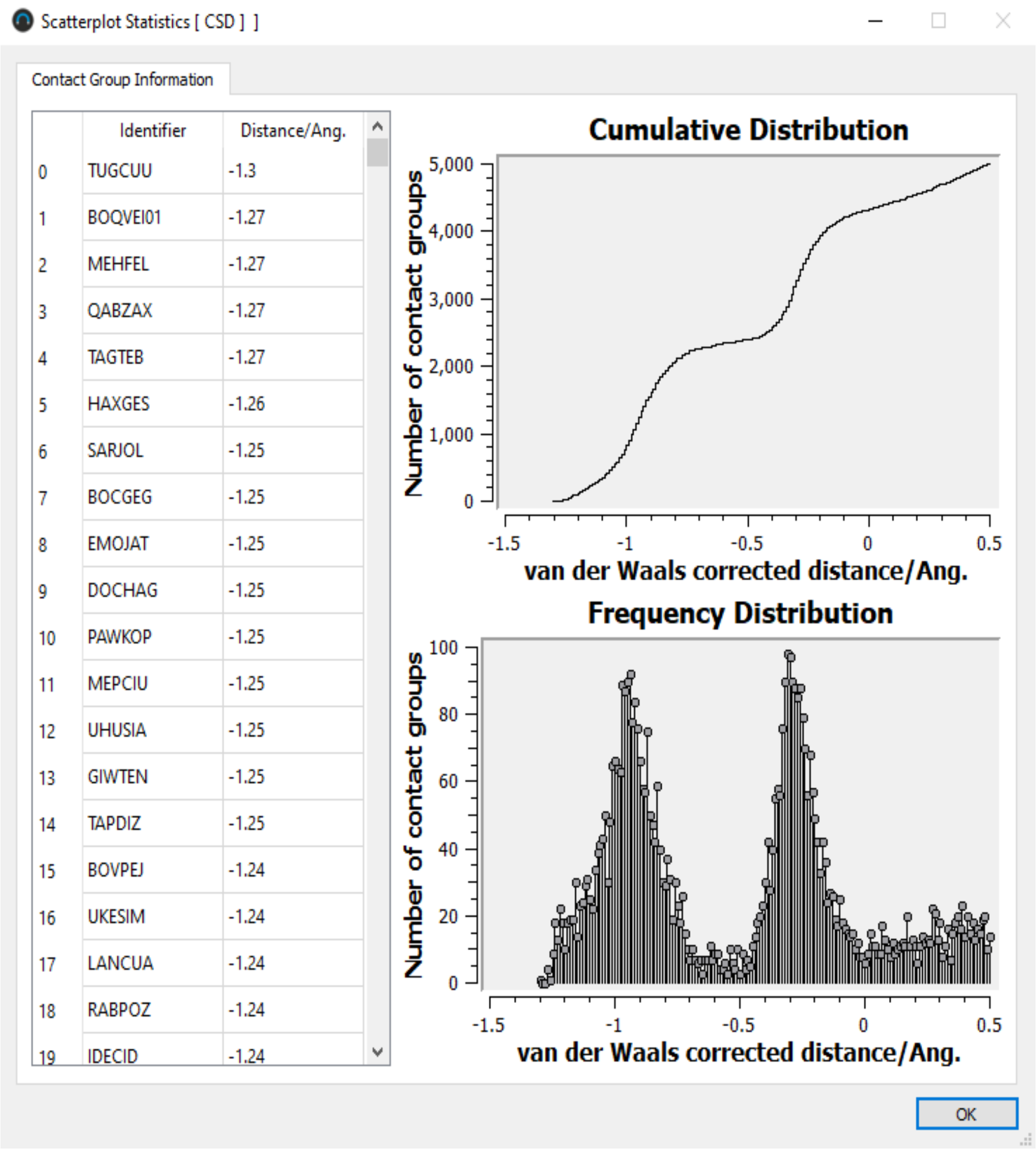


You can type distance limits directly into the white boxes, but you must terminate each number with a carriage return (i.e. **Enter**) for it to take effect.

Moving the middle grey rectangle with the mouse alters both sliders simultaneously. By setting a narrow range (e.g. the above setting corresponds to a range of 0.3Å), this therefore allows you to "slice through" the contacts.

### 4.2.3 Viewing Scatterplot Contact Distributions

Contact distribution histogram plots of the scatterplot as well as a list of refcodes that make up the plot (ordered by shortest to longest contact distance) are provided under the **Overview** button.



| | Identifier | Distance/Ang. |
|---|---|---|
| 0 | TUGCUU | -1.3 |
| 1 | BOQVEI01 | -1.27 |
| 2 | MEHFEL | -1.27 |
| 3 | QABZAX | -1.27 |
| 4 | TAGTEB | -1.27 |
| 5 | HAXGES | -1.26 |
| 6 | SARJOL | -1.25 |
| 7 | BOCGEG | -1.25 |
| 8 | EMOJAT | -1.25 |
| 9 | DOCHAG | -1.25 |
| 10 | PAWKOP | -1.25 |
| 11 | MEPCIU | -1.25 |
| 12 | UHUSIA | -1.25 |
| 13 | GIWTEN | -1.25 |
| 14 | TAPDIZ | -1.25 |
| 15 | BOVPEJ | -1.24 |
| 16 | UKESIM | -1.24 |
| 17 | LANCUA | -1.24 |
| 18 | RABPOZ | -1.24 |
| 19 | IDECID | -1.24 |

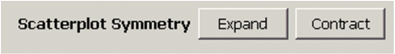
Clicking on a refcode e.g. TULWOM in the **Identifier** list selects the contact group in the scatterplot.

The **Cumulative Distribution** provided in the top right of the dialogue graphs the number of contacts versus the contact distance (the contact distance is expressed relative to the sum of van der Waals radii of the contact atoms).

The **Frequency Distribution** shows the number of contacts at each contact distance.

### 4.2.4 Expanding and Contracting Scatterplot Symmetry

In IsoStar scatterplots are symmetrised i.e. all the contacts are reflected into an area of space based on the symmetry of the central group (e.g. only one quarter of an aromatic iodine group with its contact group is shown since the central group has two planes of mirror symmetry whereas half of all contacts to an azide group are shown since it has one plane of mirror symmetry).



The symmetry can be expanded if required using the **Expand** button, next to **Scatterplot Symmetry**.

To contract the symmetry (i.e. to return the scatterplot to its original settings) hit the **Contract** button.

## 4.2.5 Display of Hydrogen Atoms in Scatterplots

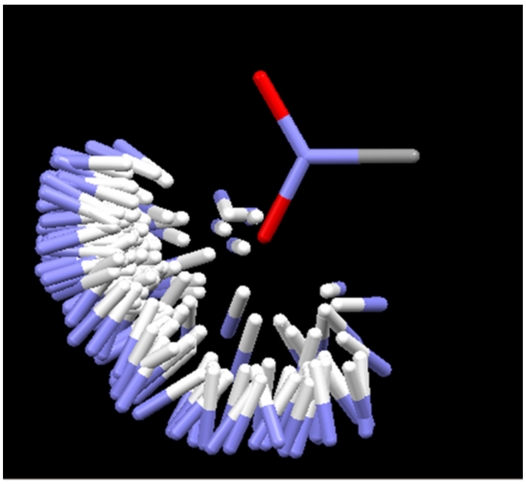
Hydrogen atoms are always included in CSD-based scatterplots but never in PDB-based plots.

This is because hydrogen atoms are almost never located in protein crystal structures whereas, usually, they are in small-molecule structures (and when IsoStar scatterplots are created, CSD structures will only be used if all the relevant hydrogen-atoms positions are available).

The lack of hydrogen atoms in PDB-based plots sometimes makes the plots difficult to interpret (see Effects of Missing Hydrogen Atoms on PDB-Based Scatterplots).

## 4.2.6 Display of Scatterplots for Symmetrical Central Groups

Many of the central groups in IsoStar are symmetrical. For these, only the asymmetric unit of the scatterplot is shown by default. For example, the scatterplot of N-H around nitro has all the N-H groups reflected into one quadrant since the nitro group has mm symmetry:



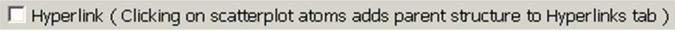
Linear groups such as cyano and ethynyl pose a problem, as they have cylindrical symmetry. The problem is overcome by including enough extra atoms to break this infinite symmetry. For example, IsoStar contains C(ar)-CN and C($sp^3$)-CN entries, having mm and 3m symmetry, respectively.

### 4.2.7 Customising a Scatterplot Display

Clicking anywhere in a visualizer with the right mouse button will display a menu which allows you to alter the visual appearance (e.g. atom colours and styles) of a scatterplot (see Using the 3D Visualiser).

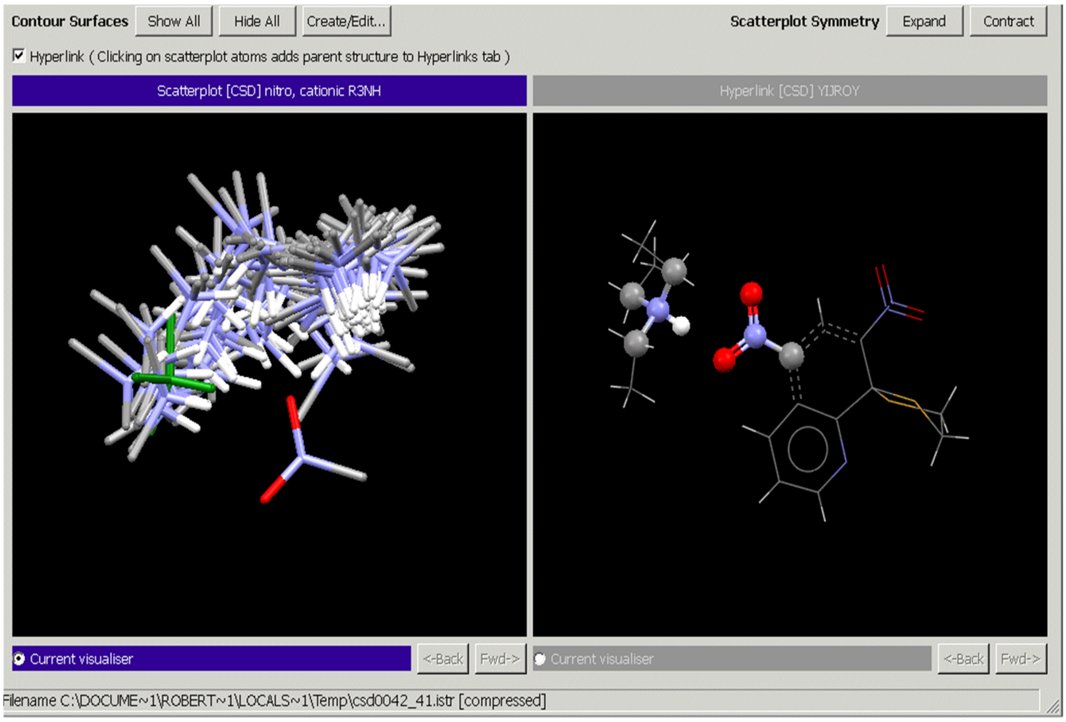
## 4.3 Hyperlinking from a Scatterplot to a Crystal Structure

Hyperlinking is only operational when the current visualizer (see Controlling which Visualiser is Used) contains a scatterplot and when the **Hyperlink** check-box has been switched on:



□ Hyperlink ( Clicking on scatterplot atoms adds parent structure to Hyperlinks tab )

In this situation, clicking on any contact-group atom in the scatterplot will cause the crystal structure from which that contact was taken to be displayed in the other visualiser window, e.g.



☑ Hyperlink ( Clicking on scatterplot atoms adds parent structure to Hyperlinks tab )

Scatterplot [CSD] nitro, cationic R3NH

Hyperlink [CSD] YIJROY

● Current visualiser

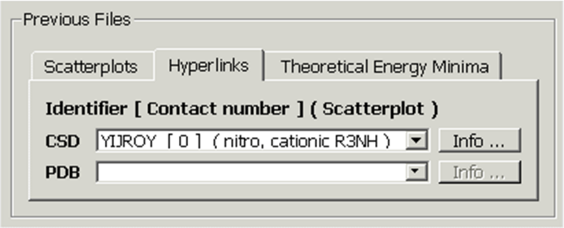
○ Current visualiser

Filename C:\DOCUME~1\ROBERT~1\LOCALS~1\Temp\csd0042_41.istr [compressed]

The contact that was selected from the scatterplot will be highlighted in green; this makes it easier to select contacts from a cluster without accidentally picking the same one twice. The highlighted contacts can be set to their original colours by right-clicking in the scatterplot visualiser, selecting **Colours** from the resulting menu, and then **Colour by Element** from the next menu.

In the crystal-structure display, the atoms involved in the contact are highlighted by being displayed in ball-and-stick mode.

The database identifier of the structure that has been loaded will be displayed in the **Hyperlinks** area (top left of interface), either in the CSD or PDB box, depending on which database the structure came from:



Identifier [ Contact number ] ( Scatterplot )

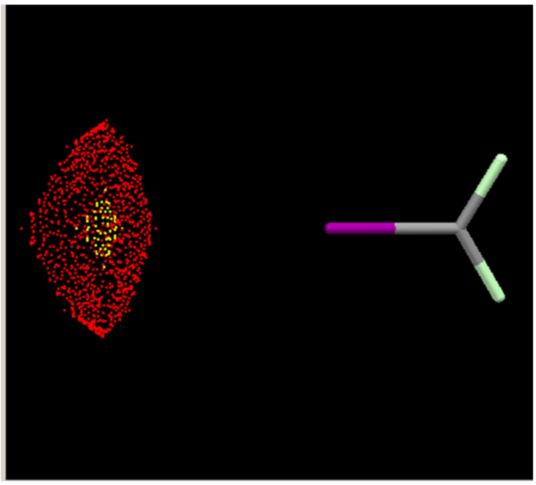
CSD YIJROY [ 0 ] ( nitro, cationic R3NH )

PDB

Hitting the **Info...** button next to the structure identifier will display further information (e.g. literature reference) about the structure.

# 4.4 Displaying Contoured Density Surfaces

Scatterplots can be converted to contoured density surfaces. For example, the scatterplot of carbonyl groups around iodo substituents (see Introduction to Scatterplots in IsoStar), can be displayed as shown below. This surface highlights the regions in space where the carbonyl oxygen atoms accumulate:

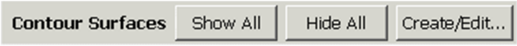


Here, the contour colours have been chosen so that red denotes regions preferred by the O atoms and yellow denotes the most preferred region. In a similar way, a surface could be produced showing where the carbonyl carbon atoms tend to be.

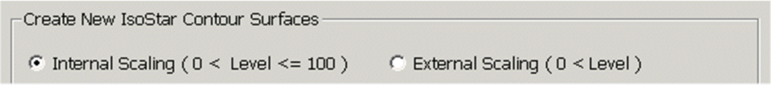
The surfaces are calculated by dividing the space around the central group into a regular grid and counting the observed number of contact-group atoms (carbonyl oxygens or carbons in this case) in each grid volume. This produces an empirical estimate of the density of contact-group atoms at each of a regular matrix of grid points. After application of a smoothing algorithm, the densities are contoured at user-specified levels and displayed as surfaces.

## 4.4.1 Creating Surfaces

With a scatterplot loaded (see Loading and Reloading Scatterplots) in the current visualiser (see Controlling which Visualiser is Used), click on the **Create/Edit** button adjacent to Contour Surfaces.



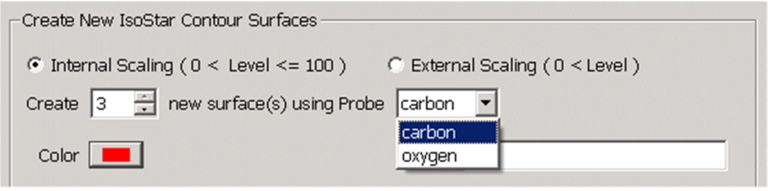
In the resulting dialogue window, pick whether you want to use internal or external scaling (see Internal and External Scaling of Surfaces), e.g.



Create New IsoStar Contour Surfaces

● Internal Scaling ( 0 < Level <= 100 ) ○ External Scaling ( 0 < Level )

Depending on which scatterplot is being contoured, you may be given a choice of probes. For example, if the contact group is C=O, it will be possible to contour on the positions of the carbon or oxygen atoms. Select the desired probe from the pull-down menu, e.g.



Create New IsoStar Contour Surfaces

● Internal Scaling ( 0 < Level <= 100 ) ○ External Scaling ( 0 < Level )

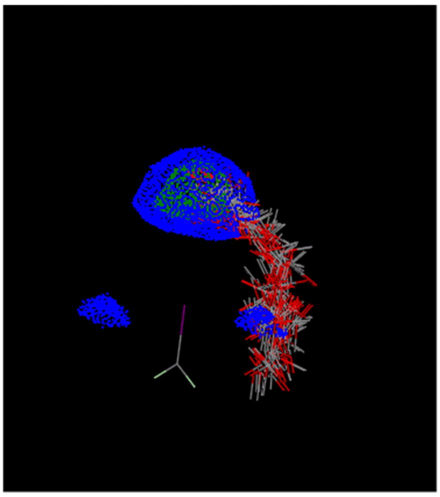
Create 3 new surface(s) using Probe carbon

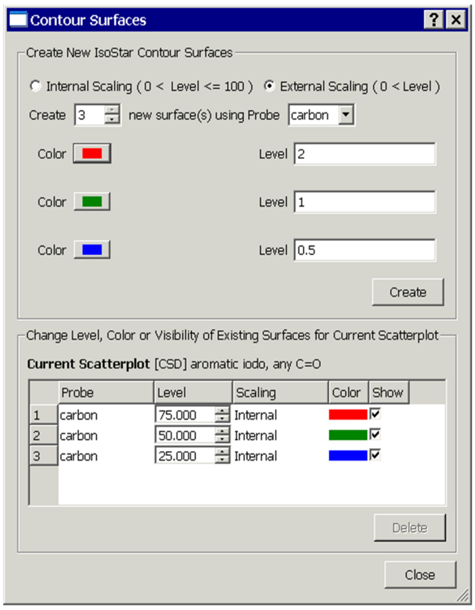
Color 🟥

Contour levels are provided however these may be user-defined by typing new values into the Level boxes. For internal scaling, levels of between 20 to 90 are usually suitable; for external scaling, values between 1 to 4 are often appropriate.

Contour colours can be changed by clicking on the **Color** button adjacent to each contour. This opens a palette from which any desired colour may be chosen.

Once all the levels and colours are as you want them, click the **Create** button. The contours at the specified levels and colours will appear in the 3D view.





Create New IsoStar Contour Surfaces

○ Internal Scaling ( 0 < Level <= 100 ) ● External Scaling ( 0 < Level )

Create 3 new surface(s) using Probe carbon

Color 🟥 Level 2

Color 🟩 Level 1

Color 🟦 Level 0.5

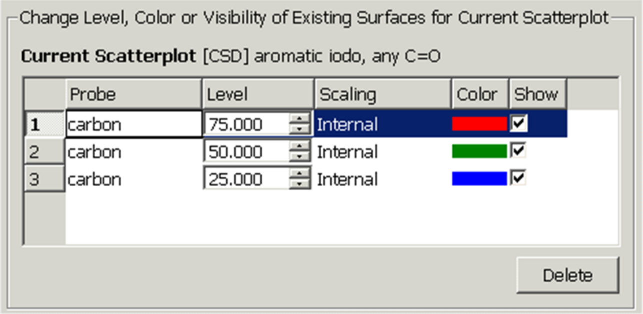
Change Level, Color or Visibility of Existing Surfaces for Current Scatterplot

**Current Scatterplot** [CSD] aromatic iodo, any C=O

| | Probe | Level | Scaling | Color | Show |
|---|---|---|---|---|---|
| 1 | carbon | 75.000 | Internal | 🟥 | ☑ |
| 2 | carbon | 50.000 | Internal | 🟩 | ☑ |
| 3 | carbon | 25.000 | Internal | 🟦 | ☑ |

If you want to see the surface but not the contacts, the latter may be hidden by switching on the **Hide All** check box, which is near the top-right of the main graphical interface.

The bottom of the Contour Surfaces dialogue lists the contours that have been generated, their levels and their colours.



Change Level, Color or Visibility of Existing Surfaces for Current Scatterplot

**Current Scatterplot** [CSD] aromatic iodo, any C=O

|   | Probe | Level | Scaling | Color | Show |
|---|---|---|---|---|---|
| 1 | carbon | 75.000 | Internal | 🟥 | ☑ |
| 2 | carbon | 50.000 | Internal | 🟩 | ☑ |
| 3 | carbon | 25.000 | Internal | 🟦 | ☑ |

Several options are available from within the Contour Surfaces dialogue (the display updates automatically):

- Level: use the up and down arrows to modify the contour levels.
- Color: click on the coloured rectangles to change the contour colours.
- Show: control the display of individual contours by switching the Show tickbox on or off. Note that the display of contour surfaces can also be controlled using the **Show All** and **Hide All** buttons on the main graphical interface.

To delete a contour, first select it by clicking on the appropriate **Contour** entry in the **Current Scatterplot** table, then hit **Delete**.

More surfaces, at different contour levels, can be added by repeating the above process.

If you calculate a surface and nothing appears in the visualiser, the probable reason is that the contour level is too high.

## 4.4.2 Internal and External Scaling of Surfaces

When contouring a scatterplot, there are two ways of defining the contour levels. In **Internal scaling**, the position of maximum contact-atom density is arbitrarily assigned a contour level of 100. Every other density value is then scaled relative to this. **External scaling** (only available for CSD-based scatterplots) is more complicated. An estimate is made of the average density of contact-group atoms in all CSD structures containing both the central group and the contact group. Then, the actual density of contact-group atoms at each grid point is divided by the average density. This gives a figure that represents whether the density at that grid point is higher or lower than would be expected by chance. For example, a figure of 2.0 would imply that contacts at that grid point were twice as frequent as would be expected; a figure of 0.5 would suggest that the density of contacts was only half the random expectation. Contouring is then done on these values.

The practical difference between internal and external scaling is seen when two different scatterplots are contoured and compared. If internal scaling is used, the scaling of each is purely local, so the contour levels of one bear no relation to those of the other. If external scaling is used, both plots are nominally on the same scale: in each case, a contour level of x means that contacts are x times as frequent as would be expected by chance.

## 4.5 Saving Plots

Files can be saved out via **File, Save As**. The menu will change depending on the type of plot currently on display.

### 4.5.1 Saving Scatterplots

To save a scatterplot, first ensure that it is being displayed in the current visualiser (see Controlling which Visualiser is Used). Then select **File** from the top level menu, followed by **Save As**, then select one of the following options from the resultant menu:

- **IsoStar (istr)**: this will save the scatterplot in .istr format. This is basically the Certara (formerly Tripos) mol2 fomat with some extra information held in comment lines. Hence, it should be possible to read files of this type into a variety of visualisers, including the CCDC programs Mercury, https://www.ccdc.cam.ac.uk/solutions/csd-core/components/mercury/, and Hermes (the CCDC's dedicated protein structure visualizer: https://www.ccdc.cam.ac.uk/solutions/csd-core/components/hermes/).

- **Hyperlink Identifiers (.txt)**: depending on whether you are viewing a CSD-based or PDB-based scatterplot, this option will output a list of CSD refcodes or PDB codes in .txt. format.

- **mol2**: the scatterplot will be saved in mol2 format (i.e. the Certara format for 3D molecules and crystal structures).

- **pdb**: the scatterplot will be saved in the Protein Data Bank’s format for 3D molecules.

- **jpg**: a graphical image of the scatterplot as it appears in the 3D display will be saved in .jpg format.

All atoms will be written out, including any that are hidden (i.e. currently not displayed because of the settings that have been chosen for the distance slider).

### 4.5.2 Saving Contour Plots

To save a contour plot, first ensure that it is being displayed in the current visualiser (see Creating Surfaces). Then select **File** from the top level menu, followed by **Save As**, then select one of the following options from the resultant menu:

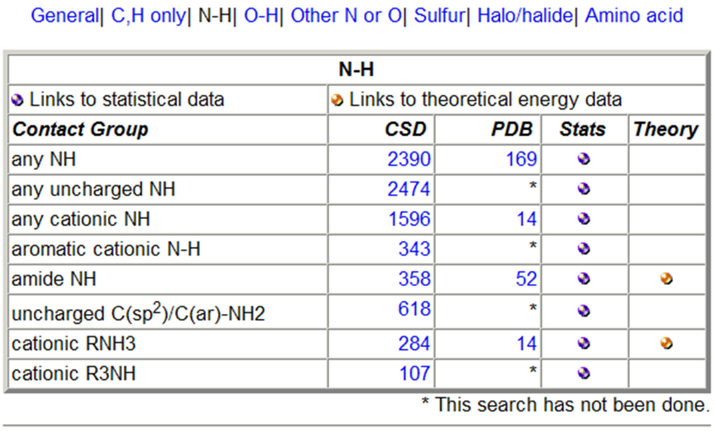
- **acnt**: a Sybyl ASCII contour file will be written out, containing details of the contour settings. This file can be read into other software packages, e.g. Hermes https://www.ccdc.cam.ac.uk/solutions/csd-core/components/hermes/ and the contours displayed.
- **jpg**: a graphical image of the contour plot as it appears in the 3D display will be saved in .jpg format.

# 5 Viewing Other Data

## 5.1 Theoretical Data

IsoStar contains the results of theoretical energy calculations performed on various model systems. Each system comprises two molecules that model a particular intermolecular interaction. For example, the interaction between the central group C(sp$^3$)COC(sp$^3$) (i.e. aliphatic ketone) and the contact group NH (amide) has been modelled by the molecular complex acetone - N-methylacetamide. Using a model system such as this, relatively quick calculations are first performed to search the potential energy hypersurface of the dimer. This enables low-energy orientations to be located. Higher quality ab-initio-based intermolecular energies are then computed for these low-energy dimer orientations, using intermolecular perturbation theory (see Details of Molecular-Orbital Methodology).

Theoretical results are associated with the scatterplot (see Introduction to Scatterplots in IsoStar) to which they are most relevant. Navigate to the web-browser page for the central group of interest (see Accessing Data for a Particular Central Group) and click on one of the gold buttons (which are only displayed when relevant theoretical results are available), e.g.

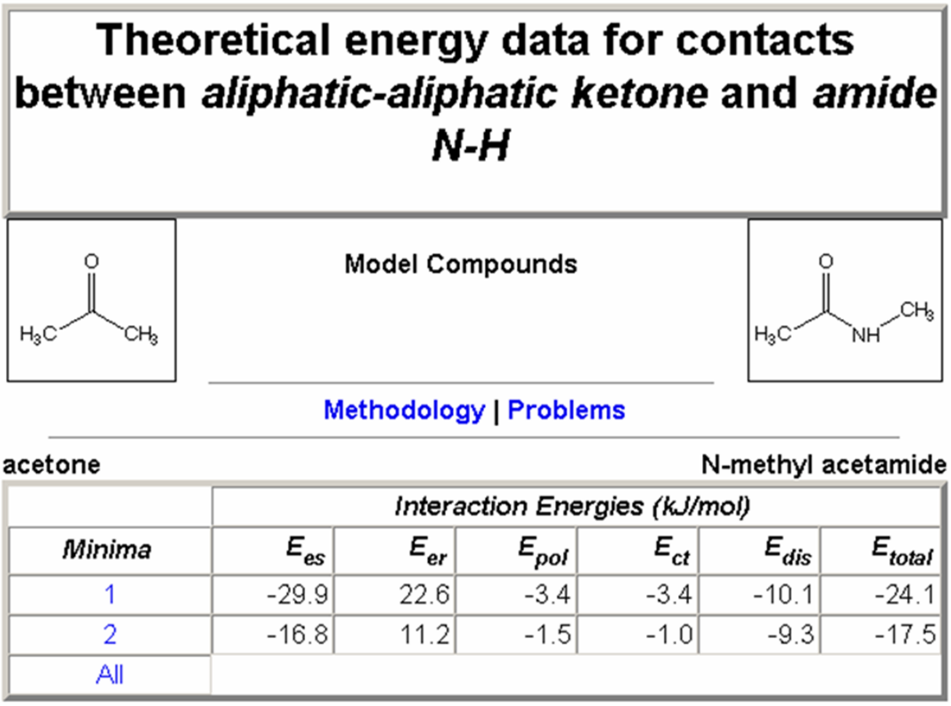


General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

| N-H | | | | |
|---|---|---|---|---|
| ● Links to statistical data | ● Links to theoretical energy data | | | |
| ***Contact Group*** | ***CSD*** | ***PDB*** | ***Stats*** | ***Theory*** |
| any NH | 2390 | 169 | ● | |
| any uncharged NH | 2474 | * | ● | |
| any cationic NH | 1596 | 14 | ● | |
| aromatic cationic N-H | 343 | * | ● | |
| amide NH | 358 | 52 | ● | ● |
| uncharged C($sp^2$)/C(ar)-NH2 | 618 | * | ● | |
| cationic RNH3 | 284 | 14 | ● | ● |
| cationic R3NH | 107 | * | ● | |

* This search has not been done.

This produces a table listing calculated dimer energies (units kJ/mol). These include the total interaction energy of each geometrical minimum together with a breakdown into the component physical contributions (see Details of Molecular-Orbital Methodology), e.g.



**Theoretical energy data for contacts between *aliphatic-aliphatic ketone* and *amide N-H***

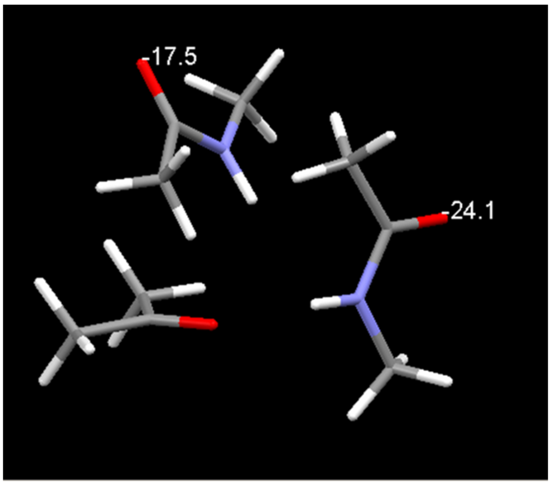
**Model Compounds**

**Methodology | Problems**

**acetone** **N-methyl acetamide**

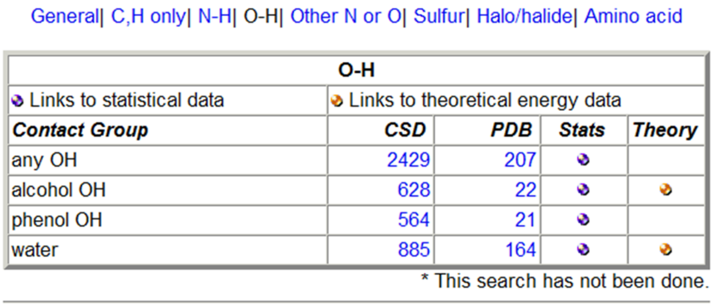
| | *Interaction Energies (kJ/mol)* | | | | | |
|---|---|---|---|---|---|---|
| ***Minima*** | $E_{es}$ | $E_{er}$ | $E_{pol}$ | $E_{ct}$ | $E_{dis}$ | $E_{total}$ |
| 1 | -29.9 | 22.6 | -3.4 | -3.4 | -10.1 | -24.1 |
| 2 | -16.8 | 11.2 | -1.5 | -1.0 | -9.3 | -17.5 |
| All | | | | | | |

Calculated minimum-energy dimer geometries can be visualised by clicking on one of the numbers in the Minima column (displays an individual minimum) or clicking on **All** (displays all minima), e.g.



## 5.2 Statistical Data

Some simple statistics are included in IsoStar. They can be displayed by navigating to the web-browser page for the central group of interest (see Accessing Data for a Particular Central Group) and clicking on one of the blue buttons in the **Stats** column, e.g.

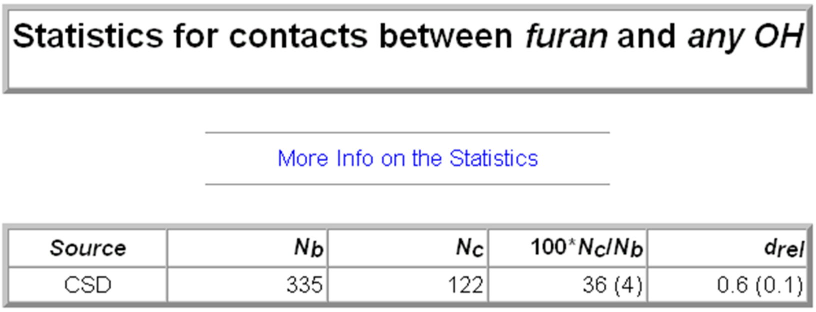


General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

| O-H | | | | |
|---|---|---|---|---|
| ● Links to statistical data | ● Links to theoretical energy data | | | |
| ***Contact Group*** | ***CSD*** | ***PDB*** | ***Stats*** | ***Theory*** |
| any OH | 2429 | 207 | ● | |
| alcohol OH | 628 | 22 | ● | ● |
| phenol OH | 564 | 21 | ● | |
| water | 885 | 164 | ● | ● |

* This search has not been done.

The following example is taken from a typical statistics page (furan...any OH):



**Statistics for contacts between *furan* and *any OH***

More Info on the Statistics

| *Source* | *Nb* | *Nc* | *100*Nc/Nb* | *drel* |
|---|---|---|---|---|
| CSD | 335 | 122 | 36 (4) | 0.6 (0.1) |

The statistics include:

The number of crystal structures ($N_b$) in the CSD that contain both groups, irrespective of whether they form a contact. In the above example, this would be the number of CSD structures containing both a furan ring and a hydroxyl group.

The number of crystal structures ($N_c$) in which the two groups form at least one intermolecular interaction that is shorter than the sum of van der Waals radii. In the above example, this would be the number of CSD structures containing a furan ring and a hydroxyl group in van der Waals contact.

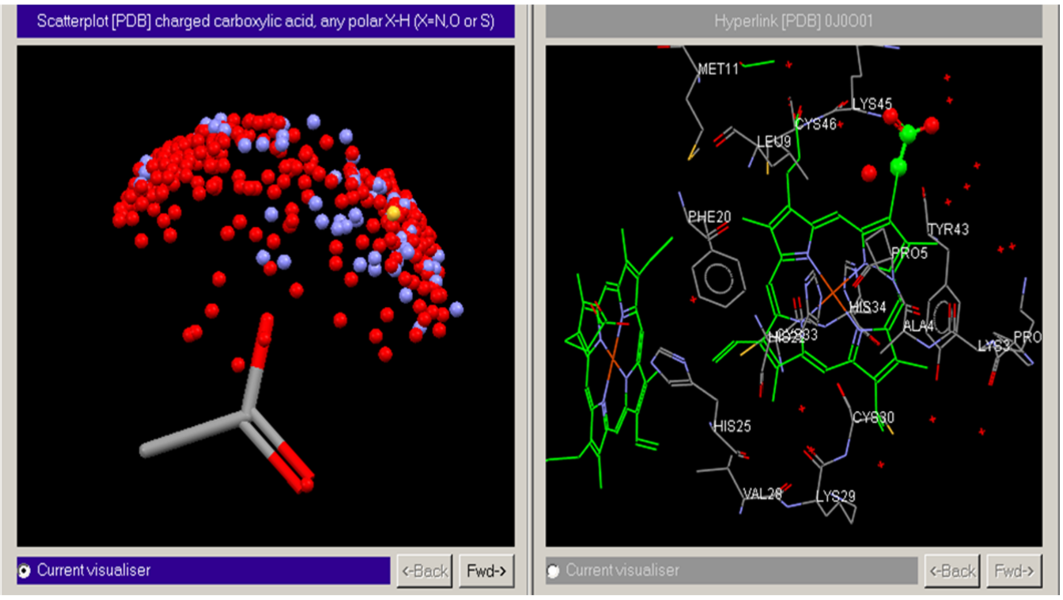
The ratio $N_c/N_b$, expressed as a percentage. The bigger the ratio, the more often the groups form short interactions. Obviously, the ratio says something about the extent to which the groups attract each other. However, it is only a very crude measure, as it does not take stoichiometric factors into account. For example, $N_c/N_b$ tends to be high when the contact group is C(ar)-H, simply because most structures contain many C(ar)-H groupings. An estimated standard deviation for the ratio is given in brackets, assuming Poisson statistics.

The relative density, $d_{rel}$. This is a somewhat less crude measure of the tendency for the groups to form short interactions. It is defined as $d_{rel} = d_{short}/d_{long}$, where $d_{short}$ is the density of contacts within the sum of van der Waals radii, V, and $d_{long}$ is the density of contacts between V and V+Tol. The larger $d_{rel}$ is, the greater the tendency for the groups to form short interactions. Typical values for $d_{rel}$ are 2-8 for a strong hydrogen bond (e.g. an H-bond where one or both of the groups is/are charged), 1-2 for an average hydrogen bond, 0.7-1 for a weak H-bond, and 0.3-0.7 for a very weak attractive interaction. An estimated standard deviation for $d_{rel}$ is given in brackets, assuming Poisson statistics.

# 6 Using the 3D Visualiser

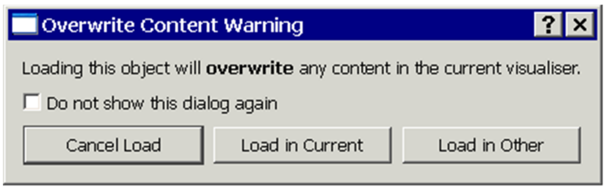
## 6.1 Controlling which Visualiser is Used

The IsoStar graphical interface contains two visualisers. At any given time, one of them is the current visualiser, denoted by blue bars above and below the visualiser area. For example, in the screenshot below, the current visualiser is the one on the left:



If you click in a visualiser, it becomes the current visualiser.

The first scatterplot or theoretical energy minimum to be loaded into the IsoStar graphical interface will be placed in the left-hand visualiser. By default, if a second scatterplot or theoretical energy minimum is loaded, a dialogue will appear which allows you to specify which visualiser is to be used (i.e. whether to overwrite the first scatterplot or use the empty visualiser):



Overwrite Content Warning

Loading this object will **overwrite** any content in the current visualiser.

□ Do not show this dialog again

Cancel Load | Load in Current | Load in Other

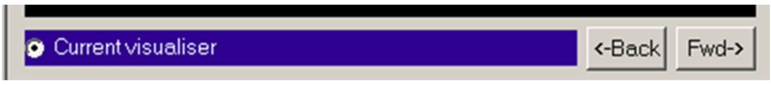
If you switch on the **Do not show this dialog again** check box (or if you select **Options** from the top-level menu and disable **Show Overwrite Visualiser Warning**), this dialogue will not be shown again and any further scatterplots or theoretical energy minima will be loaded into the current visualiser.

When you hyperlink (see Hyperlinking from a Scatterplot to a Crystal Structure) from a scatterplot, the crystal structure will always be loaded into the other visualiser (i.e. the visualiser not containing the scatterplot).

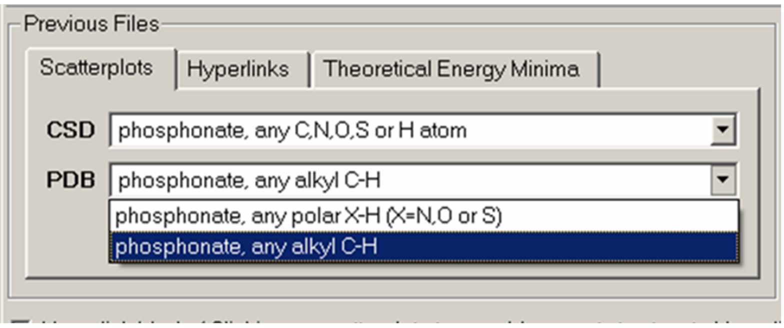
## 6.2 Reloading Previously Viewed Scatterplots, Structures and Minima

Scatterplots, theoretical energy minima and crystal-structures that have been previously loaded in a session may be reloaded in two ways:

- By use of the **Back** and **Fwd** buttons underneath the visualiser; these move backwards or forwards through the items that have been loaded into the visualiser:



- By use of the **Previous Files** options in the top-left of the IsoStar graphical interface. For example, to view a previously-loaded PDB-based scatterplot, hit the **Scatterplots** tab (this may not be necessary if the Scatterplots "pane" is already at the front) and then select the desired PDB scatterplot from the pull-down menu:


Previous Files
Scatterplots
Hyperlinks
Theoretical Energy Minima

CSD phosphonate, any C,N,O,S or H atom

PDB phosphonate, any alkyl C-H

phosphonate, any polar X-H (X=N,O or S)
phosphonate, any alkyl C-H

Access to previously-loaded CSD-scatterplots can be obtained via the **CSD** pull-down menu, and the **Hyperlinks** and **Theoretical Energy Minima** tabs similarly give access to pull-down menus of previously-loaded crystal structures and energy minima.

## 6.3 Rotating, Translating and Scaling

Rotate by moving the cursor around in the visualiser while keeping the left-hand mouse button pressed down. Rotate around the z-axis (the axis perpendicular to the screen) by keeping the left-hand mouse button and the **Shift** key pressed down.

Translate by moving the cursor in the visualiser with the centre mouse button depressed (requires 3-button mouse). Alternatively, use the left-hand button with the Control key pressed down.

Zoom in or out by moving the cursor up and down in the visualiser while keeping the right-hand mouse button pressed down.

## 6.4 Selecting and Deselecting Atoms

Selection of atoms and molecules is useful for changing properties such as display style. Atoms may be selected or deselected in several ways:

- Right-click anywhere in the visualiser (atom, bond or background) and choose **Selection** from the resulting menu. Various options for selecting atoms can then be chosen from the resulting menus.
- Right-click in the visualiser, select **Picking Mode** from the resulting menu and set the mode to **Default Picking Mode**. Click on individual atoms with the left mouse button to select them. Once selected, an atom can be deselected by clicking on it again.

When the picking mode is set to **Default Picking Mode**, all atoms become deselected if you left-click anywhere in the display-area background.

## 6.5 Setting Global Display Options

A choice of four display styles is offered: wireframe, capped stick, ball-and-stick, and space-filling. To change the display style globally, right click anywhere in the visualiser. Select **Styles** from the resulting pull-down menu and then pick the required style.

Aromatic rings are displayed with a circle within them by default. If this is not desired, right-click anywhere in the visualiser and select **Display Aromatic Rings** from the resulting menu.

Bond types (i.e. single, double, triple etc.) are displayed by default. If this is not desired, right-click anywhere in the visualiser and select **Display Bond Types** from the resulting menu.

By default, atoms are coloured by element. Scatterplots distinguish between polar hydrogen (X-H, where X is N,O or S) and non-polar hydrogen by colouring the polar H light green and the non-polar hydrogen white. To change the colour used for a particular element, right-click in the visualiser, pick **Colours** and then **Element colours...** from the resulting menus, and then hit the coloured box next to the element you want to change. This displays a colour palette from which a new colour can be chosen. Hitting **Defaults** at the bottom of the element-colour dialogue will reset all element colours to their default settings.

If the colours of some atoms have been altered (see Setting Display Properties for Particular Atoms and Bonds), you can return to colouring by element by right-clicking in the visualiser and picking **Colours** and then **Colour by Element**.

You can switch between the default black background and a blue gradient background by right-clicking in the visualiser and hitting **Draw Backdrop**.

## 6.6 Setting Display Properties for Particular Atoms and Bonds

To set the display properties of particular atoms, select them (see Selecting and Deselecting Atoms), right-click in the visualiser, select the appropriate option from the pull-down menu (**Styles, Colours, Labels, Show/Hide**) and then select the desired display-property setting from the next menu.

## 6.7 Labelling Atoms

Right-click anywhere in the visualiser and select **Labels** from the resulting pull-down menu, then hit **Show Labels** to label the atoms currently selected (see Selecting and Deselecting Atoms). If no atoms are selected, then all atoms will be labelled.

You can also switch individual atom labels on and off by right-clicking in the visualiser and selecting **Picking Mode** and then **Pick Labels**. When the **Pick Labels** option is active, left-clicking on an atom will toggle its label on and off.

To remove all labels, right-click in the visualiser, select **Labels** from the resulting pull-down menu, then hit **Hide Labels.**

The same pull-down menu gives access to options for changing the size and colours of labels.

## 6.8 Measuring Distances, Angles and Torsions

To measure distances, angles and torsions:

1. Right-click anywhere in the visualiser, select either **Measure** (or **Picking Mode)** from the resulting pull-down menu, then select either **Measure Distances, Measure Angles** or **Measure Torsions (or Pick Distances, Pick Angles, Pick Torsions)**. Depending on which mode has been chosen, you can then click on two, three or four atoms, respectively, to measure a distance, angle or torsion.

2. You will remain in the chosen measurement mode, so after measuring the first distance (or angle or torsion), you can continue measuring others.

3. To cancel the measurement mode, right-click anywhere in the visualiser and select **Measure** or **Picking Mode** followed by **Default Picking Mode**.

4. To clear all measurements from the display, right-click anywhere in the visualiser and select **Clear Measurements**.

# 7 Experimental Details and Limitations

## 7.1 Details of Scatterplot Methodology

### 7.1.1 Structures Used for Calculating CSD-Based Scatterplots

The Cambridge Structural Database (CSD; http://www.ccdc.cam.ac.uk/products/csd/) contains the results of organic and metallo-organic crystal structure determinations. Standard software (ConQuest) can be used to search the database for intermolecular contacts between any specified pair of chemical groups.

If the structure of a compound occurs more than once in the CSD (duplicate structure determinations of the same crystal form or determinations of different crystal forms), only one example is included in the searches used to set up the IsoStar scatterplots.

With this exception, all CSD structures are used for the IsoStar searches, regardless of crystallographic R-factor, presence of disorder, etc. However, CSD structures with missing hydrogen atoms are not used in creating IsoStar scatterplots unless the missing hydrogen atoms are irrelevant to the plot (i.e. not part of either the central or contact group).

### 7.1.2 Structures Used for Calculating PDB-Based Scatterplots

The Protein Data Bank (PDB; https://www.wwpdb.org/) contains the results of protein structures, including many protein-ligand complexes. Only protein-ligand, ligand-water and protein-water contacts are included in IsoStar scatterplots, i.e. contacts between protein atoms are ignored.

A subset of the database is used for IsoStar. Amongst the restrictions applied are:

- Crystallographic determinations only (no NMR structures).
- Resolution less than or equal to 2.5Å.
- Protein-ligand complexes only. A ligand is defined as a peptide of up to ten residues or a non-peptide or non-haem molecule of at least nine atoms.
- DNA and RNA complexes excluded.
- Most covalently-bonded complexes excluded.
- Most duplicate and near-duplicate structures excluded (e.g. structures with the words mutant, mutation or mutated in their PDB files).
- No ligands with atoms of partial occupancy.
- For structures containing more than one binding site with the same ligand, only one of the binding sites is used.

To generate the IsoStar plots, ligands of interest and their surrounding residues (i.e. binding sites) are identified in all structures. Crystallographic symmetry operators are applied to generate the full ligand environment where appropriate.

In addition, a few structures have been eliminated because they caused processing problems. This still leaves several thousand protein-ligand complexes for use in building IsoStar.

### 7.1.3 Nonbonded Contact Definition; van der Waals Radii

Only intermolecular contacts are included in IsoStar; intramolecular contacts are not considered.

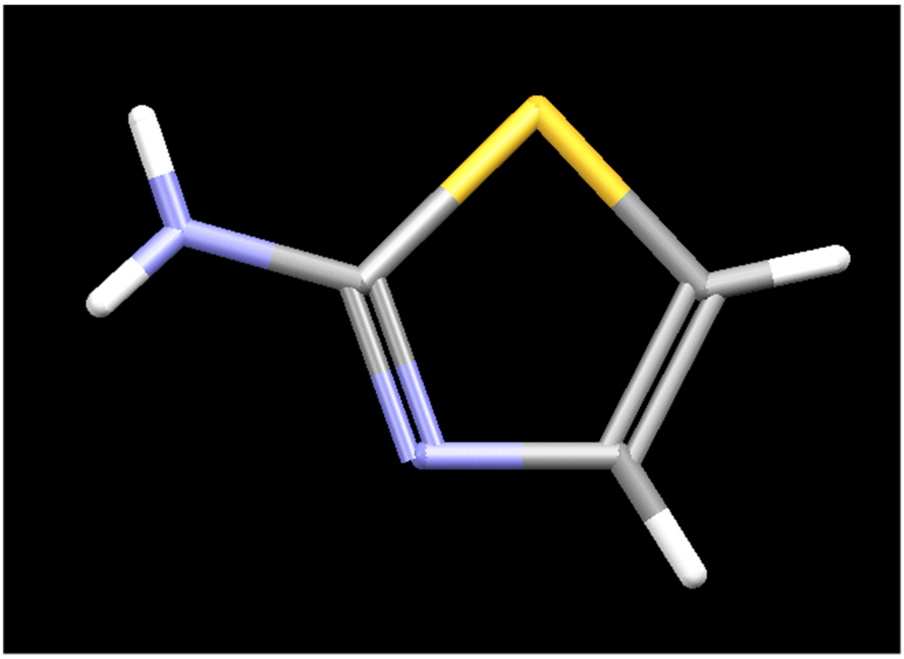
In both the CSD and PDB sections of IsoStar, nonbonded contacts are included up to a maximum distance of V + Tol, where V is the sum of the van der Waals radii of the atoms involved and Tol is a distance tolerance. Currently, a distance tolerance of 0.5 Å is used. Van der Waals radii are taken from S. Alvarez, Dalton Trans.., **42,** 8617-8636, 2013:

| Atom | Van der Waals radii |
|---|---|
| H | 1.20 Å |
| C | 1.77 Å |
| N | 1.66 Å |
| O | 1.50 Å |
| F | 1.46 Å |
| Si | 2.19 Å |
| P | 1.90 Å |
| S | 1.89 Å |
| Cl | 1.82 Å |
| Br | 1.86 Å |
| I | 2.04 Å |

### 7.1.4 Substructure-Search Details

A central group (see Overview of IsoStar) in IsoStar usually contains two types of atoms - target atoms and non-target atoms. Non-target atoms are there to help define the chemical environment of the group - e.g. the carbon atom in C(sp$^3$)-NH$_2$ - but are not included in nonbonded searches. In other words, only nonbonded contacts to target atoms are shown on IsoStar scatterplots. For example, scatterplots for C(sp$^3$)-NH$_2$ only show contacts to the nitrogen and hydrogen atoms.

Many of the central groups are ring systems, e.g. thiazole. In any particular CSD or PDB crystal structure, the ring is almost certain to be substituted. For example, one of the structures containing thiazole in the CSD, BAWKEP10, is substituted at the 2-position:



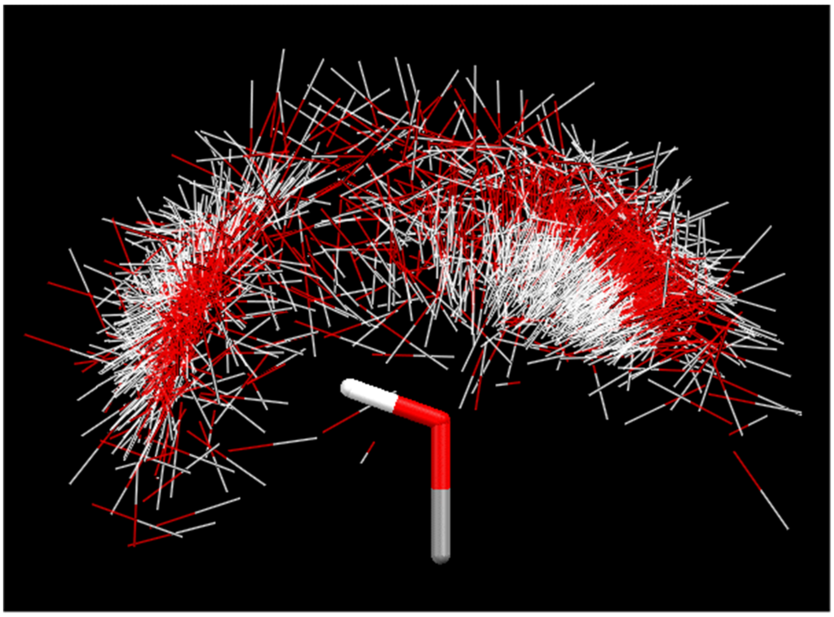
In IsoStar, contacts to ring substituents (e.g. the amino group above) are not shown. However, contacts to hydrogen atoms bonded to rings (e.g. the H atoms in the 4- and 5- positions above) are shown.

### 7.1.5 Treatment of Hydrogen-Atom Positions

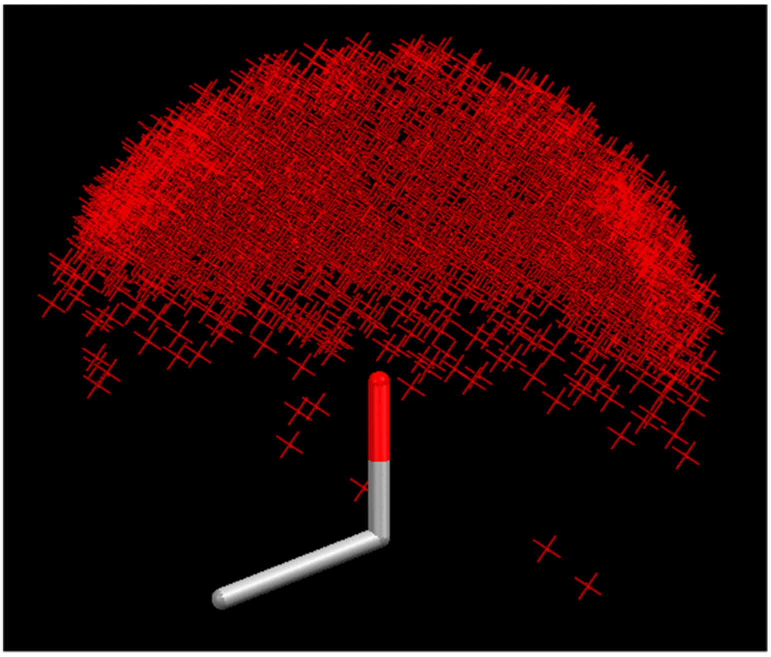
X-H covalent bond lengths in IsoStar scatterplots are normalised. Normalisation of X-H distances involves moving the H atom along the X-H vector until the X-H distance is equal to standard values (C-H = 1.083 Å, N-H = 1.009 Å, O-H = 0.983 Å). This corrects for the systematic underestimation of X-H bond lengths by X-ray diffraction.

### 7.1.6 Constructing Scatterplots when Central-Group Hydrogen Atoms are Missing

In some of the CSD-based scatterplots, the central group contains only three atoms, one of which is a hydrogen, e.g. C(sp$^3$)-OH. In such cases, the hydrogen atom must be used for overlaying when the scatterplot is constructed. Clearly, this is not possible with PDB data because the H-atom coordinates will be missing. In these cases, an extra non-hydrogen atom is therefore added to the central group, e.g. the CSD central group C(sp$^3$)-OH is changed to C-CH$_2$-OH for the PDB plot. The scatterplot below shows the distribution of oxygen atoms around an aliphatic alcohol, C(sp$^3$)-OH (CSD-based); all contacts are shorter than the sum of van der Waals radii:



The scatterplot below shows the distribution of oxygen atoms around an aliphatic alcohol, C-$CH_2$-O(H) (PDB-based); all contacts are shorter than the sum of van der Waals radii:

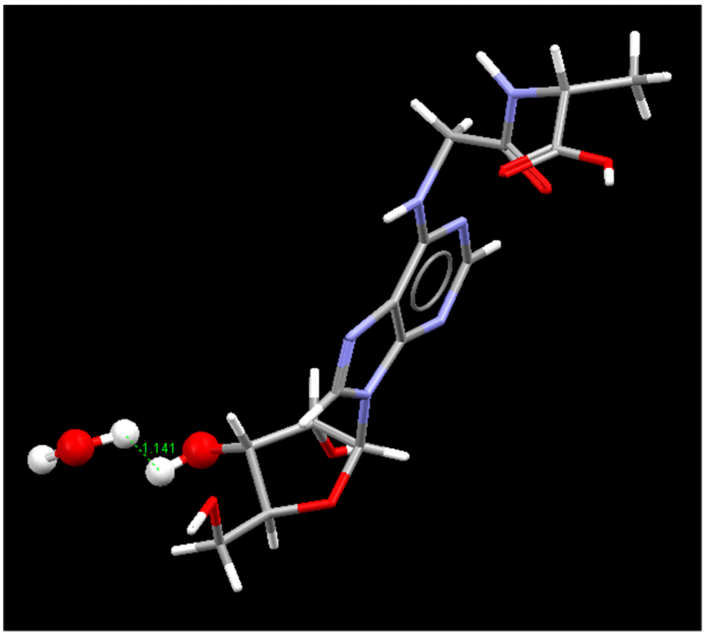


The absence of H atoms in protein structures means that there are no PDB plots for the central group water, since the water molecules cannot be overlaid.

## 7.2 Scatterplot Problems

### 7.2.1 Experimental Errors

Although CSD crystal structures are generally of high accuracy, the database inevitably contains some experimental errors. These are usually in the placement of hydrogen atoms. For example, the CSD crystal structure ADGLAL10 shows an impossibly short H...H contact of 1.14 Å. Clearly, a hydroxyl or water H atom has been misplaced in this structure.



Errors in PDB crystal structures are much more common because of their poorer resolution. However, the increasing use of refinement programs such as X-PLOR has tended to reduce the number of impossibly short contacts in protein crystal structures by including an empirical potential-energy term in the minimization function. While reducing the number of obvious errors, this introduces a subjective bias into the results since they reflect, to some degree, the empirical parameters used in the force field.

Experimental errors can be very noticeable in scatterplots because they often appear as outliers. Some obvious errors have been removed from the IsoStar library (specifically, any contact shorter than V - 1.3 Å has been removed, where V is the sum of the van der Waals radii of the atoms involved).

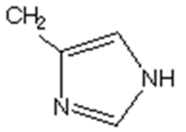
### 7.2.2 Effects of Missing Hydrogen Atoms on PDB-Based Scatterplots

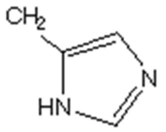
The fact that hydrogen atoms are almost never located in protein crystal-structure analysis causes difficulty in calculating PDB-based scatterplots for groups that (i) can exist in different ionisation states, (ii) have distinct conformations differing only in the positions of one or more hydrogen atoms, or (iii) have alternative tautomeric forms.

The ionisation states of acidic groups and basic groups cannot be determined unambiguously from protein crystal structures. They can sometimes be inferred by chemical reasoning, but not always and often not with confidence. In contrast, ionisation states are usually determined reliably in CSD structures. Accordingly, there are separate CSD-based scatterplots in IsoStar for, e.g., carboxylate ($CO_2^-$) and carboxylic acid ($CO_2H$) but only a single, combined PDB-based plot. Since the $pK_a$ of carboxylic acids is relatively low, most of the contributors to the PDB-based $CO_2^-/CO_2H$ plot will be ionised, but probably not all.

Similarly, alternative geometries that differ only in H-atom positions cannot be distinguished in protein structures. Thus, for example, there are separate CSD-based scatterplots for planar and pyramidal aromatic amino groups, but only a single PDB-based plot.

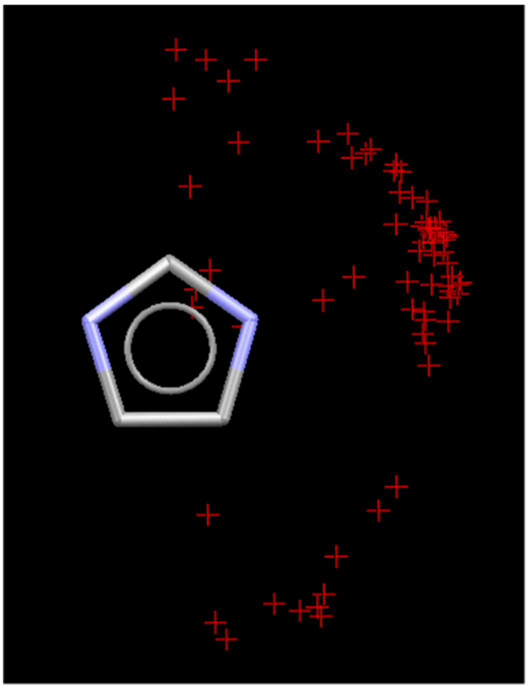
Tautomeric states are often uncertain in protein structures, whereas they are usually determined with confidence in CSD structures. For example, imidazol-4-yl (left below) and imidazol-5-yl (right below) can be distinguished in the CSD but not in the PDB:





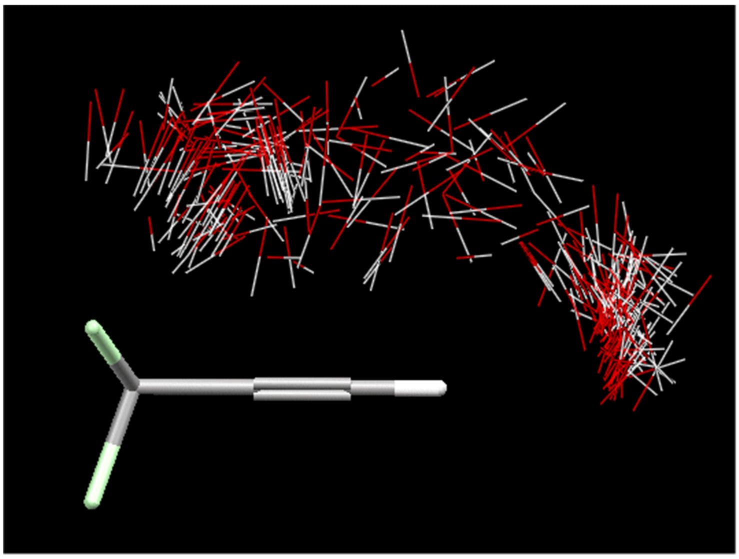
Consequently, IsoStar contains separate CSD-based plots for different tautomers but only a single, combined PDB-based plot. For this and other reasons, the symmetry of CSD and PDB scatterplots sometimes differs.

Imidazole (as in histidine) illustrates several of the problems with PDB-based scatterplots. There is a tautomeric ambiguity (4-yl or 5-yl), an uncertainty in ionisation state (neutral or cationic) and, at poor resolution, the crystallographer may have had difficulty in distinguishing ring nitrogens from ring carbons. This makes PDB-based plots for this ring particularly difficult to interpret. For example, in the plot of OH around imidazole (from histidine sidechains), it is not clear which OH groups are donating to ring nitrogens and which are accepting from them. There are also some short contacts on this plot between hydroxyl oxygens and the imidazole C2 atom; these may be CH...O hydrogen bonds or may be errors resulting from the imidazole ring being fitted to the electron density the wrong way round. The contacts shown in the scatterplot below are shorter than the sum of van der Waals radii - 0.2 Å.

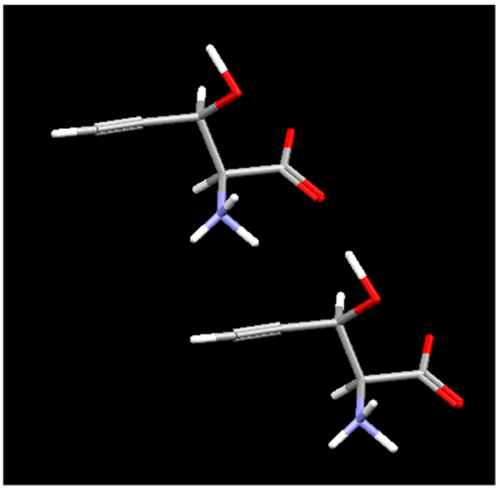


### 7.2.3 Biases in CSD-Based Scatterplots

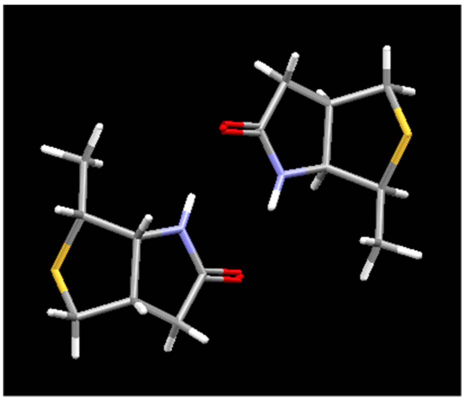
There are two reasons why a CSD-based scatterplot might be misleading: because of a common substitution pattern in the database or because of a frequent crystal-packing motif. The first is illustrated by the distribution of carbonyl groups around ethynyl. There is a surprising cluster of oxygens near the tetrahedral carbon:



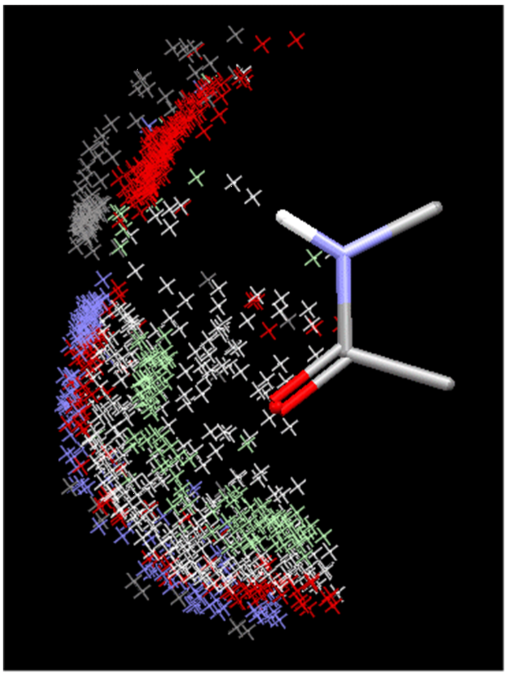
This is due to the fact that many of the CSD structures contain an OH on this carbon, i.e. C(OH)-CCH; for example, CSD entry APYNAC:



Secondly, CSD-based scatterplots may be misleading because of the existence of a common crystal-packing motif. This is well illustrated by ring systems that contain a cis-amide function, e.g. lactams and hydantoins. These often dimerise by hydrogen bonding, e.g.

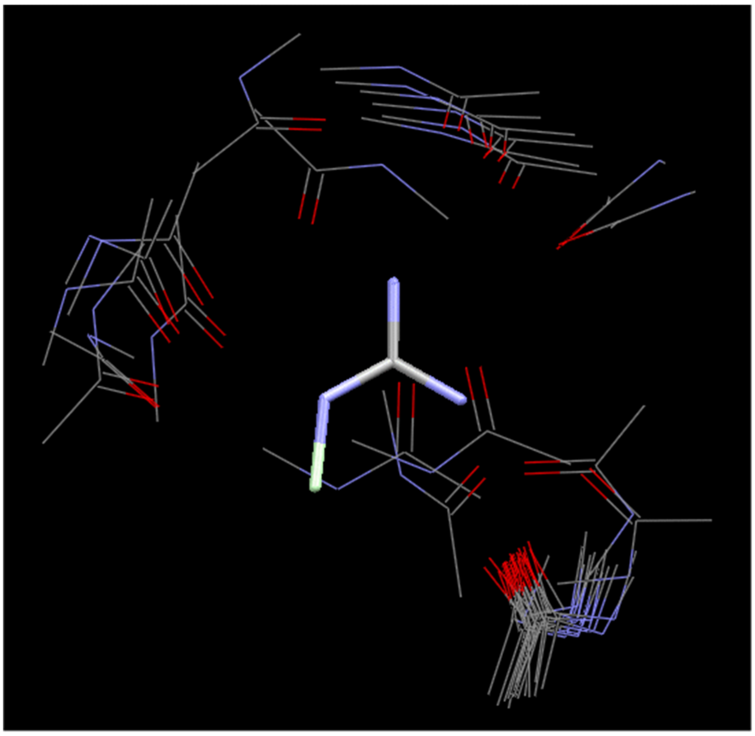


This results in strong clustering on the corresponding scatterplots. Thus, the distribution of the contact group any C,N,O,S or H around gamma lactams shows clusters of oxygen and nitrogen atoms near the N-H and C=O groups, respectively, of the lactam. Many (though by no means all) of these contacts arise from H-bonded dimers. This is shown in the image below (all contacts are shorter than the sum of van der Waals radii - 0.1 Å):



### 7.2.4 Biases in PDB-Based Scatterplots

An important source of bias is the number of closely related entries in the PDB. For example, the distribution of amide carbonyl groups around arginine guanidiniums shows a tight cluster of contacts between the two $-NH_2$ groups (all contacts are shorter than the sum of van der Waals radii - 0.1 Å):



On hyperlinking, however, it is found that most of these contacts come from a series of very closely related thermolysin complexes.

### 7.2.5 Biases in Central-Group Geometries

The central group geometry in a scatterplot is obtained by translating and rotating individual structures so that the central group moiety is in a standard position and orientation. A mean geometry is then obtained by averaging in Cartesian coordinate space. For a rigid group, this produces an average geometry very close to that which would be obtained if the averaging were done in internal coordinate (bond length, bond angle, torsion angle) space. For more flexible groups, however, the averaging of Cartesian coordinates results in a slight shrinkage of the bond lengths of the group, similar to that produced by the librational effect in crystallography. This has an insignificant effect on the IsoStar scatterplots. However, the central-group geometries in IsoStar should not be taken as the statistically optimum geometries of these groups.

## 7.3 Details of Molecular-Orbital Methodology

The ab initio package CADPAC 6.0 (Amos, CADPAC6.0: The Cambridge Analytical Derivatives Package, University of Cambridge, Cambridge, UK, 1996) is used to calculate the wave functions of individual monomers and to optimize their geometries at the 6-31G** basis set level. The charge distributions are corrected for electron correlation by the MP2 method. A distributed multipole analysis (DMA) is then carried out on the wavefunctions to obtain atomic multipoles up to hexadecapole.

Calculations of minimum energy orientations for dimers are carried out using the ORIENT 3.0 (Stone, Popelier & Wales, ORIENT 3.0: A Program for the Calculation of Electrostatic Interactions between Molecules, University of Cambridge, Cambridge, UK, 1994) program. This program can efficiently minimise the interaction energy of a given dimer starting orientation using a model intermolecular potential. The model potential describes the long-range electrostatic energy, using the DMA multipoles, by including all terms of the multipole expansion up to and including $r^{-5}$. Repulsion and dispersion are described by an empirical, isotropic exp-6 term. This potential, FIT (Coombes, Price, Willcock & Leslie, J. Phys. Chem., **100**, 7352, 1996), reproduces a wide range of polar organic crystal structures, and so will provide a reasonable first estimate of the intermolecular separations, but ignores any anisotropy in the atomic repulsive wall. The methodology described is adequate to compute long-range interactions and, therefore, results in well orientated monomer positions for energy minimized dimers. Since the method is fast, it can be used to scan and minimize quickly the full intermolecular hypersurface. Typically, starting with 100 - 500 initial orientations, the global and 5 - 20 local minima are obtained for any particular dimer.

Intermolecular interaction energies are then calculated for selected minimum-energy dimer orientations, using the Intermolecular Perturbation Theory (IMPT) method, as programmed in the CADPAC 6.0 package. This method is free of basis-set superposition error, a major problem within ab initio supermolecule calculations. It provides an estimate for each significant contribution to the total interaction energy, viz:

- The electrostatic energy ($E_{es}$) term describes the attractive or repulsive classical Coulombic interaction.
- The exchange-repulsion term ($E_{er}$) is the sum of (i) an energy lowering due to exchange of electrons of parallel spin between the molecules, and (ii) a repulsive term arising from the Pauli exclusion principle.
- The polarisation energy ($E_{pol}$) term accounts for the energy gain caused by the change of the intramolecular wave function of one molecule due to the presence of the undistorted charge distribution of the second molecule.

- The charge-transfer energy ($E_{ct}$) term is the attractive energy due to actual charge transfer between molecules.
- The dispersion energy ($E_{dis}$) is calculated at the second order double excitation level and is the result of instantaneous correlation of fluctuating electron density distributions.

Due to the limited quality of the empirical repulsion-dispersion parameters used in the ORIENT program, it is necessary to optimize the energy as a function of the distance between the model molecules, maintaining fixed angular orientations, using the IMPT method. The energies of the resulting distance-optimized dimers and the corresponding dimer orientations are stored in the library.

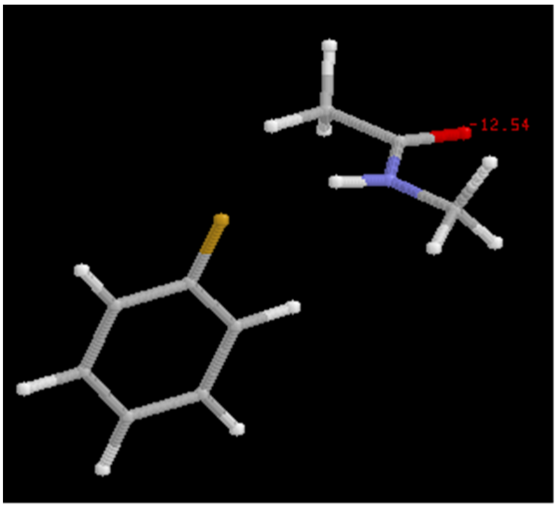
## 7.4 Limitations of Molecular-Orbital Results

Calculations are usually performed on the simplest possible model compounds. Consequently, steric effects are usually less important in the model systems than in the many highly substituted molecules found in the CSD and PDB.

As with all molecular orbital calculations, the accuracy of the results is limited by the basis set used and the methodology employed. However, comparisons between the method used in IsoStar and extended basis-set (triple zeta plus polarization) supermolecule calculations show an encouraging degree of concordance (Brode & Lommerse, unpublished work, 1996).

There is no way of ensuring that all minima in a potential energy surface are found. For any given model system, therefore, it is always possible that some important energy minima have been missed. In cases where this seems particularly likely, an appropriate warning message has been included in IsoStar.

By far the biggest limitation in the theoretical calculations is that they model an in vacuo situation. They are therefore unsuitable for indicating the gain or loss of free energy that occurs when a group is taken out of solution to form a nonbonded interaction to a protein. Also, weakly attractive interactions that occur in vacuo are much less likely to be found in condensed phases, where competing interactions are possible. For example, the calculated minimum-energy orientation of fluorobenzene - N-methylacetamide involves an N-H...F hydrogen bond (shown below). Such an interaction is extremely unlikely to occur in condensed phases, where better hydrogen-bond acceptors will successfully compete for the NH proton.



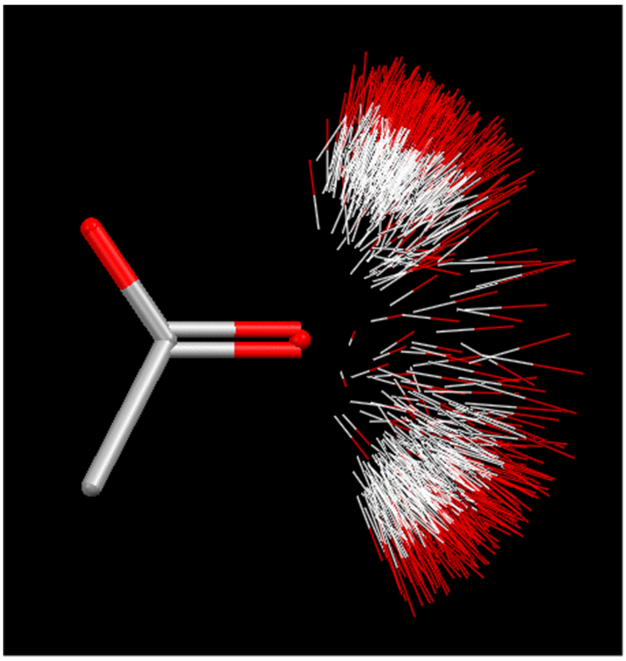
Calculated interaction energies are always much larger (i.e. more negative) for model systems in which one or both of the interacting species is an ion. Again, this is a consequence of the gas-phase nature of the calculations.

The value of the MO results is that they can provide independent evidence that the features seen on CSD and PDB scatterplots reflect the intrinsic preferences of the interacting species rather than, say, crystal-packing effects. A good example is provided by the phenomenon of H-bond lone-pair directionality to carboxylate oxygens.

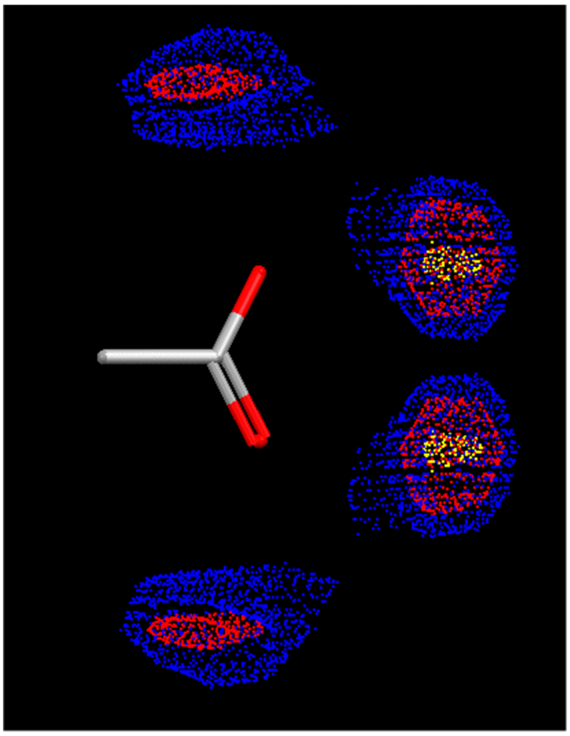
# 8 Example Results

## 8.1 Carboxylate ... OH

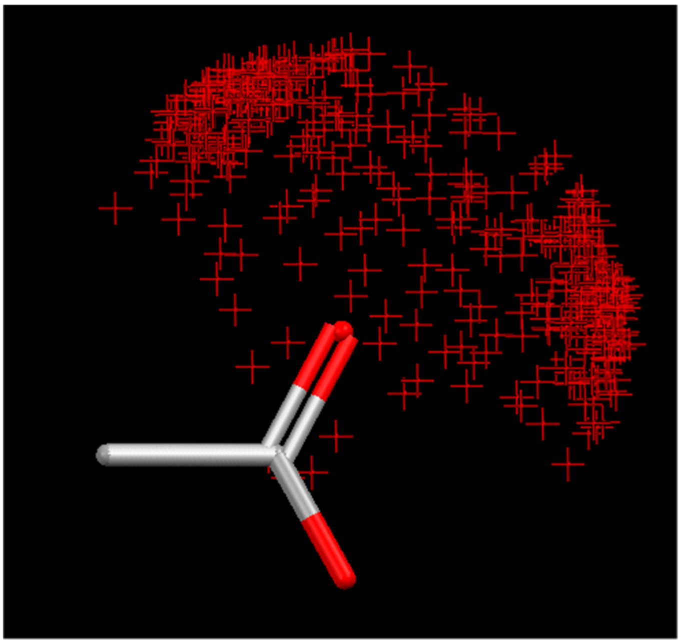
Many hydrogen-bonding groups are included in the IsoStar library. Inspection of the CSD-based plot of OH around carboxylate shows a distinct preference for H-bonding in the oxygen lone-pair directions. The preference is somewhat obscured when all the contact groups are displayed, but it becomes clear if the distance slider is manipulated so that only short contacts are shown (i.e. shorter than the sum of van der Waals radii - 0.9 Å):



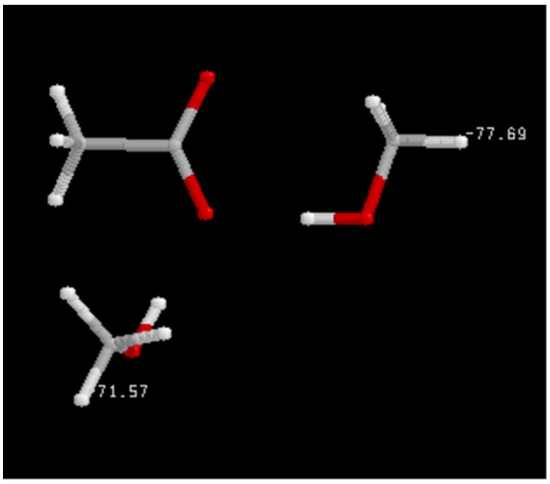
The preference for lone-pair directions becomes obvious if the scatterplot is converted to a contoured density surface:



The corresponding PDB-based scatterplot (below) shows similar features (contacts are shorter than the sum of van der Waals radii - 0.5 Å). Now, however, only the oxygen atoms of the OH groups are shown, as H atoms are rarely located in protein crystal structures. For the same reason, we cannot be sure that all the contacts on this plot are to ionised rather than neutral acid groups.

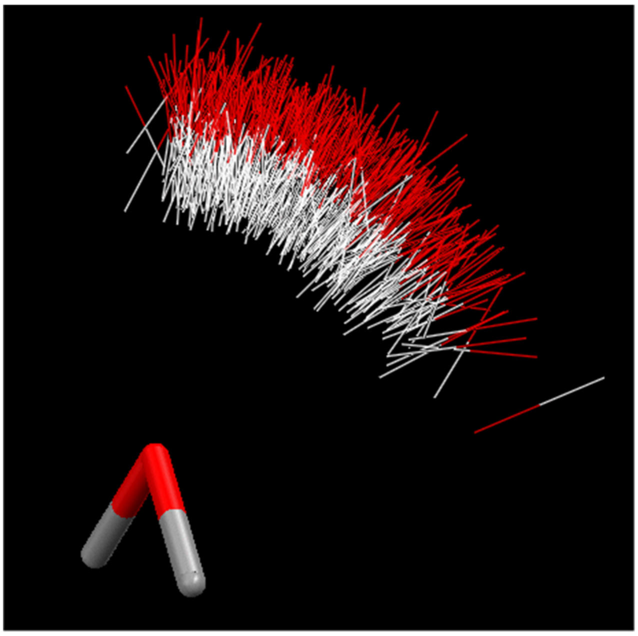


Theoretical IMPT calculations on the model system acetate - methanol also support the conclusion that hydrogen bonds to carboxylates form preferentially along the lone-pair directions. The two lowest minima occur in these positions:

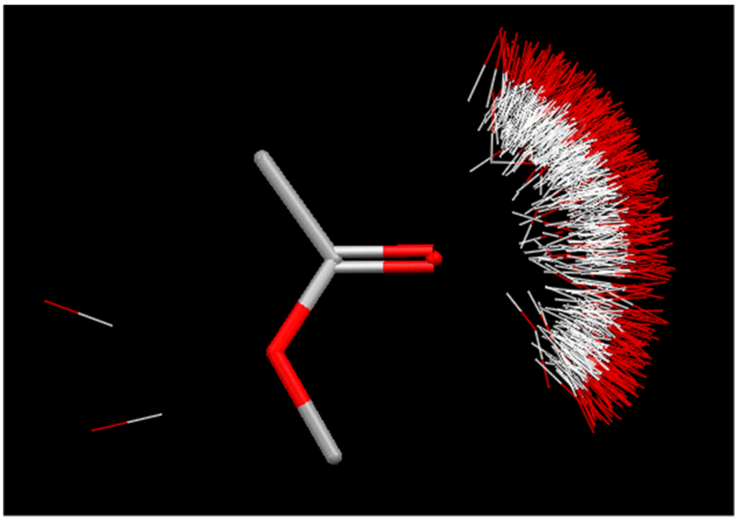


## 8.2 Aliphatic ether ... OH

Not all H-bond acceptors show lone-pair directionality. For example, the CSD-based plot of OH around aliphatic ether shows no particular clustering in the lone-pair (tetrahedral) directions (contacts shorter than the sum of van der Waals radii - 0.5 Å):

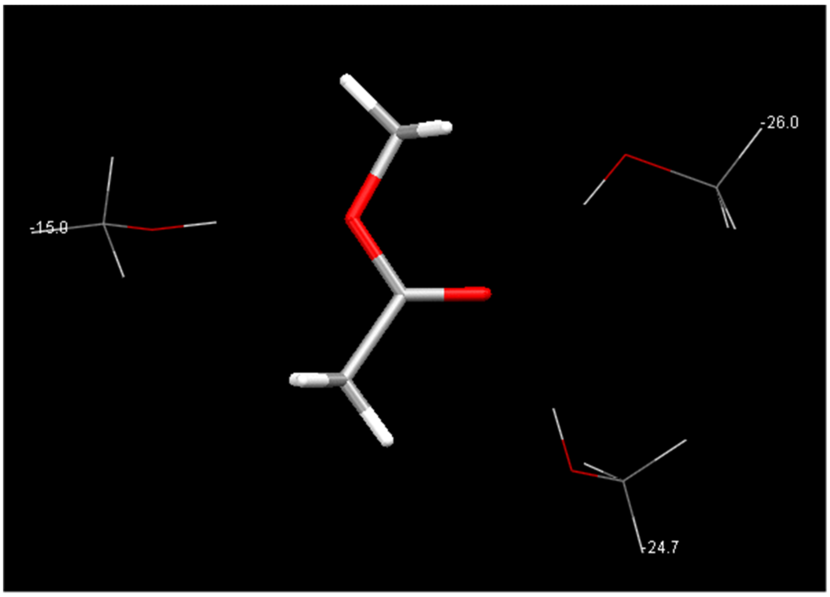


The CSD-based plot of OH around aliphatic ester has several interesting features (contacts shorter than the sum of van der Waals radii - 0.7 Å):

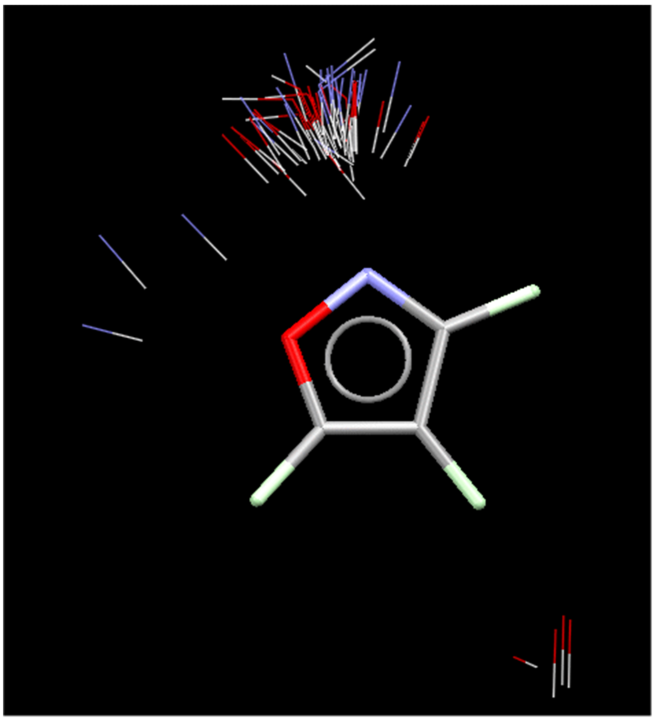


First, the carbonyl oxygen accepts H-bonds frequently but the C-O-C ester oxygen almost never accepts. Secondly, one of the carbonyl lone pair directions is preferred to the other (this becomes clear if the scatterplot is contoured).

The first observation is confirmed by IMPT calculations on the model system methyl acetate - methanol. However, the MO calculations predict the opposite lone-pair preference to that observed in crystal structures. This is probably because steric effects are usually more important in the CSD structures than in the very unhindered model system used for the calculations. The three lowest potential-energy minima of the methyl acetate - methanol complex are shown below:

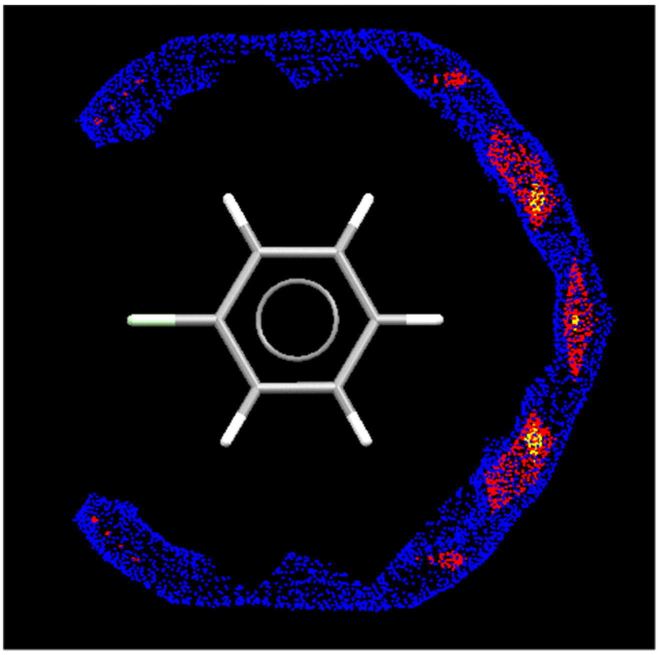


Other oxygens in conjugating environments are also poor acceptors. For example, the CSD-based plot of (N/O/S)H groups around isoxazole shows that hydrogen bonding almost always occurs to the ring nitrogen, not the ring oxygen (contacts shorter than the sum of van der Waals radii - 0.25 Å are shown):

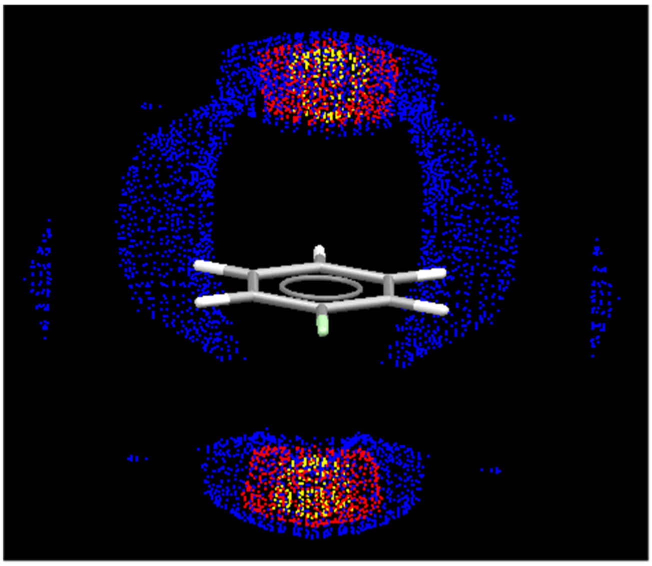


## 8.3 Phenyl ... C=O; phenyl ... CH(aliphatic)

One of the most striking features in IsoStar is the surprisingly strong directional preferences of contacts to hydrophobic groups. For example, the contoured distribution of carbonyl oxygens around phenyl shows an overwhelming preference for these atoms to occur around the edge of the phenyl ring:



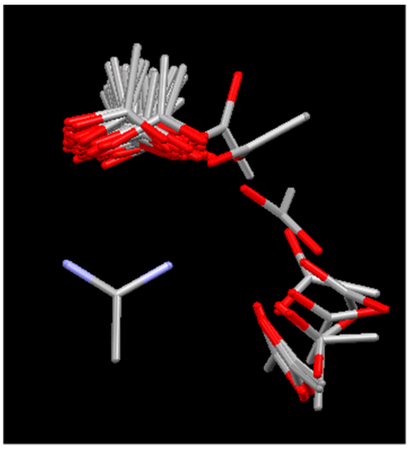
In contrast, the distribution of aliphatic C-H around phenyl shows a peak above the ring:



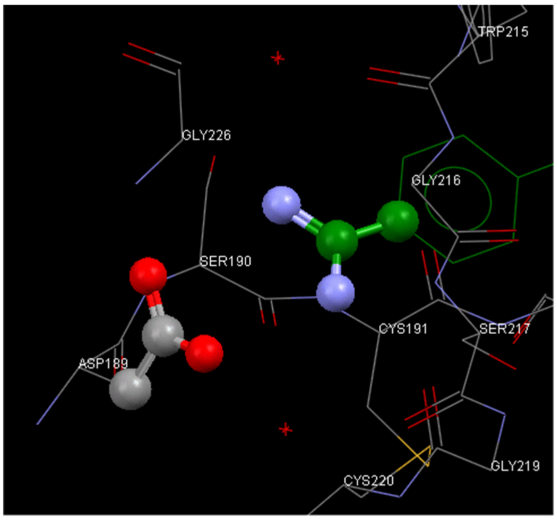
This is consistent with the high quadrupole moment of benzene.

## 8.4 Amidino ... Carboxylate

In the PDB-based plot of carboxylate groups around amidine, there are some short contacts above the amidine carbon (contacts shorter than the sum of van der Waals radii are shown):



Use of the hyperlinking facility shows that one of these occurs in the structure 1PPC (a trypsin complex) and involves the close approach of the backbone carboxylate of Asp189:

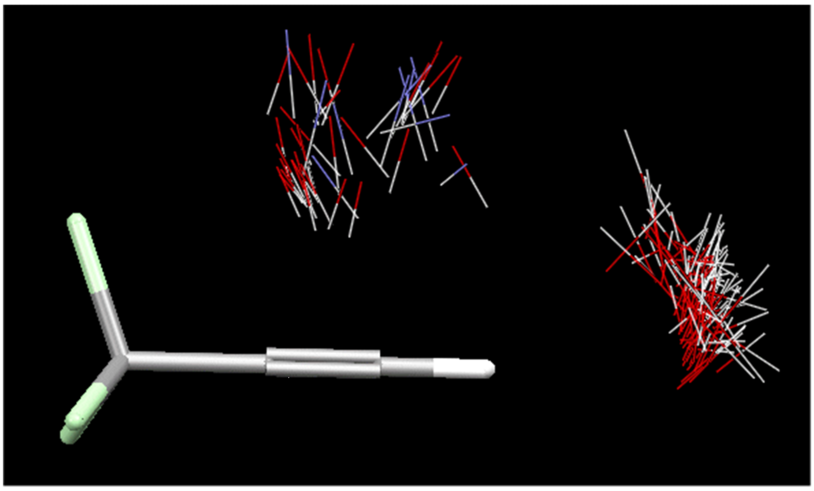


## 8.5 Ethynyl ... (O,N,S)H

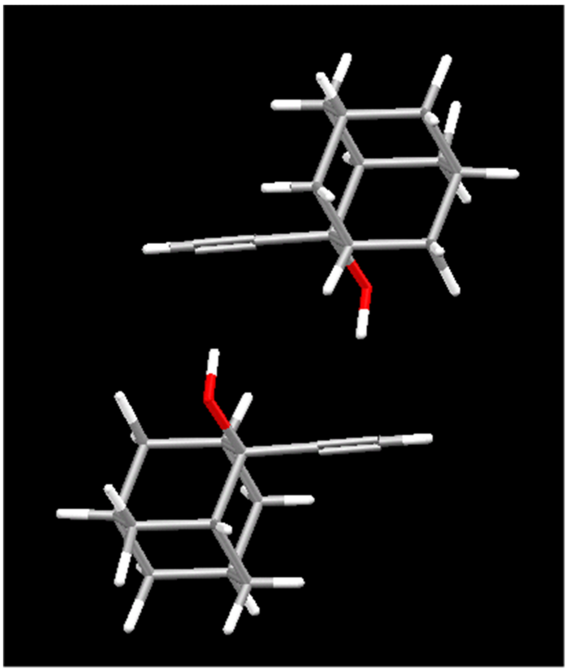
The CSD-based plot of (O/N/S)H groups around ethynyl shows two interesting features.

There are some hydrogen bonds to the acetylene pi system.

There are several CH...O interactions, where the relatively acidic acetylenic hydrogen acts as an H-bond donor to hydroxyl oxygen (contacts shorter than the sum of van der Waals radii - 0.2 Å are shown).



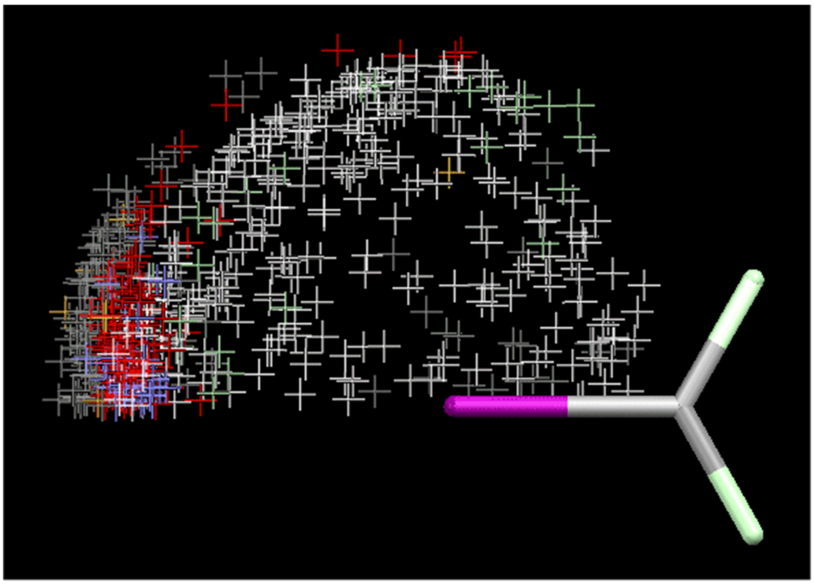
Again, the hyperlink function can be used to inspect some of these. For example, one of the O-H...pi H-bonds occurs in the CSD structure BETXAZ:



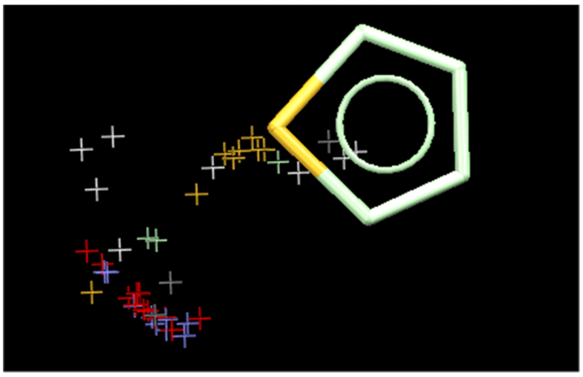
## 8.6 Aromatic I ... any C,N,O,S or H

A particularly good way of spotting novel interactions is to look at scatterplots involving the contact group any C,N,O,S or H.

For example, the CSD-based plot for any C,N,O,S or H around aromatic iodine shows several C-I..O contacts along the extension of the C-I bond (Lommerse et al, J. Am. Chem. Soc., **118**, 3108, 1996). Some of these are extremely short (contacts less than the sum of van der Waals radii are shown below).



Short, highly directional sulfur...oxygen and sulfur...nitrogen contacts (Burling & Goldstein, J. Am. Chem. Soc., **114**, 2313, 1992) are found in the plot of any C,N,O,S or H around sulfur heterocycles (contacts shorter than the sum of van der Waals radii - 0.3 Å are shown below).



# 9 IsoGen: Quick Summary

Please note that IsoGen is available on Linux platforms only.

To generate a scatterplot from the CSD, the following simple recipe will usually suffice:

1. Perform a ConQuest nonbonded search for the contact of interest.
2. Save the 3D parameters using the **Export Parameters and Data** option from the **File** menu. This will result in 3 files, `<search_name>.tab`, `<search_name>.fgn` and `<search_name>.fgd`.
3. Then you will need to save the fractional fragments using the **COORD: CSD Coordinate file** option in the **Export Entries As...** part of the **File** menu.
4. After the search has run, browse to the directory where the files are stored and type:

   ```
   isogen <search_name>
   ```

   at the Unix prompt, where `<search_name>` is the name used for the ConQuest search.
5. Pick **Run** from the IsoGen top-level menu. If all is well, the scatterplot will be calculated and saved as a file with the extension `<search_name>.istr`.

# 10 IsoGen: Quick Summary

IsoGen is the program that was used to calculate all the scatterplots in the IsoStar database. It is released to users so that they can calculate their own customised scatterplots, e.g. for groups which are not included in the standard IsoStar library. Plots calculated by users with IsoGen can be viewed in the IsoStar graphical interface. IsoGen is available for Linux only.

Two restrictions should be noted. First, IsoGen can only be used to calculate scatterplots from the Cambridge Structural Database (CSD), not from the Protein Data Bank (PDB). Secondly, the external scaling method of contouring density surfaces is unavailable for user-generated scatterplots.

There are three main stages in producing a scatterplot with IsoGen:

1. A ConQuest search must first be performed to find structures in the CSD that contain the nonbonded contact of interest. The crystallographic fractional coordinates of the hits must be saved as a `.cor` file.

2. IsoGen must then be run, using the `.cor`, `.tab`, `.fgn` and `.fgd` files from the ConQuest search as input. Usually, all the default settings of the program can be used - all the user has to do is identify the relevant search name and then pick **Run** from the top level IsoGen menu. The program will read the nonbonded fragments from the `.cor` file, superimpose the central group, detect and eliminate conformational outliers, identify and deal with symmetry, and calculate the scatterplot.

3. The plot may then be visualised and manipulated in the normal way, using the IsoStar graphical interface.

A Custom Plots directory is included in the IsoStar directory structure for storage of user-generated scatterplots. Assuming file protections are set appropriately, plots placed in this directory are accessible to all IsoStar users within an organisation. For more details, click on **Custom Plots** (at the bottom of the IsoStar central-group menu) and look at the ReadMe file.

# 11 Introduction to IsoGen

It is assumed that the reader has some knowledge of ConQuest - in particular, how to build fragments and do simple substructure searches. If information is required about these basic features, the reader is referred to the relevant documentation: [www.ccdc.cam.ac.uk/solutions/csd-core/components/conquest/]https://www.ccdc.cam.ac.uk/solutions/csd-core/components/conquest/.

Alternatively, the basics of searching using ConQuest are covered in tutorials provided at the end of the documentation. ConQuest Tutorial 5 covers nonbonded contact searching.

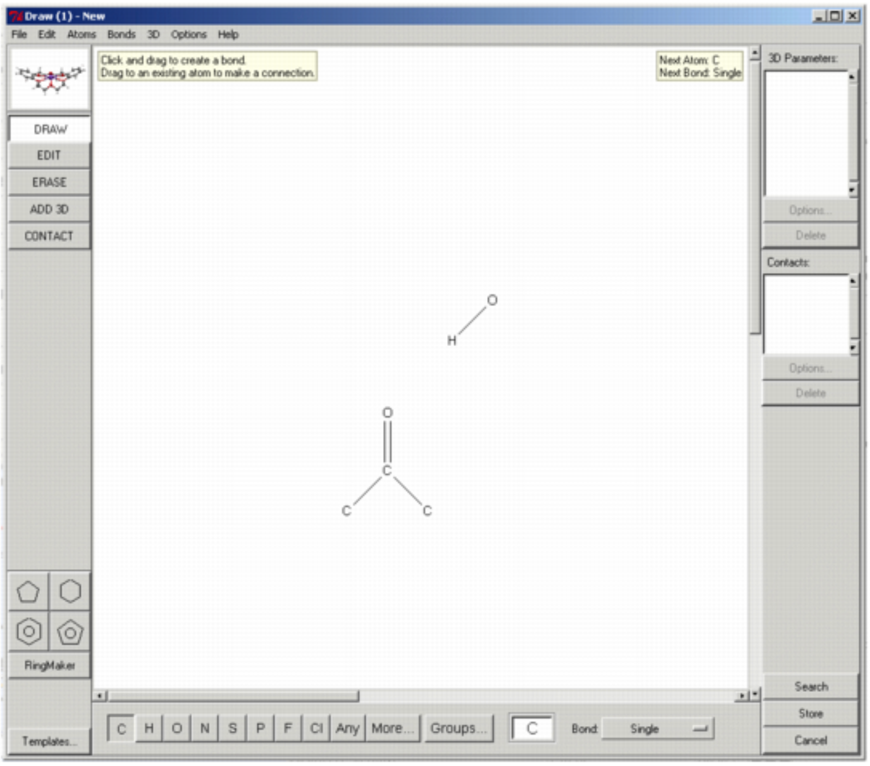
The essential features of nonbonded searching in ConQuest will be illustrated with an example, viz. a very simple search for close nonbonded interactions between ketone oxygens and hydroxyl hydrogens (see A Simple Nonbonded Search).

Some advanced search options are also covered (see Advanced Nonbonded Search Options).

1. IsoGen must then be run, using the .cor, .tab, .fgn and .fgd files from the ConQuest search as input. Usually, all the default settings of the program can be used - all the user has to do is identify the relevant search name and then pick **Run** from the top level IsoGen menu. The program will read the nonbonded

fragments from the .cor file, superimpose the central group, detect and eliminate conformational outliers, identify and deal with symmetry, and calculate the scatterplot.

2. The first step is to draw the two groups using the ConQuest sketcher, accessed from the **Draw** button in the main ConQuest interface:

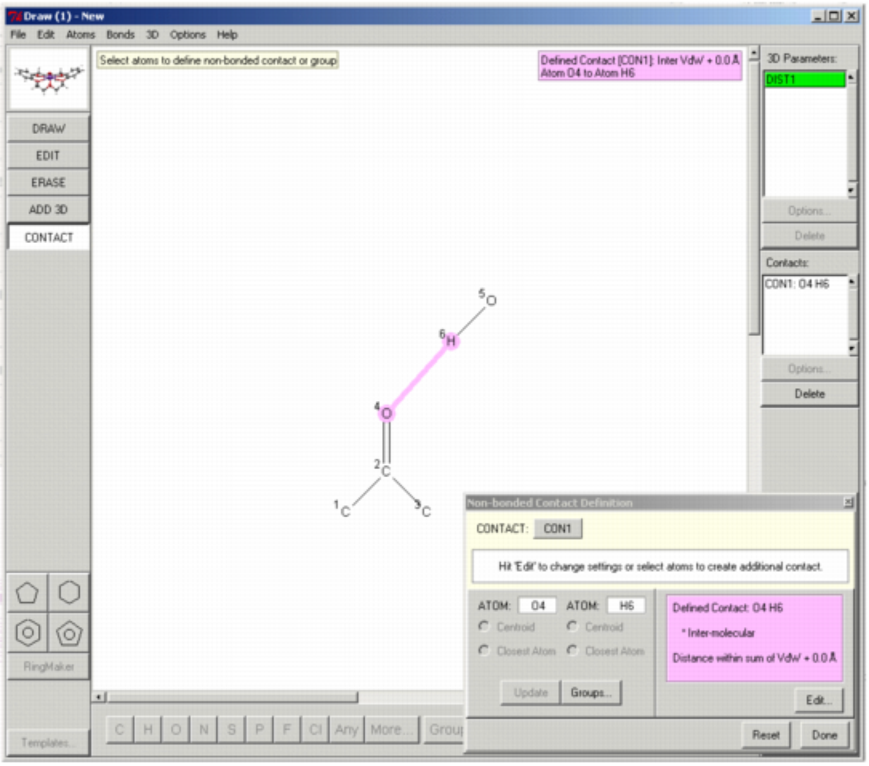


Click and drag to create a bond.
Drag to an existing atom to make a connection.

Next Atom: C
Next Bond: Single

C Bond: Single

3. The next step is to define a contact between the oxygen atom of the ketone group and the hydrogen of the O-H group using the **CONTACT** button in the sketcher:



Select atoms to define non-bonded contact or group

Defined Contact [CON1]: Inter VdW + 0.0 Å
Atom O4 to Atom H6

CON1: O4 H6

CONTACT: CON1

Hit 'Edit' to change settings or select atoms to create additional contact.

ATOM: O4 ATOM: H6
○ Centroid ○ Centroid
○ Closest Atom ○ Closest Atom

Defined Contact: O4 H6
* Inter-molecular
Distance within sum of VdW + 0.0 Å

4. Once the contact has been defined, the search can be started by clicking on **Search**.

5. It is important that the **Normalise terminal H positions** tickbox is turned on under the **Advanced Options** tab in the **Search Setup** window.

6. The search can be started by clicking **Start Search**.

## 11.1 Advanced Nonbonded Search Options

### 11.1.1 Defining Distance Ranges

When a contact is defined in ConQuest, the default distance of less than the sum of Van der Waals radii is used unless the user inputs alternative values.

An alternative contact distance range can be input by clicking on **Edit** in the **Non-bonded contact definition** window and activating the **Distance range** radio button. The desired distance range can then be input, e.g. if two distances are typed in: 2.0 to 3.0, then all contacts in this range will be found.

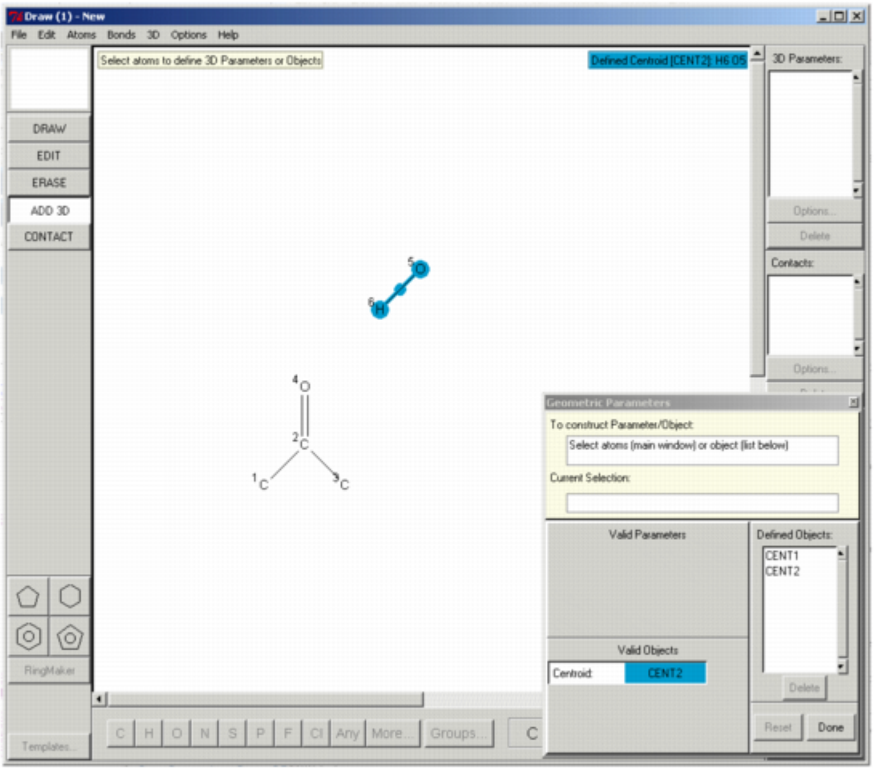
## 11.1.2 Searching for Contacts between Groups Rather than Atoms

In the above example, a search was performed for contacts between the oxygen atoms of ketone groups and the hydrogen atoms of hydroxyl groups. An alternative is to search for contacts between any atom of the ketone fragment and any atom of the hydroxyl group, or between the centroids of the ketone and hydroxyl groups.

To do this:

1. Click on the **ADD 3D** button, clicking on each atom of the fragment in turn and then selecting the **Define...** button next to **Centroid** on the **Geometric Parameters** window.

2. Then, do the same for the hydroxyl group:



Select atoms to define 3D Parameters or Objects

Defined Centroid [CENT2]: H6 O5

To construct Parameter/Object

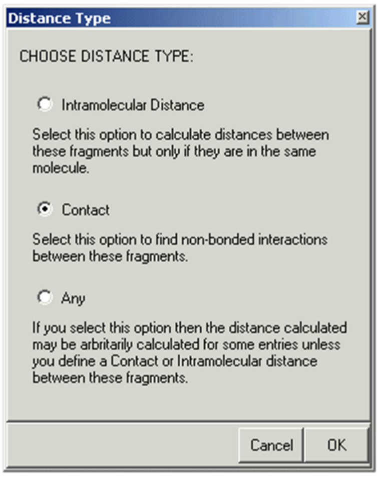
Select atoms (main window) or object (list below)

Current Selection:

Centroid: CENT2

3. Leave the **Geometric Parameters** window open and select both **CENT1** and **CENT2** in the **Defined Objects:** window. Then click on **Define...** next to **Distance:**.

4. A **Distance Type** window will pop up:



Distance Type

CHOOSE DISTANCE TYPE:

○ Intramolecular Distance

Select this option to calculate distances between these fragments but only if they are in the same molecule.

● Contact

Select this option to find non-bonded interactions between these fragments.

○ Any

If you select this option then the distance calculated may be arbitrarily calculated for some entries unless you define a Contact or Intramolecular distance between these fragments.

5. Select the radio button next to **Contact** and click **OK**.

6. In the resultant window, click on the **Define** button. This will allow you to input a distance range for the contact e.g. 0.0 to 3.0. Once this has been done, the search can be run.

### 11.1.3 Correcting Hydrogen Atom Positions

Hydrogen X-H valence bond distances are usually too short when measured by X-ray diffraction. To correct for this systematic error, ensure the **Normalise terminal H positions** tickbox is turned on under the **Advanced Options** tab in the Search Setup window.

# 12 Isogen: Basic Features

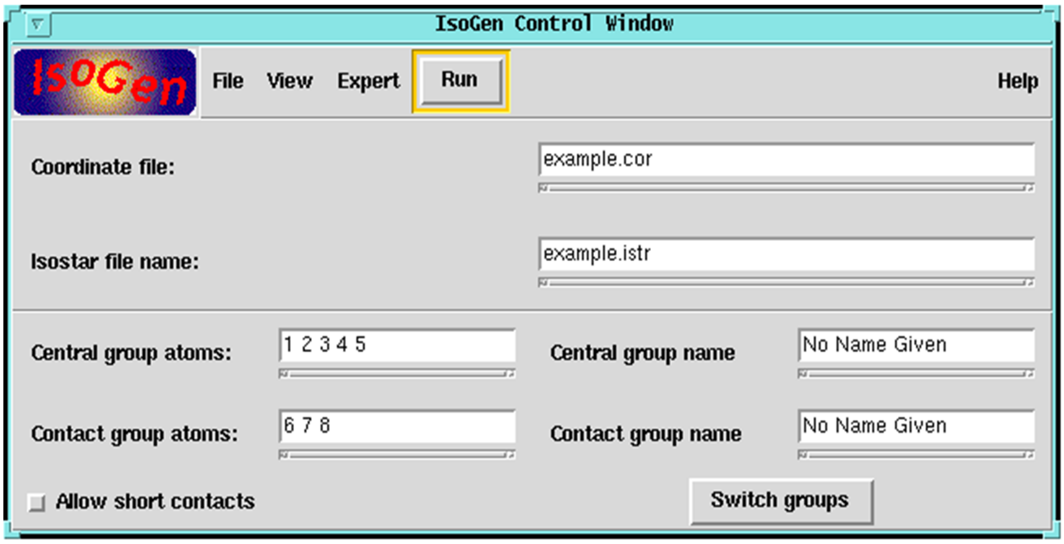
## 12.1 Starting IsoGen

Once a ConQuest nonbonded search has been run, and assuming IsoStar is installed, the IsoGen program can be started on Linux platforms only by typing:

```
isogen <search_name>
```

## 12.2 Basic Options in the IsoGen Interface

On starting IsoGen, the user is presented with a graphical interface from which to produce scatterplots:



Coordinate file: example.cor

Isostar file name: example.istr

Central group atoms: 1 2 3 4 5

Central group name No Name Given

Contact group atoms: 6 7 8

Contact group name No Name Given

□ Allow short contacts

The IsoGen interface, in its Basic mode, contains a number of items, the functions of which are summarised below:

- Coordinate file*: The name of the .cor file produced by a ConQuest search.
- IsoStar file name*: (see Isostar file name)
- Central group atoms*: (see Central group atoms)
- Contact group atoms*: (see Contact group atoms)
- Central group name: The name that will be used for the central group in the IsoStar graphical interface. By default, this is set to No Name Given.
- Contact group name: The name that will be used for the contact group in the IsoStar graphical interface. By default, this is set to No Name Given.

Those superscripted by a star are mandatory for successful completion of IsoGen, but are, in most cases, read in automatically from the ConQuest output files. Usually, the defaults are adequate to produce an acceptable scatterplot. In such cases, all the user need do is click on **Run** in the top level menu. After the plot is calculated, the user will be given the option of viewing it in the usual way (i.e. using the IsoStar graphical interface).

To get help when using IsoGen, click with the right-hand mouse button on the item for which information is required.

### 12.2.1 Isostar file name

The filename to which the results of the superposition (i.e. the scatterplot) are written. This should have the extension `.istr` (the `.istr` format is closely related to the `.mol2` format of Certara (formerly Tripos Inc.).

### 12.2.2 Central group atoms

The atoms of the central group (i.e. the fragment which will be superimposed). The atom numbers are taken from the ConQuest search. There must be at least three non-linear atoms in the central group.

### 12.2.3 Contact group atoms

The atoms of the contact group (i.e. the fragment whose distribution around the central group will be shown). The atom numbers are taken from the ConQuest search.

### 12.2.4 Allow short contacts

If this button is switched on, all contacts will be included in the scatterplot, irrespective of their distance. If the button is switched off, very short contacts (viz. more than 1.3 Å shorter than the sum of van der Waals radii) will be excluded.

### 12.2.5 Switch groups

This command transposes the contact group and the central group. It is only enabled when there are sufficient atoms in the contact group for superposition (i.e. at least three).

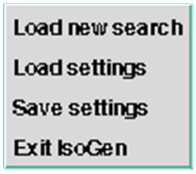
## 12.3 IsoGen Top Level Menu Options

A number of top-level menu options are provided:

- **File** (see File)
- **View** (see View)

These are only activated when the program has sufficient information for them to be relevant. For example, the **Run** button will only be active if a valid `.cor` file is named in the **Coordinate file** box.

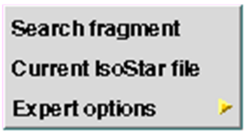
### 12.3.1 File



The **File** menu displays a sub-menu with the following options:

- **Load new search**: This command allows the user to load a new set of files from a ConQuest nonbonded search. The program will automatically read the appropriate ConQuest output files and generate a new set of defaults, e.g. for central-group and contact-group atom numbers. If a file called `<search_name>.rules` exists, then the program will prompt the user as to whether it should use this file, or re-generate the settings anew (see next item for more information on settings).
- **Load settings**: When a scatterplot is successfully calculated by IsoGen, details of the scatterplot settings (e.g. 3D symmetry, etc.) can be saved in a file called `<search_name>.rules`. The **Load Settings** command allows the user to load a settings (i.e. .rules) file saved in a previous run of IsoGen.
- **Save settings**: This command allows the user to save the current scatterplot settings (e.g. containing details of 3D symmetry, etc.) in a settings file, which should have the extension `.rules`.
- **Exit IsoGen**: This command exits IsoGen.

### 12.3.2 View



The **View** menu displays a sub-menu with the following options:

- **Search fragment:** This command allows the user to view the search query, as drawn in ConQuest, and with ConQuest atom numbers displayed. This option will be greyed out if there are no 2D coordinates in the ConQuest output files.
- **Current IsoStar file**: This command allows the user to view the current scatterplot. It is greyed out if no `.istr` file exists.

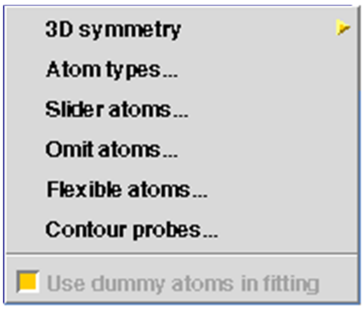
- **Expert options:** This command gives the user access to a number of expert options (see Isogen: Expert Features) that are available in IsoGen. Most of these options can also be accessed through the top-level menu item **Expert**.

- **Expert:** Gives access to various expert options (see Isogen: Expert Features).

- **Run:** Initiates calculation of the scatterplot. In most cases, the IsoGen program defaults are adequate to produce an acceptable scatterplot. In such cases, clicking on **Run** in the top level menu is all the user need do once the IsoGen interface has been opened. After the plot has been calculated, the user will be given the option of viewing it in the usual way (i.e. using the IsoStar graphical interface).

- **Help:** Displays a sub-menu with the following options:
  - **Getting help:** Displays a pop-up window which explains the ways of getting help in IsoGen, i.e. via the main **Help** document or via context-dependent help. The context-dependent help facility can be accessed by clicking with the right-hand mouse button on any option in the IsoGen interface. This opens a pop-up which gives a brief reminder of the function of the chosen option.
  - **About IsoGen:** Displays a pop-up window which explains what IsoGen is, who distributes it, and who should be contacted in case of difficulty.

# 13 Isogen: Expert Features



A number of options are available under the top-level **Expert** menu option. Their functions and use are described in the following sections:

- 3D Symmetry (see 3D Symmetry).

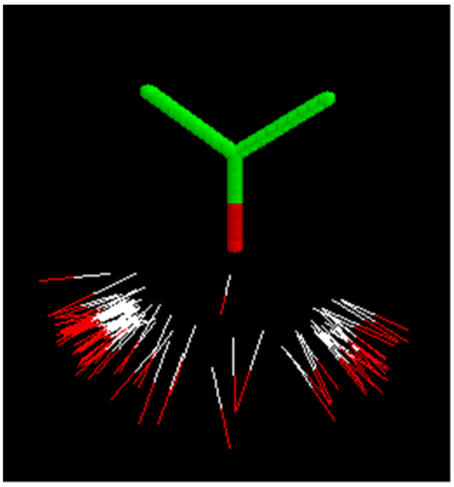
- Atom types (see Atom types).

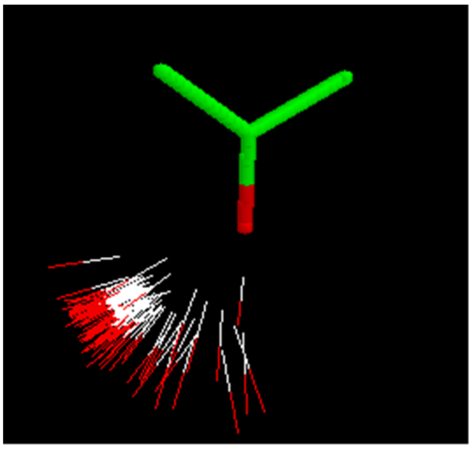
- Slider atoms (see Slider atoms).
- Omit atoms (see Omit atoms).
- Flexible atoms (see Flexible atoms).
- Contour probes (see Contour probes).
- Use dummy atoms in fitting (see Use dummy atoms in fitting).

## 13.1 3D Symmetry

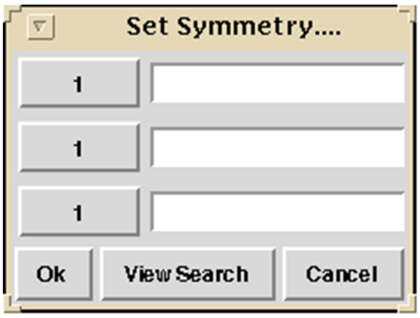
The 3D symmetry of the central group is used by the program for two reasons:

- To calculate an average geometry for the central group.
- To reduce the scatterplot so that all contact groups lie in one asymmetric unit of the central group's point group. This is illustrated by the following raw and symmetry reduced plots; in the latter, all contacts have been reflected into one quadrant, since the central group has mm symmetry.





The 3D symmetry menu option offers three possible selections: **Derive, Define,** and **Do not use.** The default option is **Derive**. This makes the program deduce the 3D symmetry automatically from the ConQuest output, a procedure which usually gives satisfactory results. **Do not use** instructs the program to produce a raw plot, i.e. with no determination of 3D symmetry. As the name suggests, the remaining option, **Define**, allows the user to specify explicitly a particular symmetry. Selecting this option causes the following window to be opened:



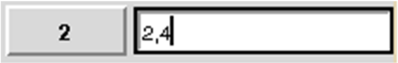
This window enables the user to specify explicitly up to three symmetry elements. The required element is selected by clicking on one of the left-hand buttons and choosing from the resulting list of elements. The direction in which the symmetry element lies is then typed into the corresponding white box. Directions are defined using central-group atom-atom vectors or centroid-atom vectors, see Appendix 1: Vector Direction Syntax. (Note that the central-group atom numbers may be displayed by clicking on **View Search**.) It is possible to specify sums and cross products of bond vectors as well as bond vectors themselves: this is most easily illustrated with examples.

### 13.1.1 Defining Symmetry, Example 1: Ketone

The aliphatic ketone fragment, by default, has mm2 symmetry. The two-fold axis lies along the C=O bond. A mirror lies in the plane formed by the four atoms. This can also be regarded as a two-fold rotation-inversion axis along any vector perpendicular to the plane. Since symmetry definition in IsoGen must be done with respect to a direction, not a plane, we must regard mirror planes in this way. The second mirror plane contains the C=O bond, and is perpendicular to the first.

Only the first two symmetry elements are needed to describe the full point group, since the third is generated by the program from the first two. The syntax for vector specification is provided later (see Appendix 1: Vector Direction Syntax).

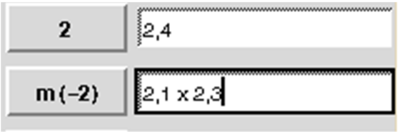
The direction of the two-fold axis is along the C=O bond and so is easily specified:



2,4

This defines a two-fold axis along the bond between atoms 2 and 4 (the C=O bond in this case).

The first mirror plane is specified using the cross product of the two C-C bonds (the bonds between atoms 2 and 1, and 2 and 3):



| | |
|---|---|
| 2 | 2,4 |
| m(-2) | 2,1 x 2,3 |

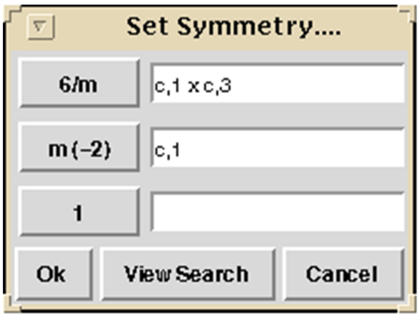
Since the third symmetry element is automatically generated from these two, it does not need to be specified explicitly. Thus, the above input is adequate to describe the mm2 point group.

### 13.1.2 Defining Symmetry, Example 2: Benzene

The symmetry of benzene can be described in the following terms:

- A mirror plane passing through all atoms (mirror 1).
- A six-fold axis along a direction passing through the molecular centroid and perpendicular to mirror plane 1. Note that, in axial terms, mirror 1 can be regarded as lying along the same axis as the six-fold, and so one can describe this axis as having 6/m symmetry.
- Mirror planes perpendicular to the centroid - carbon atom vectors, passing through the centroid.

Assuming the atoms of the benzene ring are numbered 1 to 6, the full point-group symmetry can be specified as:



Set Symmetry....

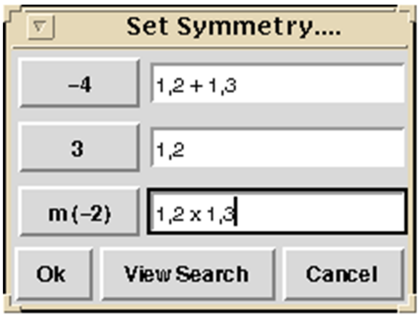
| | |
|---|---|
| 6/m | c,1 x c,3 |
| m(-2) | c,1 |
| 1 | |

Note here that the c is used to specify the centroid of the benzene fragment.

### 13.1.3 Defining Symmetry, Example 3: Sulfate

A sulfate anion has -43m symmetry. The -4 axes lie in the directions of the bisectors of the O=S=O angles, the three-fold axes lie in the direction of the S=O bond vectors, the mirror plane axes are perpendicular to the O=S=O planes.

To specify such symmetry, three elements are required: a -4 axis, a three-fold axis and a mirror plane. For example (and assuming the sulfur is atom number 1 and the oxygens atom numbers 2 to 5):



-4 1,2 + 1,3

3 1,2

m(-2) 1,2 x 1,3

The symbol + indicates a vector sum (in this case, the sum of 1,2 and 1,3).

## 13.2 Atom types

Each central-group atom is assigned an atom type. The atom types used are taken from the program SYBYL-X [Certara, formerly Tripos Inc.] and are summarised below:

H
    Hydrogen
C.3, C.2, C.1
    sp$^3$, sp$^2$, sp hybridised carbon, respectively
C.ar
    aromatic carbon
C.cat
    cationic (e.g. guanidinium) carbon
N.3, N.2
    sp$^3$ (pyramidal) and sp$^2$ hybridised nitrogen, respectively
N.4
    tetrahedral quaternary nitrogen
N.ar
    aromatic nitrogen
N.am
    amide nitrogen
N.pl3
    planar trigonal nitrogen
O.3, O.2
    sp$^3$, sp$^2$ hybridised oxygen, respectively
O.co2
    carboxylate/phosphate oxygen
F, Cl, Br, I
    fluorine, chlorine, bromine, iodine
P.3
    phosphorus
S.3, S.2
    sp$^3$ and sp$^2$ hybridised sulphur, respectively

S.o, S.o2

sulphoxide and sulphone sulphur, respectively

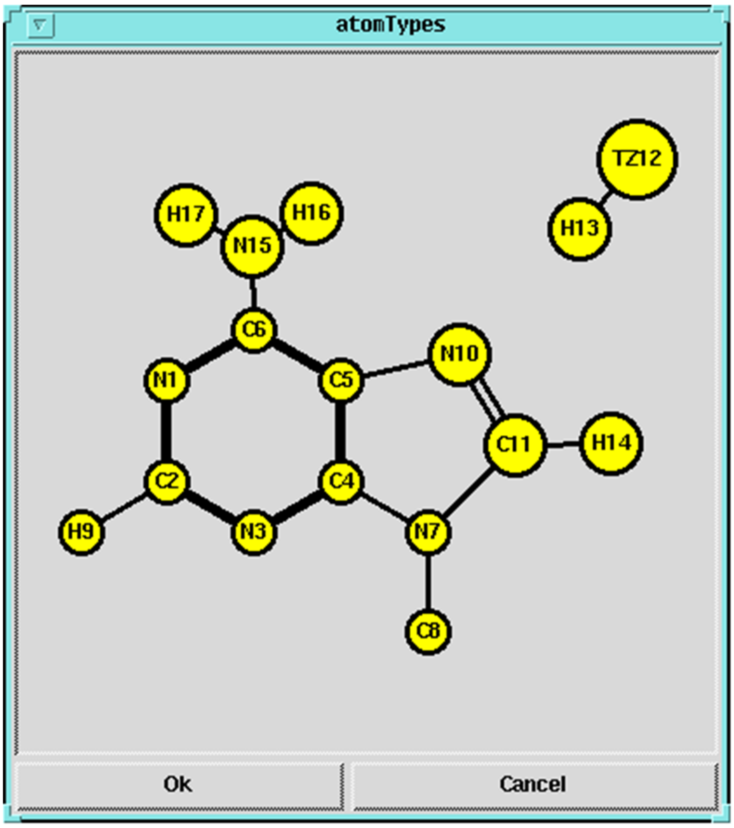
Du

dummy atom - used, e.g., for variable atom types, such as the ConQuest X (any atom) type

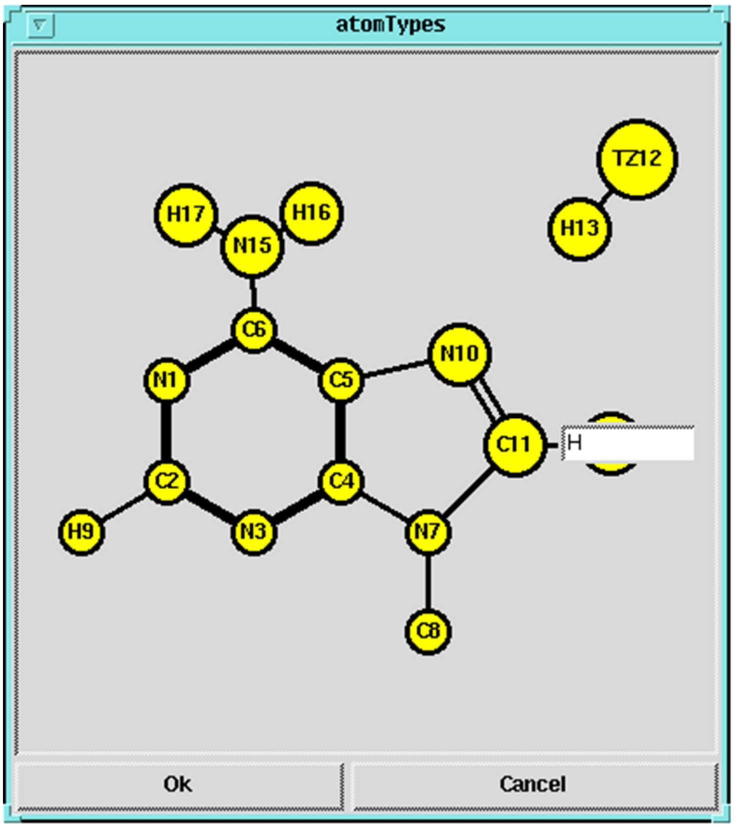
If the atom types assigned to central-group atoms are incorrect, it may affect the identification of topologically equivalent atoms (e.g. the two oxygen atoms of the carboxylate ion) and hence the determination of 3D symmetry.

By default, atom types are derived by IsoGen from the ConQuest output, usually with sufficient accuracy to avoid problems in detection of topological and 3D symmetry. When problems do occur, the user can edit atom types by selecting **Atom types...** in the **Expert** menu.

This creates a pop-up window, a typical example of which is:



Clicking on an atom then pops up a small entry box that enables the user to edit the atom type:

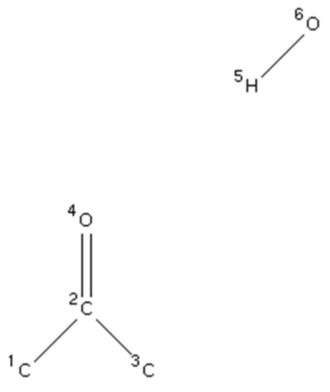


For contact-group atoms, the atom type ? is also allowed in IsoGen. This makes the program guess the atom type for each individual contact atom: an atom type is assigned which gives the atom the correct element symbol (although not necessarily the correct hybridisation state).

## 13.3 Slider atoms

A distance slider is provided in the IsoStar graphical interface. This allows users to remove from a scatterplot display all contact groups outside a specified corrected-distance range (the corrected distance between two atoms is their interatomic distance minus the sum of their van der Waals radii). Only central- and contact-group atoms defined as slider atoms are taken into account when IsoStar determines whether a given contact group should be displayed or not, given a particular setting of the distance slider.

The operation of the slider, given a particular choice of slider atoms, is best illustrated by an example. Consider the following ConQuest query:



Suppose that the first contact found in the ConQuest search had the following interatomic distances:

O4...H5 = 1.892 Å, O4...O6 = 2.709 Å, C2...H5 = 2.896 Å, C2...O6 = 3.601 Å.

Each of these values may be converted to a corrected distance by subtracting the relevant van der Waals radii. For example, the vdw radii of O and H are 1.52 and 1.20 Å, respectively, so the corrected O4...H5 distance = 1.892-1.52-1.20 = -0.828 Å. In this way, we get corrected distances of:

O4...H5 = -0.828 Å, O4...O6 = -0.331 Å, C2...H5 = -0.004 Å, C2...O6 = 0.381 Å.

Now suppose that C2, O4, H6 and O6 were all slider atoms and the distance-slider range in IsoStar was set to display all contacts between -1.0 and -0.5. In this case, the contact described above would be displayed, because the shortest corrected distance between a central-group slider atom and a contact-group slider atom (i.e. -0.828) falls in the range -1.0 to -0.5. However, if only C2, O4 and O6 were slider atoms, the contact would not be displayed, because the shortest distance between slider atoms (-0.331) would now fall outside the specified range.

By default, any atom utilised in a ConQuest CONTACT search is designated as a slider atom. The default settings can be changed by means of the **Slider atoms** option. This opens a pop-up window containing a display of the ConQuest search query; the user may then click on the desired slider atoms.

## 13.4 Omit atoms

When a scatterplot is generated, the various central-group fragments found in the ConQuest search must be superimposed and then averaged. Sometimes, a central group will contain a conformationally flexible moiety, such as a rotationally mobile methyl group. Inclusion of these moieties in the superposition process can result in distorted average central-group geometries.

In such cases, the user may wish to omit the atoms whose positions are conformationally variable. This may be done with the **Omit atoms** option. As the name suggests, this option makes it possible for user-specified central-group atoms to be omitted from the final scatterplot. Atoms which are omitted in this way are not displayed in the scatterplot, not used in superimposing the central-group fragments or calculating the average central-group geometry, and not used in generating the distance slider.

Omitted atoms can be specified by selecting the appropriate option from the **Expert** menu, and then by clicking on the relevant atoms in the resulting pop-up display.

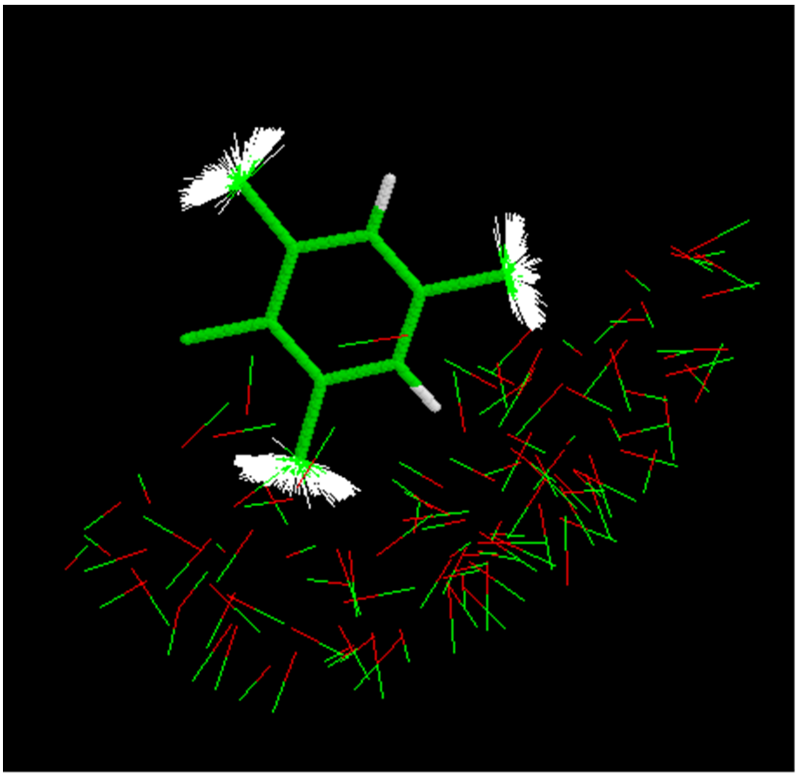
Conformationally flexible central groups may also be dealt with by means of the **Flexible atoms** option, described next.

## 13.5 Flexible atoms

When a scatterplot is generated, the various central-group fragments found in the ConQuest search must be superimposed and then averaged. Sometimes, a central group will contain a conformationally flexible moiety, such as a rotationally mobile methyl group. Inclusion of these moieties in the superposition process can result in distorted average central-group geometries.

In such cases, the user can designate the conformationally variable atoms as flexible. IsoGen will keep a separate atomic position for every flexible atom in the ConQuest output file. For example, here is a plot of carbonyl groups around a mesityl central group. The hydrogen atoms of the mesityl methyl groups are designated as flexible. Hence, every individual methyl group in the ConQuest file is shown explicitly.

Flexible atoms are not used in the central-group superposition calculation.



Flexible atoms can be specified by selecting the appropriate option from the **Expert** menu, and then by clicking on the desired atoms in the resulting pop-up display.

## 13.6 Contour probes

A contour probe is an atom of the contact group from which it is possible to compute a contoured density surface in IsoStar. For example, if the contact group were carbonyl, it would normally be possible to generate surfaces based on the positions of (1) the carbonyl carbons, or (2) the carbonyl oxygens; in this situation, both the carbon and oxygen atoms are therefore contour probes.

By default, all contact-group atoms will be defined as contour probes by IsoGen. Each probe atom will be assigned the label used in the ConQuest search.

The definition and labelling of contour probes can be changed by using the **Contour probes** option of the **Expert** menu. Selection of this option will delete the default definitions of contour probes and allow the user to select and label individual atoms. These atoms alone can then be used for the calculation of contoured density surfaces.

## 13.7 Use dummy atoms in fitting

Sometimes, a central group will contain one or more dummy atoms. This is usually when a variable element type has been defined in the ConQuest search. If the **Use dummy atoms in fitting** button is switched on, these atoms are included in the least-squares superposition of central-group fragments. They can be omitted from the fitting calculation by switching the button off.

# 14 Appendix 1: Vector Direction Syntax

## 14.1 Summary

Before IsoGen can be run, the user must have performed a ConQuest nonbonded search and save a number of output files, namely the `.tab`, `.fgd` and `.fgn` (using the **Export Parameters and Data** option from the **File** menu) and the `.cor` (using the **COORD: CSD Coordinate file** file type from the **Export Entries As...** option found under **File**). The `.cor` file contains the fractional crystallographic coordinates of each nonbonded contact found in the ConQuest search, together with some symmetry information.

From this starting point, and on Linux platforms only, the key steps in producing a scatterplot are:

1. Definition of program settings via the IsoGen graphical user interface.

2. Reading of data from ConQuest output files.

3. Detection of any topological symmetry in the central group (the group which will be overlaid).

4. Pass 1: Derivation of an initial average geometry for the central group.

5. Pass 2: Superposition of all central-group fragments in the ConQuest output file onto the average group derived in the preceding step. Conformational outliers are eliminated at this stage. Torsion angles are used to resolve ambiguities due to topological symmetry.

6. Derivation of 3D molecular symmetry (if requested).

7. Reduction of distribution into an asymmetric unit.

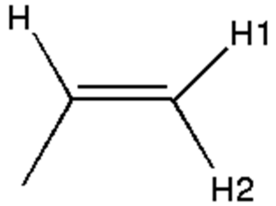
8. Writing of output files.

Further details of some of the methods used are summarised in the sections that follow:

- Derivation of Topological Symmetry (see Derivation of Topological Symmetry)
- Derivation of Initial Model (see Derivation of Initial Model)
- Superposition (see Superposition)
- Derivation of 3D Symmetry (see Derivation of 3D Symmetry)

## 14.2 Derivation of Topological Symmetry

Some chemical groups contain atoms that are topologically equivalent whilst not being symmetry equivalent in three dimensions. An example of such a fragment would be vinyl:



Here, the two hydrogen atoms H1 and H2 are not symmetry related in 3D but are topologically equivalent in 2D (i.e. the $=CH_2$ group has localised mm2 symmetry). As a result, the matching of H1 and H2 onto the atoms of the ConQuest search query is arbitrary. The consequence of this is that the Z-hydrogen (H2) and the E-hydrogen (H1) will appear in an arbitrary order in the ConQuest output files. IsoGen must detect this in order to achieve correct superposition of central-group fragments since, clearly, it must not attempt to fit the Z-hydrogen of one vinyl fragment onto the E-hydrogen of another.

Topologically equivalent atoms are detected by using a modified version of the Morgan algorithm [Morgan, 1965]. An explanation of this algorithm is provided later (see Appendix 2: Algorithm for Identifying Topological Equivalence).

## 14.3 Derivation of Initial Model

An approximate average geometry for the central group is derived as follows. The first central-group fragment in the ConQuest coordinate file is used as a mode onto which the second fragment is fitted. The average geometry of these two fragments is computed. The third fragment is then fitted onto this average. The average geometry is again updated. The process is repeated until all fragments in the ConQuest file have been fitted, giving a final average model for use in the next step.

Problems may occur with central groups that contain conformationally flexible moieties (e.g. rotationally mobile methyl groups). In such cases, the fitting procedure may result in artificially shortened average bond distances.

## 14.4 Superposition

A second pass through the data is carried out, in which each fragment is fitted onto the average model obtained above. Standard methodology is used for the fitting [Mietzner, 1997]. Fragments are rejected (i.e. eliminated completely from the scatterplot calculation) if they fit onto the average geometry with an RMS deviation exceeding 0.5 Å. Fragments are also rejected if, after fitting, they contain any individual atom that lies more than 1.0 Å from the corresponding atom in the average model. Currently, these tolerances cannot be changed by the user.

In cases where topologically equivalent atoms have been detected, torsion angles are used to resolve which atoms match onto which. For example, the Z-hydrogen of vinyl (H2 in the diagram above) can be distinguished from the E-hydrogen (H1) because the torsion angle C-C=C-H2 will be close to zero whilst the C-C=C-H1 torsion will be approximately 180 degrees. When topologically equivalent atoms cannot be distinguished on the basis of torsion angles, they are assumed to be symmetrically related in 3D.

# 15 Derivation of 3D Symmetry

The 3D symmetry is derived in a series of steps. First, the 2D-connectivity Morgan-algorithm scores are used to deduce the maximum order of symmetry that the central group could possibly have (assuming that the group is not completely linear). If there are no topologically equivalent atoms, the central group can have nothing higher than mirror symmetry, with all atoms lying in the mirror plane. The number of topologically equivalent atoms in a group defines the maximum order of symmetry by which these atoms can be related, and so defines the maximum order for searching.

For lower-symmetry central groups, the first step in locating symmetry is to transform the group to its moments of inertia. In higher-symmetry systems, the moments of inertia are not defined uniquely and so are inappropriate. (This can be visualised by considering elliptical solids. An ellipsoid that has three radii of differing lengths has mmm symmetry: the directions of the 3 principal axes are unique. Next, consider the case of an ellipsoid with 2 radii of equal length, and one different. In this case, the two degenerate axes can be placed anywhere in a circular plane. Finally, in the case where all three radii are equal - a sphere - the axes can be placed anywhere. In mathematical terms, this can be regarded as a problem of degeneracy: the eigenvalues of the inertia tensor are degenerate. The same situation occurs with molecules that have high-order symmetry elements.)

To resolve such situations, vectors from the molecular centroid to each atom in turn are calculated and tested for symmetry. The centroid-atom vectors to topologically equivalent atoms are then summed to produce further directions for symmetry testing. Additional directions for symmetry testing are calculated by cross-multiplying pairs of centroid-atom vectors.

The following algorithm is used to test for a rotational symmetry element of order n around any particular direction:

1. Matrix operations are generated for each operator in the symmetry element.
2. Each operator is applied in turn to each coordinate in turn.
3. A search is made to determine whether the atomic positions generated in step 2 ($G_i$) are equal, within defined tolerances, to any of the original atomic positions ($O_i$). The tolerances for non-hydrogen atoms are:
    - The angle $G_i$...centroid...$O_i$ is < 5.0 degrees.
    - The difference between the $G_i$...centroid distance and the $O_i$...centroid distance is < 0.1 Å.
    - Hydrogen atoms have larger tolerances to allow for the fact that their positions are usually determined with lower experimental precision.

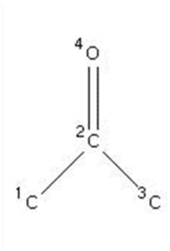
4. If any of the original central-group atoms cannot be matched with a symmetry-generated atom, the symmetry element is rejected.

5. If every operator for a symmetry element passes, the operator is accepted.

At the end of the process, a symmetry expansion routine is used. This computes the products of elements found in the above steps, and then tests for further symmetry.

# 16 Vector Direction Syntax

Vector directions can be specified in IsoGen using the numbers of the atoms from the ConQuest search query. So, if a ketone group had atoms C1, C2, C3 and O4, with C2 double bonded to O4 and single bonds C1-C2, C3-C2:



- The C2=O4 vector would be denoted as 2,4.
- The O4=C2 vector would be denoted as 4,2.
- The C1-C3 vector would be denoted as 1,3 etc.

Vectors can also be referenced with respect to the centroid of the molecule. For example, the ketone centroid-O4 vector would be denoted as c,4.

Three operations can be used to combine vectors: summation, subtraction, and cross-multiplication. For example:

The vector sum of the C2=O4 and C2-C3 directions would be denoted 2,4 + 2,3 (each vector is delimited from the operator with a space).

A vector perpendicular to the C2=O4 and C2-C3 directions (i.e. their cross product) would be denoted 2,4 x 2,3.

Currently, operator precedence is from left to right. Brackets are not currently supported, so if one specifies a vector as:

```
2,4 + 2,5 x 2,3
```

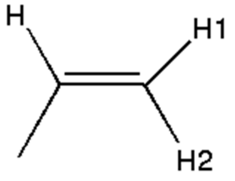
the program will interpret this as the cross product of the sum of 2,4 and 2,5 with 2,3 (i.e. [a + b]*c, where a, b and c are the three vectors). Conversely,

```
2,4 x 2,5 + 2,3
```

would be interpreted as the cross product of 2,4 and 2,5, summed with 2,3 (i.e. [a*b] + c).

# 17 Appendix 2: Algorithm for Identifying Topological Equivalence

The algorithm used is a modification of the Morgan algorithm and is best explained with an example:



Let us label the carbons C, $C_H$ and $C_{H1H2}$, and the hydrogens H, $H_1$ and $H_2$. Obviously, only atoms of the same elemental type can be topologically equivalent. Thus, it is immediately clear that the carbon atoms can be separated from the hydrogen atoms.

The algorithm proceeds by analyzing the extended connectivity in the following way. A score is assigned to each atom. Initially, the scores are computed by counting the number of bonds formed by each atom: i.e. C = 1, $C_H$ = 3 and $C_{H1H2}$ = 3. This tells us that C is unique; hence, amongst the carbons, only $C_H$ and $C_{H1H2}$ can possibly be topologically equivalent. All the hydrogens have a score (i.e. sum connectivity) of 1.

In the second iteration, the new score of each atom is calculated by summing the first-iteration scores of all the atoms to which it is bonded. CH gets a score of 1 (C) + 1 (H) + 3 ($C_{H1H2}$) = 5. $C_{H1H2}$ gets a score of 3 ($C_H$) + 1 ($H_1$) + 1 ($H_2$) = 5. H gets a score of 3. $H_1$ and $H_2$ also get scores of 3. Scores based on summing the atomic numbers of bound atoms are also computed: $C_H$ gets a score of 13, $C_{H1H2}$ gets a score of 8 and the protons all score 6. This means that $C_H$ is distinct from $C_{H1H2}$.

In the third cycle of iteration, the scores based on numbers of bonds become 5 for all the protons, but the scores based on atomic numbers become 13 for H, and 8 for $H_1$ and $H_2$. Thus, H is distinct from $H_1$ and $H_2$.

The termination criterion for the iterative process is when no further atoms can be assigned as unique by an iteration. At this point, we know which atoms are grouped together: those that had the same score at each iteration are topologically equivalent. In this example, the fourth pass shows that $H_1$ and $H_2$ are equivalent.

# 18 Appendix 3: References

Isostar: A library of information about non-bonded interactions I. J. Bruno, J. C. Cole, J. P. M. Lommerse, R. S. Rowland, R. Taylor and M. L. Verdonk, J. Comput.-Aided Mol. Des., **11**, 525-537, 1997.

Learning about Intermolecular Interactions from the Cambridge Structural Database G. M. Battle, F. H. Allen, J. Chem. Ed., **89**, 38-44, 2012, [DOI: 10.1021/ed200139t]

The Protein Data Bank H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, Nucleic Acids Res., **28**, 235-242, 2000.

# 19 Appendix 4: Acknowledgements

The CCDC gratefully acknowledges:

- The members of the Research Collaboratory for Structural Bioinformatics (RCSB), responsible for the maintenance of the PDB.
- Roger Amos and Anthony Stone (Theoretical Chemistry Department, University of Cambridge, UK) and Sally Price (Department of Chemistry, University College, London) who provided software for and advice about molecular orbital calculations.

- Gerhard Klebe (University of Marburg, Germany), Stefan Brode and Hugo Kubinyi (both of BASF-AG) who provided benchmark molecular orbital data.
- David Watson (CCDC) and Manfred Hendlich (University of Marburg) for providing information about the chemical connectivities of protein ligands in the PDB.

# 20 Appendix 5: IsoStar Tutorials

## 20.1 Tutorial 1: Navigating the IsoStar Web Pages

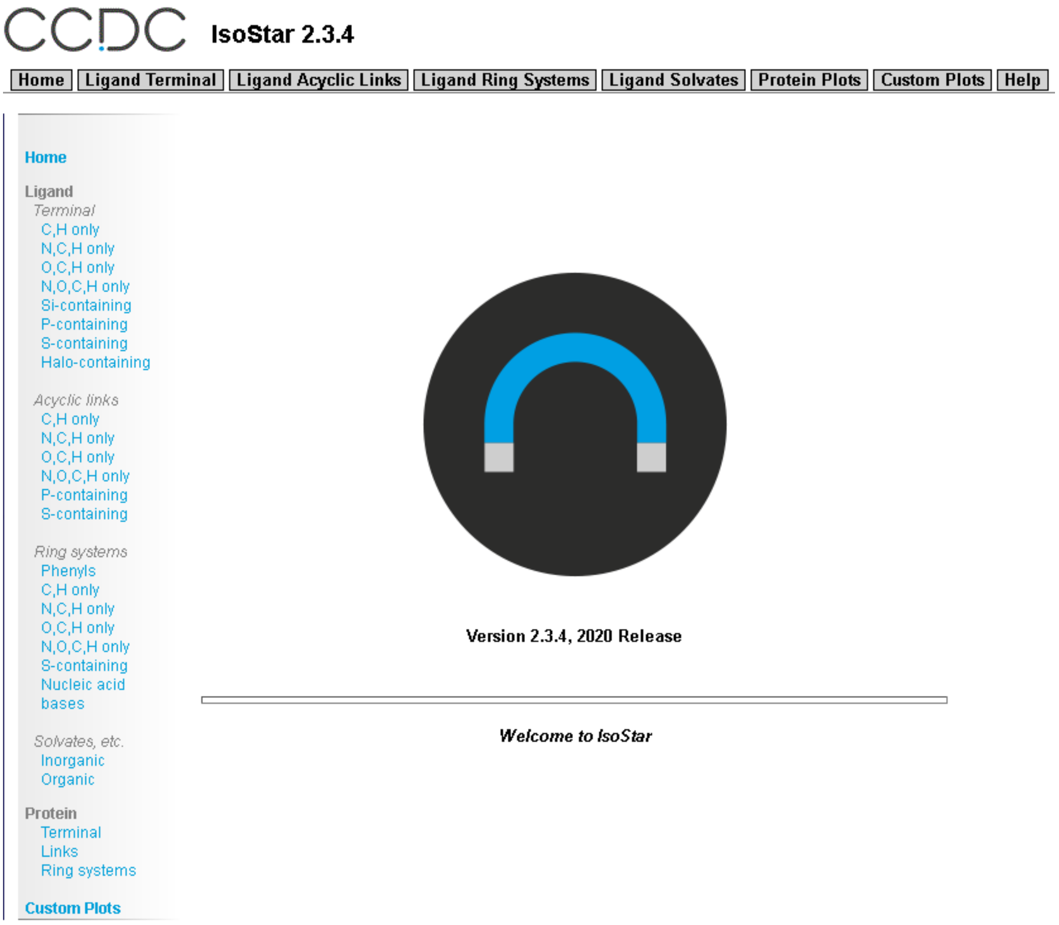
### 20.1.1 The Example

IsoStar is a library of information about the nonbonded interactions formed by a wide variety of chemical groups. It is based on data from the Cambridge Structural Database (CSD), the Protein Data Bank (PDB) and molecular orbital calculations. This example will demonstrate how to navigate the IsoStar webpages.

### 20.1.2 Selecting a Central Group

In your browser open the IsoStar homepage, e.g. http://isostar.ccdc.cam.ac.uk/html/isostar.html. Note: The above URL is the public IsoStar server.

IsoStar contains information about a large number of chemical groups, including terminal substituents, acyclic linking groups and ring systems. The major categories are listed on the left-hand side of the IsoStar Web page.



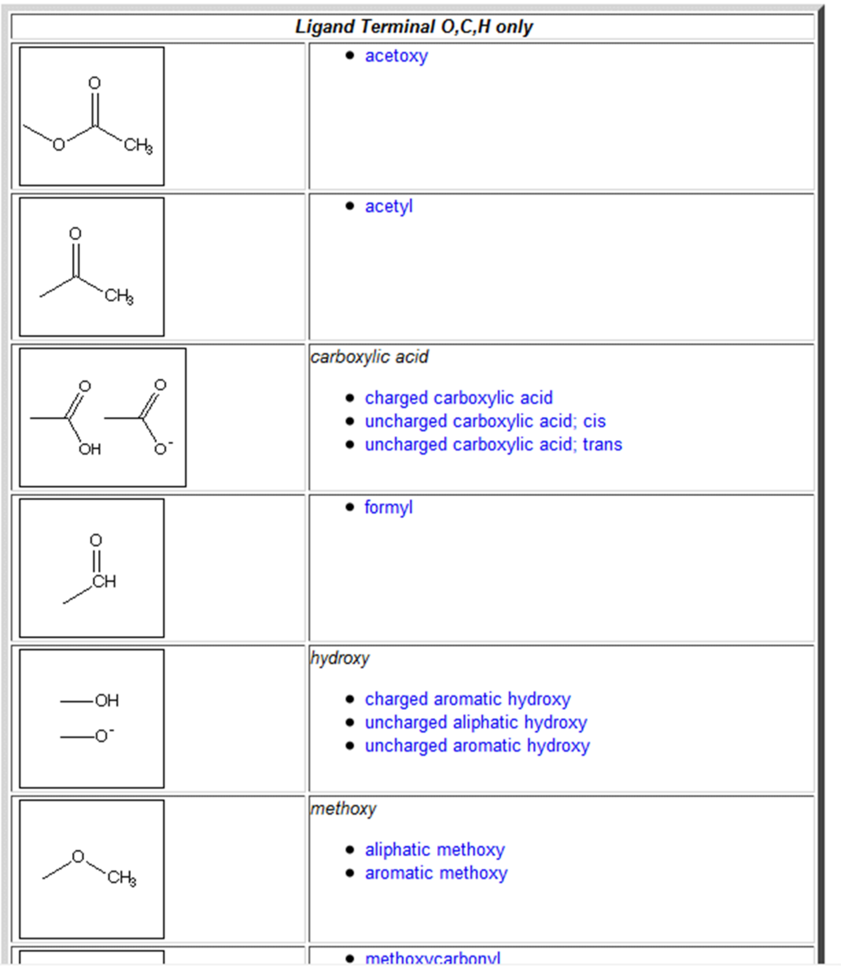
IsoStar 2.3.4

Version 2.3.4, 2020 Release

*Welcome to IsoStar*

Do not worry at this stage about the distinction between Ligand and Protein. We will start off in the **Ligand** section.

Since we are going to investigate carboxylic acids (i.e. a terminal group containing only O, C and H atoms), start by clicking on **O,C,H only** in the **Terminal** section of **Ligand**.

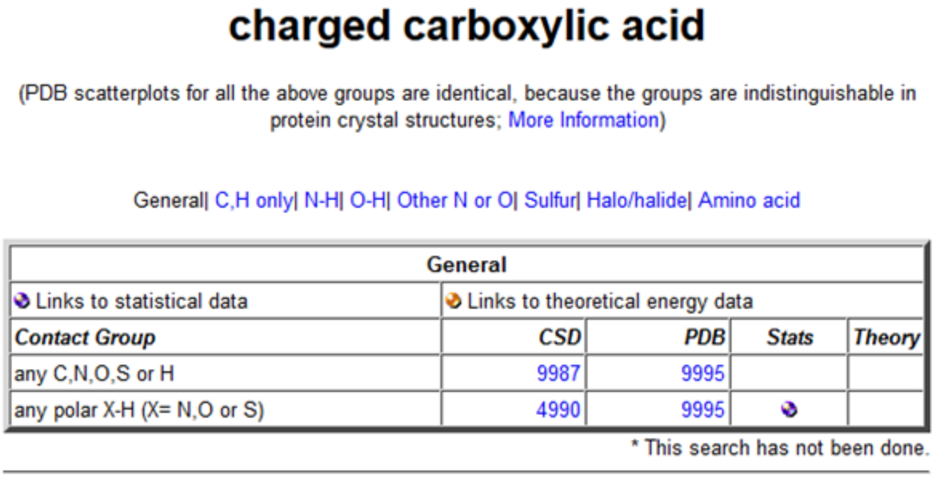


| Ligand Terminal O,C,H only | |
|---|---|
| | - acetoxy |
| | - acetyl |
| | *carboxylic acid*<br>- charged carboxylic acid<br>- uncharged carboxylic acid; cis<br>- uncharged carboxylic acid; trans |
| | - formyl |
| | *hydroxy*<br>- charged aromatic hydroxy<br>- uncharged aliphatic hydroxy<br>- uncharged aromatic hydroxy |
| | *methoxy*<br>- aliphatic methoxy<br>- aromatic methoxy |
| | - methoxycarbonyl |

Scroll through the list of functional groups that appears in the table. These are all the terminal O,C,H groups for which IsoStar contains information. We will start off by looking at carboxylate anions, so scroll back towards the top of the list and click on the **charged carboxylic acid** link.

This completes the selection of the central group.

## 20.1.3 Displaying CSD Scatterplots



### charged carboxylic acid

(PDB scatterplots for all the above groups are identical, because the groups are indistinguishable in protein crystal structures; More Information)

General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

| General | | | | |
|---|---|---|---|---|
| ● Links to statistical data | ● Links to theoretical energy data | | | |
| *Contact Group* | *CSD* | *PDB* | *Stats* | *Theory* |
| any C,N,O,S or H | 9987 | 9995 | | |
| any polar X-H (X= N,O or S) | 4990 | 9995 | ● | |

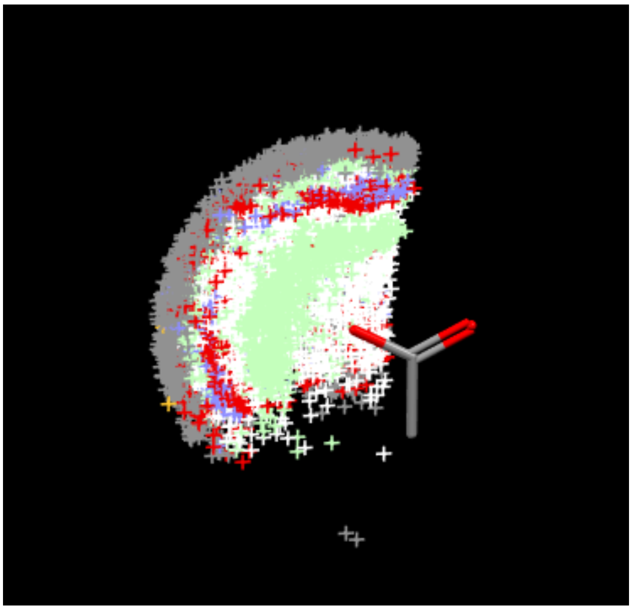
* This search has not been done.

The new page lists the available contact groups which are displayed in contact-group tables. The page is subdivided into a number of sections (headed **General**, **C,H only**, etc.). The contact-group table lists the types of contacts for which IsoStar contains information.

Launch the CSD-based scatterplot for the any C,N,O,S or H contact group, i.e. click on the hyperlink in the row marked **any C,N,O,S or H** and the column headed **CSD**. If IsoStar is installed correctly, this will launch the IsoStar visualiser displaying a scatterplot that shows the distribution of C, N, O, S and H atoms around carboxylate anions in the CSD. Note you may need to select **Open with CCDC.IsoStar (default)** in an **Opening csd0016_01.istr** window to view the plot. If you are running IsoStar on Linux you may need to manually associate the `.istr` file format with IsoStar.

The plot may take some time to load. If the central group is unclear, its style can be changed e.g. from **wireframe** to **capped stick** by selecting each atom (or shift-click one atom) and picking the relevant option from the right-click menu.

In the plot, non-polar hydrogens are coloured white, polar hydrogen are light green, carbons are grey, oxygens red, sulfur yellow and nitrogen atoms are blue.

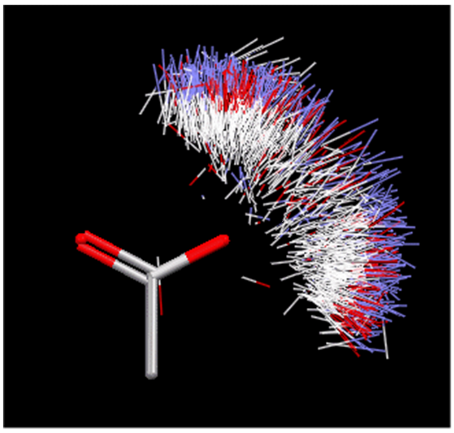


You can see the clusters of O, N and H atoms in the carboxylate-oxygen lone-pair directions, corresponding to OH...carboxylate and NH...carboxylate H-bonds.

Look back at the number you hit in the Web browser to display this scatterplot. This is the number of contact groups in the plot. Because carboxylate is a common group, the scatterplot would contain tens of thousands of atoms if it included all of the contacts in the CSD. This would make it very slow to load and manipulate. In setting up IsoStar, therefore, the plot has been limited to a total of about 10,000 atoms. This was done by taking a pseudo-random subset of the available CSD data.

Close the IsoStar graphical interface and return to the contact table in the Web browser.

To investigate H-bond geometries more closely, click on the number in the row marked **any polar X-H** and the column headed **CSD**. A scatterplot will appear showing the distribution of NH, OH and SH groups around carboxylates in the CSD. Note that the text at the top of the visualiser window updates to reflect the contents of the visualiser, in this case **Scatterplot [CSD] charged carboxylic acid, any polar X-H (X=N,O or S)**. Again, the plot shows that H-bonds tend to occur along the oxygen lone-pair directions. This becomes clear if you move the **Upper Limit** bar of the slider to the left, i.e. display shorter contacts.



## 20.1.4 Displaying PDB Scatterplots

1. Now go to the contact-group table and click on the number in the row marked **any polar X-H** and the column headed **PDB**.

2. In the **Overwrite Content Warning** window, click on the **Load in Other** button to load the scatterplot into the second visualiser window. You will now have the CSD-based plot adjacent to the PDB plot.

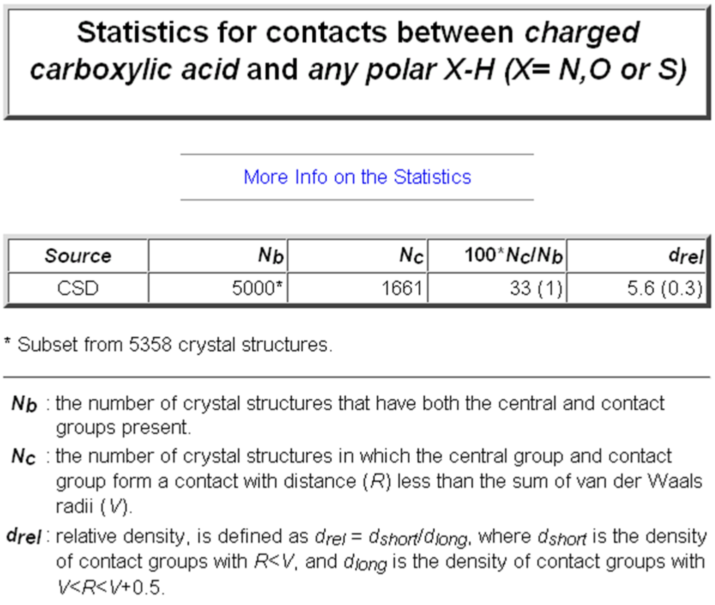
3. The PDB-based scatterplot displays the distribution of NH, OH and SH groups around carboxylates in PDB protein-ligand complexes. Since hydrogen atoms are not located in protein crystal structures, the plot shows only the non-hydrogen atoms.

4. Close the IsoStar graphical interface and return to the browser window.

5. Look at the contact-group table. Some of the cells are empty, i.e. contain no number. This means that no hits were found when the search for that contact was performed. If a cell contains a star, it means that the relevant search has not yet been done.

## 20.1.5 Displaying Statistical Data

1. Go to the contact-group table and hit the round purple button in the row marked **any polar X-H** and the column headed Stats. This displays some simple statistics about (N,O,S)H...carboxylate contacts in the CSD.



**Statistics for contacts between *charged carboxylic acid* and *any polar X-H (X= N,O or S)***

More Info on the Statistics

| *Source* | *Nb* | *Nc* | 100**Nc*/*Nb* | *drel* |
|---|---|---|---|---|
| CSD | 5000* | 1661 | 33 (1) | 5.6 (0.3) |

* Subset from 5358 crystal structures.

***Nb*** : the number of crystal structures that have both the central and contact groups present.

***Nc*** : the number of crystal structures in which the central group and contact group form a contact with distance (*R*) less than the sum of van der Waals radii (*V*).

***drel*** : relative density, is defined as $d_{rel} = d_{short}/d_{long}$, where $d_{short}$ is the density of contact groups with *R*<*V*, and $d_{long}$ is the density of contact groups with *V*<*R*<*V*+0.5.

2. The column headed 100*$N_c/N_b$ indicates that these contacts occur in a high percentage of structures, implying that they are very favourable energetically. The last column, $d_{rel}$, contains a simple statistic that shows how common short contacts are (i.e. shorter than the sum of van der Waals radii) compared to long contacts. The higher the number, the more frequently short contacts occur.

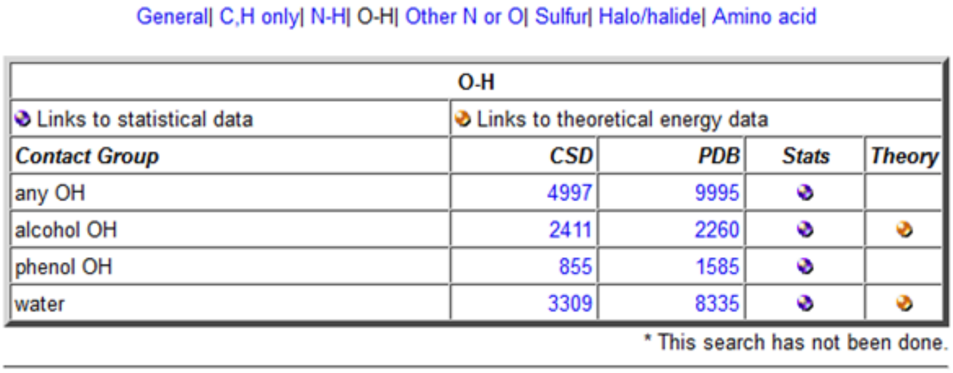
3. In this case, the number is very high (a value of <1 is probably typical for most other contacts). This suggests that short (N,O,S)H...carboxylate contacts are very common, and therefore, by implication, energetically favourable.

4. Click the browser **Back** button to return to the contact-group table.

## 20.1.6 Displaying Theoretical Data

You will notice some round, gold-coloured buttons in the contact-group table. These indicate that results from molecular orbital calculations are available.

1. Go to the section of the contact-group table headed O-H.



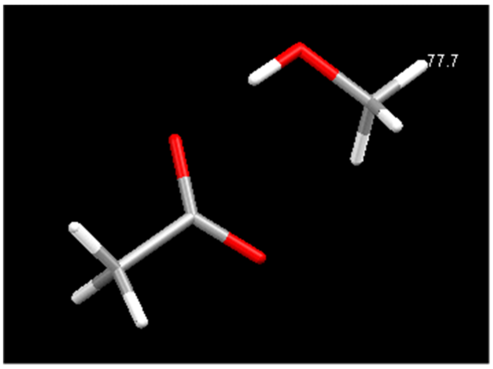
General| C,H only| N-H| O-H| Other N or O| Sulfur| Halo/halide| Amino acid

| O-H | | | | |
|---|---|---|---|---|
| ● Links to statistical data | ● Links to theoretical energy data | | | |
| *Contact Group* | *CSD* | *PDB* | *Stats* | *Theory* |
| any OH | 4997 | 9995 | ● | |
| alcohol OH | 2411 | 2260 | ● | ● |
| phenol OH | 855 | 1585 | ● | |
| water | 3309 | 8335 | ● | ● |

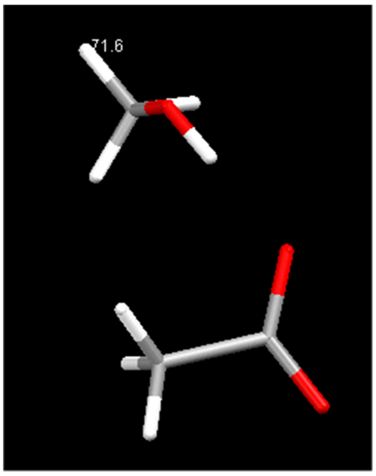
* This search has not been done.

2. Click on the gold button in the row marked **alcohol OH** and the column headed **Theory**. This displays a summary of calculations performed on the model system acetate...methanol. The table shows the interaction energies of the most important minima, broken down into different physical contributions. The total interaction energies, in the right-most column, are very large. This is because the calculations are on a gas-phase model system and one of the interacting species is charged.

3. Click on the number **1** in the column headed **Minima**. This will display the geometry of the lowest-energy minimum.



4. Now click on the number **2** to display the second-lowest minimum. The theoretical results support the crystallographic data in suggesting that the H-bonds prefer to form along lone-pair directions.

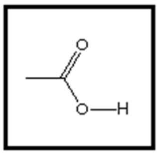


5. Close the IsoStar graphical interface and return to the Web browser window.

## 20.1.7 Looking at Closely Related Groups

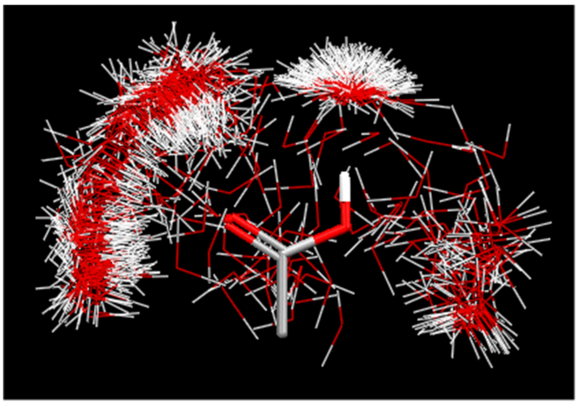
So far, we have looked at carboxylate anions. IsoStar also contains data for unionised carboxylic acids, which may exist in two different conformations (O=C-O-H cis or trans). These are shown at the top of the Web page.

1. Click on the hyperlink **uncharged carboxylic acid; cis** to display the contact-group table for this group.



uncharged carboxylic acid;
cis

2. Go to the section of the contact-group table headed **O-H** and click on the number in the row marked **water** and the column headed **CSD**. You will get a scatterplot showing the (CSD) distribution of water molecules around cis carboxylic acid groups. It shows that the acid carbonyl oxygen is a much better H-bond acceptor than the acid OH oxygen, which accepts only rarely. Note the cluster of H-bond acceptor water molecules near the acid OH hydrogen.



3. Close the IsoStar graphical interface and return to the Web browser.

## 20.1.8 PDB Plots of Ionisable Groups

Now for an important point about the PDB-based scatterplots of acidic and basic groups. Make a note of any number in the **PDB** column of the contact-group table and compare it with the corresponding numbers for **charged carboxylic acid** and **uncharged carboxylic acid; trans**. You should find that all three numbers are the same.

The reason is that the three groups carboxylate, carboxylic acid (cis) and carboxylic acid (trans) cannot be distinguished in protein X-ray structures, because the positions of hydrogen atoms are not located. Hence, the PDB data for these three groups are merged together. Thus, for any given contact group (e.g. water), you will get the same scatterplot whether you select from the charged, uncharged (cis) or uncharged (trans) contact-group tables.

Return to the **uncharged carboxylic acid; cis** page by clicking on the appropriate hyperlink at the top of the Web page.

## 20.1.9 Ligand and Protein Central Groups

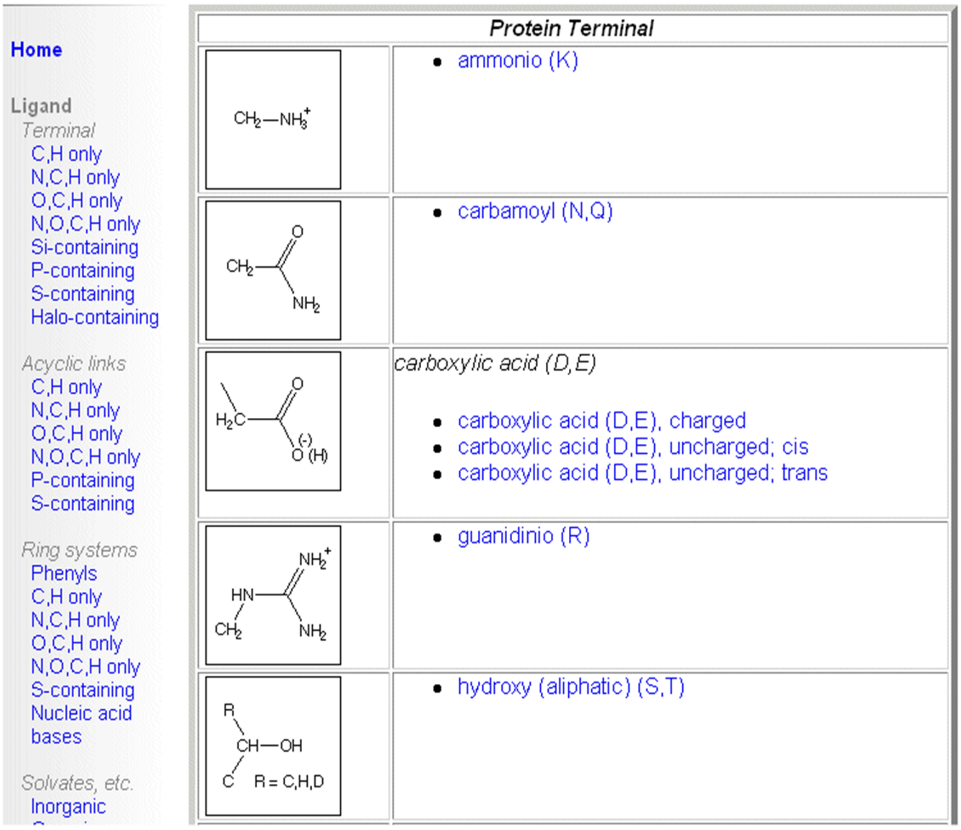
There is one more important point to be made about PDB data.

1. Look at the menu on the left-hand side of the Web page. The central groups are divided into two main sections, **Ligand** and **Protein**. Carboxylate can be found in both sections, but the PDB-based plots in the two sections are not the same. This is because the **Ligand** section contains contacts to carboxylate groups on ligand molecules. However, the plots in the **Protein** section show contacts to the carboxylate groups of Asp and Glu sidechains.

2. You are currently in the **Ligand** section of the IsoStar library. Click on any number in the **PDB** column of the contact group table (e.g. **amide N-H**) to display a PDB-based scatterplot.

3. Ensure that the hyperlink tick box is checked, then click on any contact atom in the scatterplot to hyperlink to that entry. A PDB reference code will appear in the **PDB Identifier** window and the hyperlinked structure will appear in the second visualiser window. The hyperlinked display shows a protein binding site; there is a contact between the active site ligand (which contains a cis carboxylic acid) and a residue in the protein (amide-containing).

4. Close the IsoStar visualiser window graphical interface and return to the browser. Scroll down the window until you can see the **Protein** section. Click on **Terminal**. You will obtain a list of terminal chemical groups that occur in the natural amino acids.



Home

Ligand
*Terminal*
C,H only
N,C,H only
O,C,H only
N,O,C,H only
Si-containing
P-containing
S-containing
Halo-containing

*Acyclic links*
C,H only
N,C,H only
O,C,H only
N,O,C,H only
P-containing
S-containing

*Ring systems*
Phenyls
C,H only
N,C,H only
O,C,H only
N,O,C,H only
S-containing
Nucleic acid bases

*Solvates, etc.*
Inorganic

| *Protein Terminal* | | |
|---|---|---|
| $CH_2—NH_3^+$ | | - ammonio (K) |
| $CH_2$, O, $NH_2$ | | - carbamoyl (N,Q) |
| $H_2C$, O, (-) O (H) | | *carboxylic acid (D,E)*<br>- carboxylic acid (D,E), charged<br>- carboxylic acid (D,E), uncharged; cis<br>- carboxylic acid (D,E), uncharged; trans |
| $NH_2^+$, HN, $CH_2$, $NH_2$ | | - guanidinio (R) |
| R, CH—OH, C, R = C,H,D | | - hydroxy (aliphatic) (S,T) |

5. Select any one of the carboxylic acid entries (e.g. **carboxylic acid (D,E), uncharged; cis**) from this list. The corresponding contact-group table will appear.

6. Click on any of the numbers in the **PDB** column to display a scatterplot (e.g. **amide NH** as before).

7. Hyperlink from any contact in the new scatterplot. A PDB reference code will appear in the **PDB Identifier** window and the hyperlinked plot will be displayed in the second visualiser window.

8. The resulting display will show a non-bonded contact to the carboxylic acid moiety of an Asp or Glu side chain on the protein. Remember, you are now in the **Protein** section of the library.

9. This illustrates that, for the PDB-based data, all the scatterplots in the **Ligand** section of IsoStar show contacts to functional groups on small (i.e. ligand) molecules. Conversely, all the scatterplots in the Protein section show contacts to functional groups on protein side chains (or the peptide linkages of protein backbones).

10. Strictly speaking, all the CSD-based scatterplots should be put under the **Ligand** section of IsoStar, since the CSD contains only small molecules, not proteins. However, CSD data have been included in the **Protein** section as well, so that comparisons can easily be made between nonbonded contacts to protein functional groups and contacts to similar groups in small molecules.

This ends the tutorial.

## 20.2 Tutorial 2: Manipulating IsoStar Scatterplots

### 20.2.1 The Example

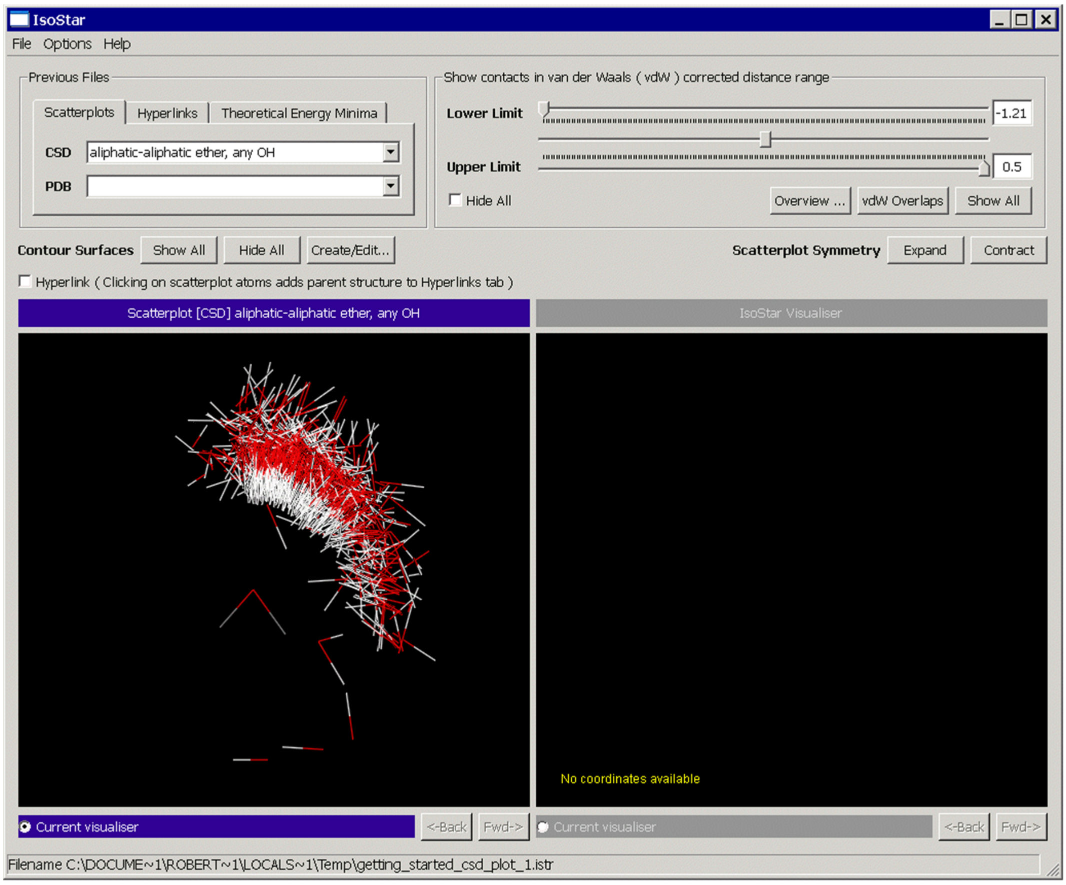
This example will introduce the user to some advanced features of the IsoStar Client software.

### 20.2.2 Getting Started

We are going to investigate how OH groups interact with aliphatic ethers. In the IsoStar home page http://isostar.ccdc.cam.ac.uk/html/isostar.html click on **O,C,H only** in the **Acyclic links** section of **Ligand.** Note: the above URL is the public IsoStar server. Scroll through the list of functional groups and click on the **aliphatic-aliphatic ether** link. Then from the **Contact group table** launch the CSD-based scatterplot for any OH contact group.

If IsoStar is installed correctly, the scatterplot will be displayed in one of the two visualisers in the IsoStar graphical interface. Move the window to a convenient location on the screen.

The visualiser window contains a scatterplot showing the distribution of OH groups around aliphatic ethers in crystal structures taken from the CSD. The plot was prepared by finding all OH...ether contacts in the CSD that lie within van der Waals radii + 0.5Å and overlaying them so that the ether portions were exactly superimposed. Because the ether group is symmetrical, all OH groups have been reflected into one quadrant.



IsoStar

Show contacts in van der Waals ( vdW ) corrected distance range

CSD: aliphatic-aliphatic ether, any OH

PDB:

Lower Limit: -1.21

Upper Limit: 0.5

□ Hide All

□ Hyperlink ( Clicking on scatterplot atoms adds parent structure to Hyperlinks tab )

Scatterplot [CSD] aliphatic-aliphatic ether, any OH

No coordinates available

Filename C:\DOCUME~1\ROBERT~1\LOCALS~1\Temp\getting_started_csd_plot_1.istr

Contact groups can be viewed for the remaining quadrants using the **Expand** button, next to **Scatterplot Symmetry**. If you do expand the symmetry, return the scatterplot to its initial state using the **Contract** button.

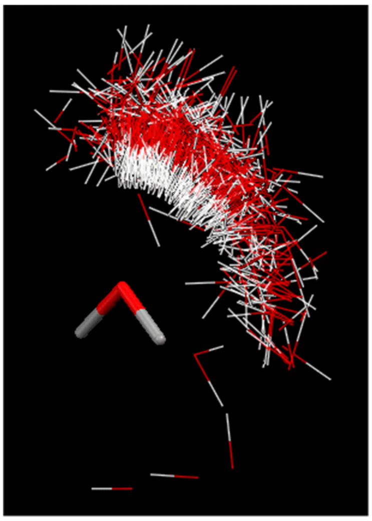
The scatterplot can be rotated with the left mouse button and translated with the middle mouse button (or using the left mouse button whilst pressing on the Control key on your keyboard). The plot can be zoomed in or out using the right mouse button.

Inspect the scatterplot. It shows that OH groups tend to lie quite close to the plane bisecting the C-O-C angle. Within that plane they can adopt a variety of positions.

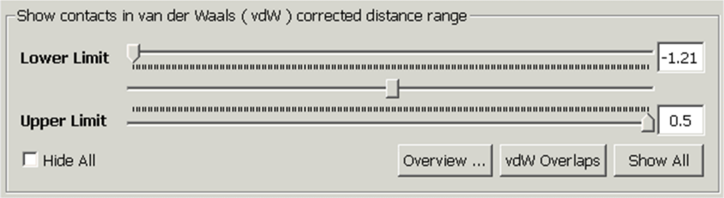
### 20.2.3 Exploring the Visualiser Interface and Interacting with a CSD Plot

A number of options for modifying the scatterplot can be displayed by right clicking in the 3D display.

1. Select the three atoms that make up the ether moiety (i.e. COC) and right-click anywhere in the 3D display. From the resultant menu, select **Styles** then **Capped Sticks**. The ether central group will now be displayed in the **capped stick** style, while the OH contact groups are displayed in the default **wireframe**.



2. The scatterplot currently displays all OH...ether contacts in which the ether oxygen is within a distance of (V + 0.5) Å to either the hydroxyl O or H atom (V = sum of van der Waals radii of O, O or O, H). This is illustrated in the slider controls at the top right of the IsoStar graphical interface.



Show contacts in van der Waals ( vdW ) corrected distance range

Lower Limit -1.21

Upper Limit 0.5

3. Hit the **Overview** button. The resultant dialogue lists all the CSD reference codes for the contact groups in the plot. The distances are reported based on how much shorter the specific contact is compared to vdW radii, thus the shortest contact, FIDVUK, is 1.21 Å shorter than the sum of van der Waals radii. Once you have finished inspecting the contact groups, close the dialogue.

4. Hit the button labelled **vdW Overlaps**. This removes all contacts that are longer than the sum of van der Waals radii from the scatterplot. Hitting the button marked **Show All** will re-display the longer contacts.

5. Alternatively, contacts can be removed by using the slider arrows. The UpperLimit arrow will remove longer contacts while the LowerLimit arrow will remove short contacts.

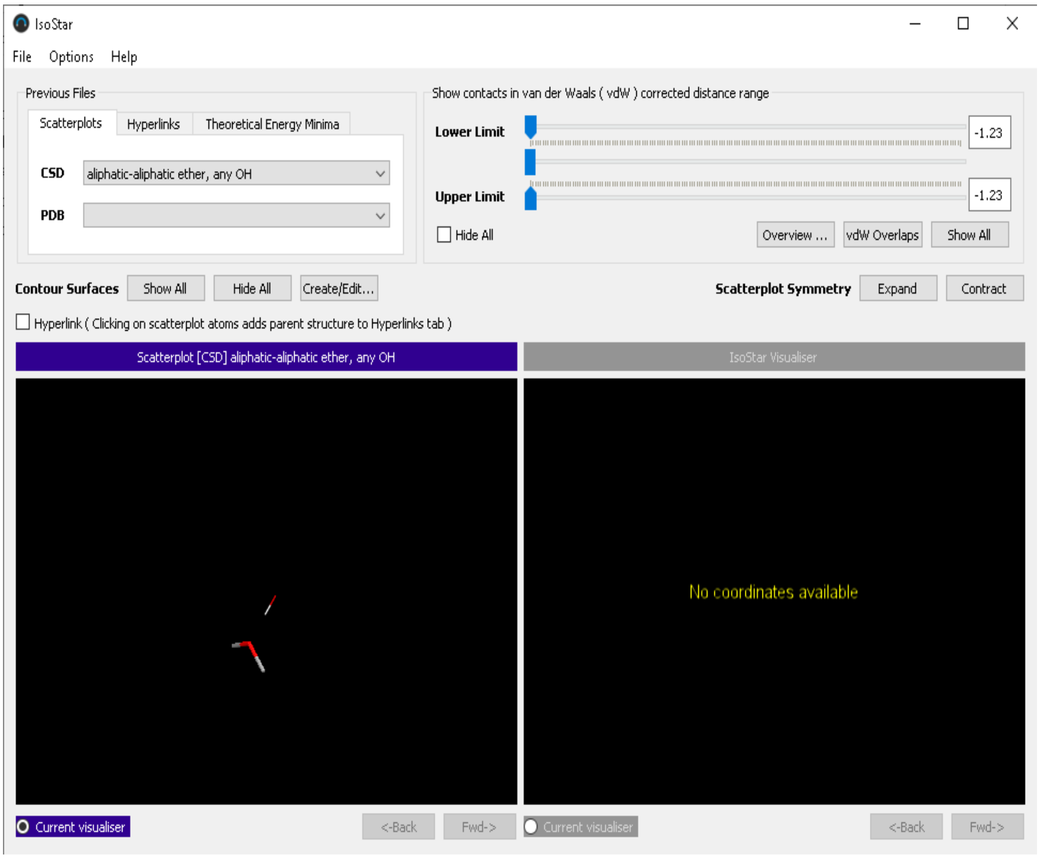
6. As the slider is moved, the two white boxes next to the slider update to reflect the range of contact distances on display.

7. For example, move the LowerLimit slider arrow until the top box reads -0.60. Now move the UpperLimit slider arrow until the number in the bottom box reads -0.50. All contacts now displayed are within the distance range (V - 0.6) Å to (V - 0.5) Å.

8. Moving the central part of the slider bar enables you to move this 0.1 Å display range upwards (i.e. up to longer contact distances) or downwards (towards shorter distances).

9. Specific values may also be typed into the white boxes.

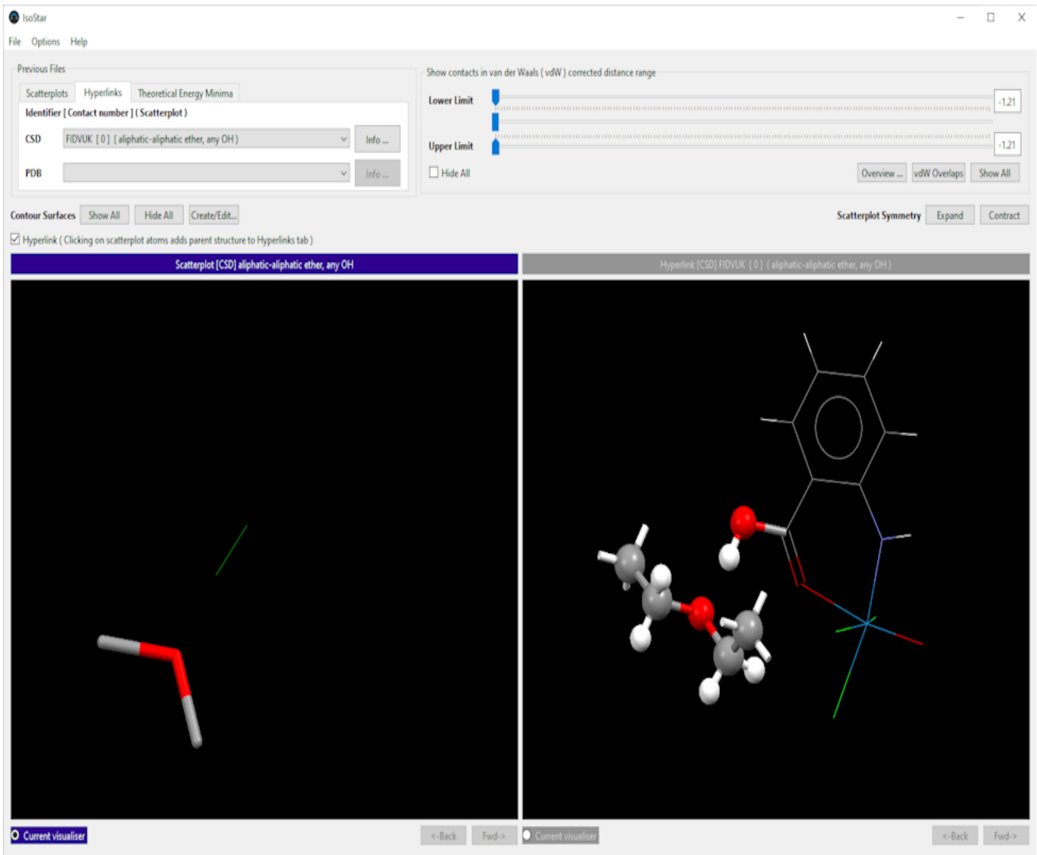
10. Hit the button labelled **Show All** to re-display all contacts. Then move the **UpperLimit** slider arrow as far left as it will go until only one contact (the shortest) is displayed.



### 20.2.4 Hyperlink to a CSD Structure

If we want to know more about this unusually short contact, we can hyperlink to the original CSD entry to view further details.

1. Activate the **Hyperlink** tick box above the 3D visualiser then click on either the H or the O atom of the hydroxyl contact group. The contact will turn green, indicating it has been selected, and the hyperlinked plot will be loaded into the second visualiser window. In addition, the tabbed view in the top left of the graphical interface has changed from **Scatterplots** to **Hyperlinks** and the CSD identifier (FIDVUK) is given in the **CSD Identifier** window.



Identifier [ Contact number ] ( Scatterplot )

CSD: FIDVUK [ 0 ] ( aliphatic-aliphatic ether, any OH )

Show contacts in van der Waals ( vdW ) corrected distance range

Lower Limit: -1.21

Upper Limit: -1.21

☑ Hyperlink ( Clicking on scatterplot atoms adds parent structure to Hyperlinks tab )

Scatterplot [CSD] aliphatic-aliphatic ether, any OH

Hyperlink [CSD] FIDVUK [ 0 ] ( aliphatic-aliphatic ether, any OH )

2. Ensure the hyperlinked display is set as the Current visualiser by clicking on the bar at the top of the visualiser.

3. The fragments involved in the contact are highlighted in ball and stick mode.

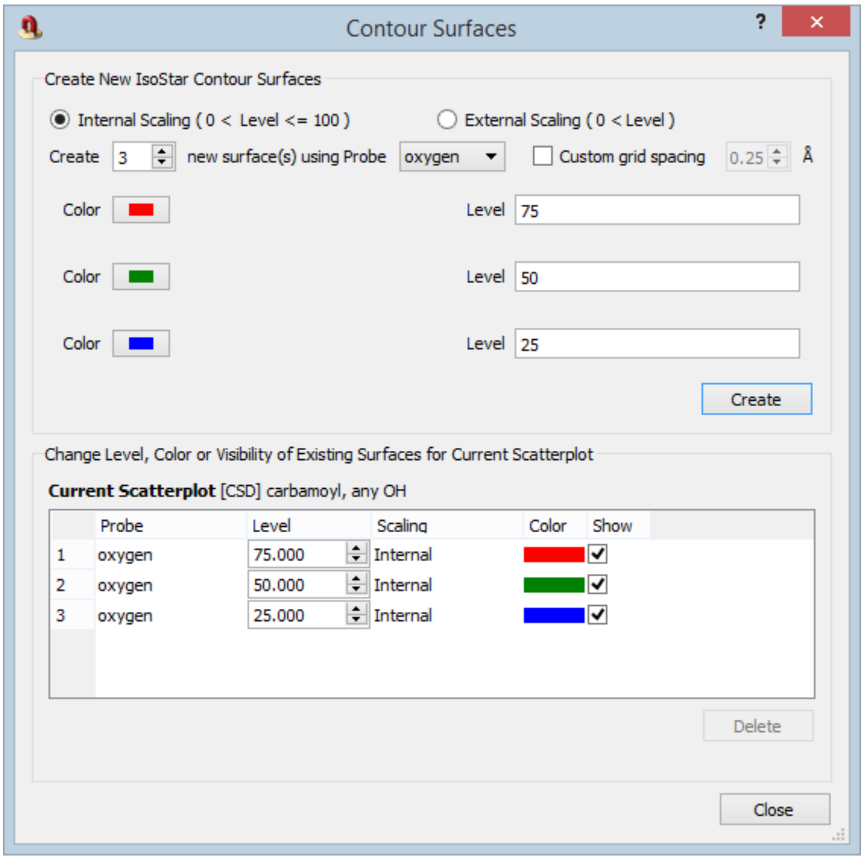
4. Measure the contact distance by right-clicking in the 3D display, selecting **Measure** and **Measure Distances** from the pull-down menu, then clicking on the ether O and hydroxyl H atoms. The distance will be reported as 1.563 Å.

5. Hit the **Info...** button to the right of the CSD Identifier display. This launches a dialogue that provides further information about the structure, e.g. the compound name, literature reference, and crystallographic information. Close this dialogue and return to the IsoStar visualiser window.

6. Return to the original scatterplot, ensure the display is set as the Current visualiser by clicking on the bar at the top of the visualiser, and display all contacts by clicking on the **Show All** button and return to the **Scatterplots** tabbed display.

### 20.2.5 Generation of Contoured Density Surfaces

Click on the **Create/Edit** button in the Scatterplots tabbed display. This launches a **Contour Surfaces** dialogue which allows you to apply contours to the current scatterplot.



Contour Surfaces

Create New IsoStar Contour Surfaces

● Internal Scaling ( 0 < Level <= 100 )    ○ External Scaling ( 0 < Level )

Create 3 new surface(s) using Probe oxygen    □ Custom grid spacing 0.25 Å

Color 🟥    Level 75

Color 🟩    Level 50

Color 🟦    Level 25

Change Level, Color or Visibility of Existing Surfaces for Current Scatterplot

**Current Scatterplot** [CSD] carbamoyl, any OH

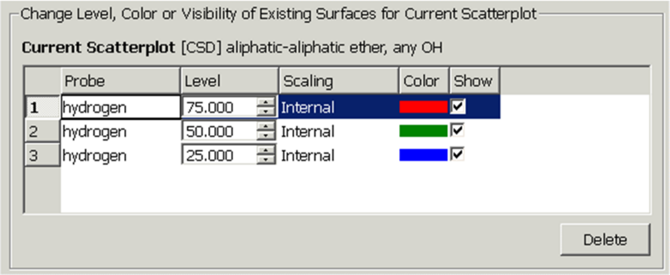
| | Probe | Level | Scaling | Color | Show |
|---|---|---|---|---|---|
| 1 | oxygen | 75.000 | Internal | 🟥 | ☑ |
| 2 | oxygen | 50.000 | Internal | 🟩 | ☑ |
| 3 | oxygen | 25.000 | Internal | 🟦 | ☑ |

Contour surfaces can be scaled in one of two ways, either **Internal** or **External**. The **Internal Scaling** radio button will already be activated. With Internal scaling, the densest region in the scatterplot is assigned a density of 100, and all other regions are scaled relative to this.

Grid spacings for the contours can be user-defined by activating the **Custom grid spacing** tickbox and modifying the grid spacing value.

The number of surfaces and the probe atom can be chosen using the **Create ... new surfaces using Probe ...** pulldowns. Note that a maximum of 3 surfaces can be generated. Keep the number of surfaces at **3** and select **hydrogen** from the probe pulldown menu.

Colours and levels are suggested for the different contours. Keep the levels as they are (i.e. at 75, 50 and 25). The colours can be modified if desired. Once the settings are as you want them, click **Create**. All three contours will be added to the **Current Scatterplot** section of the **Contour Surfaces** window.



Change Level, Color or Visibility of Existing Surfaces for Current Scatterplot

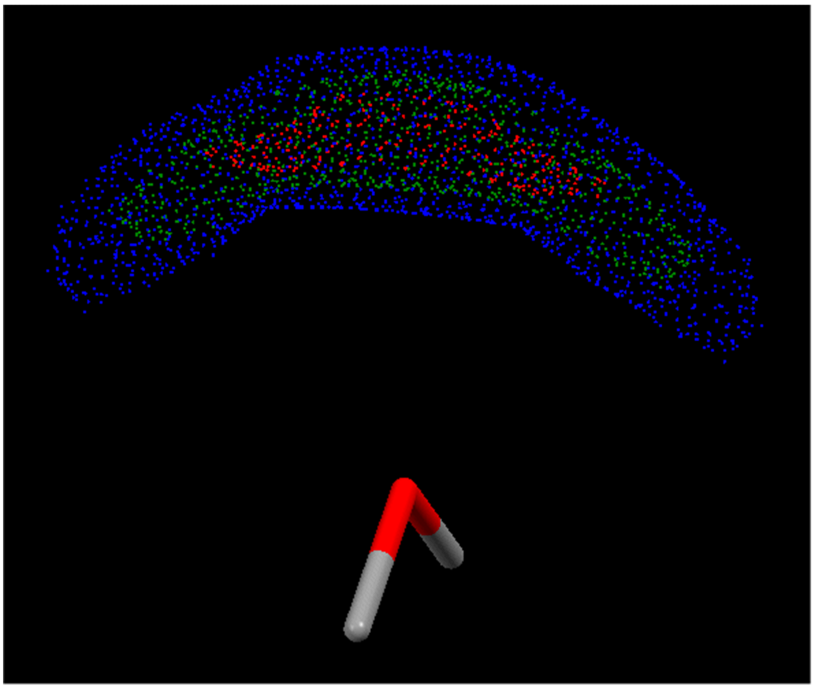
**Current Scatterplot** [CSD] aliphatic-aliphatic ether, any OH

|   | Probe | Level | Scaling | Color | Show |
|---|---|---|---|---|---|
| 1 | hydrogen | 75.000 | Internal | 🟥 | ☑ |
| 2 | hydrogen | 50.000 | Internal | 🟩 | ☑ |
| 3 | hydrogen | 25.000 | Internal | 🟦 | ☑ |

From within this dialogue it is possible to modify contour levels (using the up and down arrows adjacent to the **Level** value) and the contour colour. Whether a contour is displayed or not can also be controlled using the **Show** tickboxes.

Close the **Contour Surfaces** window using the **Close** button and return to the main IsoStar interface.

Both the scatterplot and the contour plot are present in the 3D view, so remove the scatterplot by activating the **Hide All** tick box.



You can also experiment with External Scaling settings.

Hide the existing contours by clicking on the **Hide All** button then open the **Contour Surfaces** window by clicking on the **Create/Edit** button.

In the **Contour Surfaces** window, activate the **External Scaling** radio button and ensure the probe is still set to **Hydrogen**.

As before, select new contour colours (if desired) but leave the **Level** settings as they are. Once you are happy with the settings, hit **Create**.

The new surface is scaled relative to the density of hydroxyl hydrogens that would be expected by chance, given the number of these atoms that occur in CSD structures. So, the red contour shows where the density of hydroxyl hydrogens is two times as high as would be expected by chance. The green contour corresponds to a density expected by chance and so on.

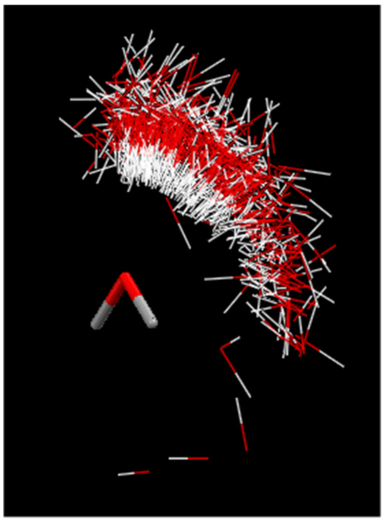
Close the **Contour Surfaces** window by hitting the **Close** button, then close the IsoStar graphical interface.

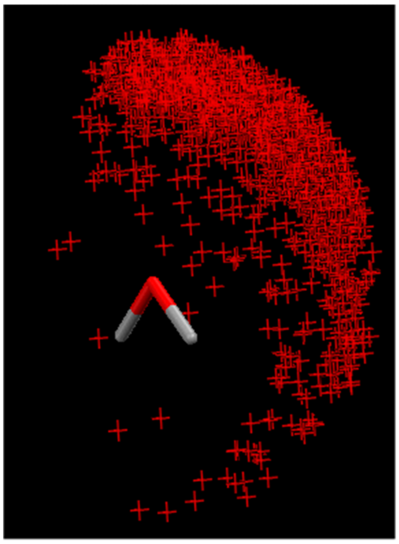
## 20.2.6 Inspecting a PDB Scatterplot

Display the distribution of OH…ether contacts as observed in protein-ligand complexes from the PDB. i.e. launch the PDB-based scatterplot for any OH contact group interacting with an **aliphatic-aliphatic ether** central group.

Because hydrogen atoms are not located in protein crystal structures, this plot does not show the H atoms of the hydroxyl groups, just the O atoms.

By comparing the PDB plot with the CSD plot, you can see that both distributions are similar geometrically.

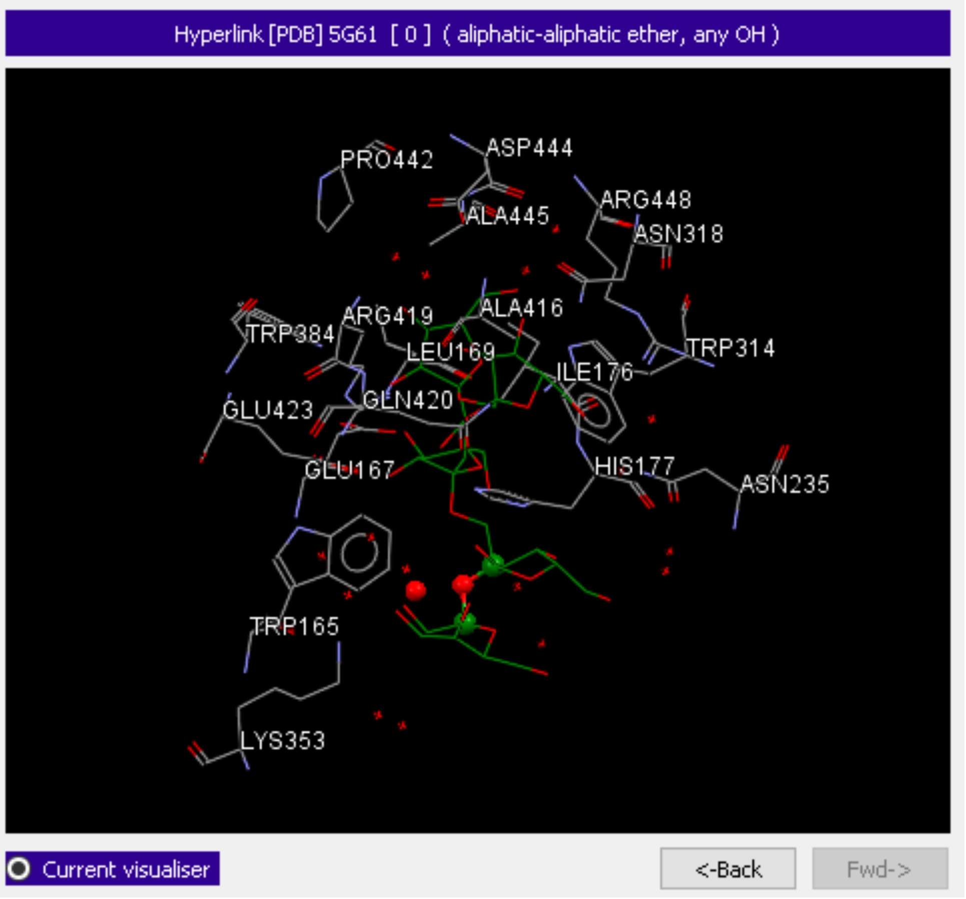




### 20.2.7 Hyperlink to a PDB Structure

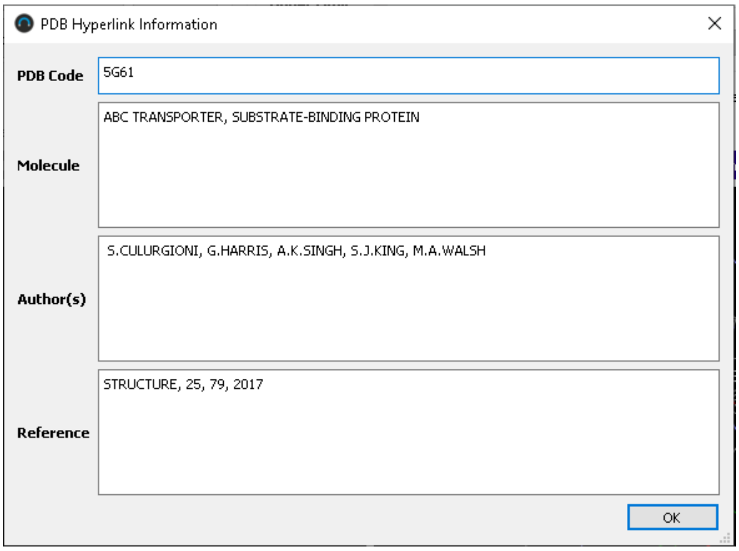
Move the **Upper Limit** slider to the far left to display the shortest contact.

Activate hyperlinking by clicking on the **Hyperlink Mode** tick box, then select the remaining single contact atom in the 3D view. The contact will turn green and the hyperlinked structure will appear in the second visualiser window.



Hyperlink [PDB] 5G61 [ 0 ] ( aliphatic-aliphatic ether, any OH )

The PDB code of the structure containing this short contact (5G61) is displayed in the PDB identifier window. Details of the structure including publication details are provided under the **Info...** button.



PDB Code: 5G61

Molecule: ABC TRANSPORTER, SUBSTRATE-BINDING PROTEIN

Author(s): S.CULURGIONI, G.HARRIS, A.K.SINGH, S.J.KING, M.A.WALSH

Reference: STRUCTURE, 25, 79, 2017

Inspect the active site in the hyperlinked structure. It shows the ligand together with nearby water molecules and protein residues. In fact the contact is between the active site ligand and a water molecule.

The contact distance can be measured in the same way as with the CSD plot, i.e. by right-clicking, selecting **Measure Distances** from the pull-down menu and selecting both atoms involved in the contact.

## 20.3 Inspection of a Theoretical Plot

Display the minimum energy geometry of the complex between methanol and dimethyl ether, i.e. launch the **Theory** plot for any alcohol OH contact group interacting with an aliphatic ether central group.

As the two visualiser windows contain the scatterplot and the hyperlinked plot, you will obtain an **Overwrite Content Warning** window, asking whether you would like to load the new plot into the **Current visualiser** (i.e. the one with the active **Current visualiser** radio button), or the other visualiser (with greyed out headers and footers). We are finished with both PDB plots so the new plot can be loaded into either window. Click on either **Load in Current** or **Load in Other**.

This plot was calculated theoretically. The number in white is the computed interaction energy between the two molecules, in kJ/mol - but note, this is a gas-phase result.

Close the IsoStar client window.

This ends the tutorial.

# 20.4 Tutorial 3: Calculating Customised Scatterplots Using IsoGen (Linux only)

## 20.4.1 The Example

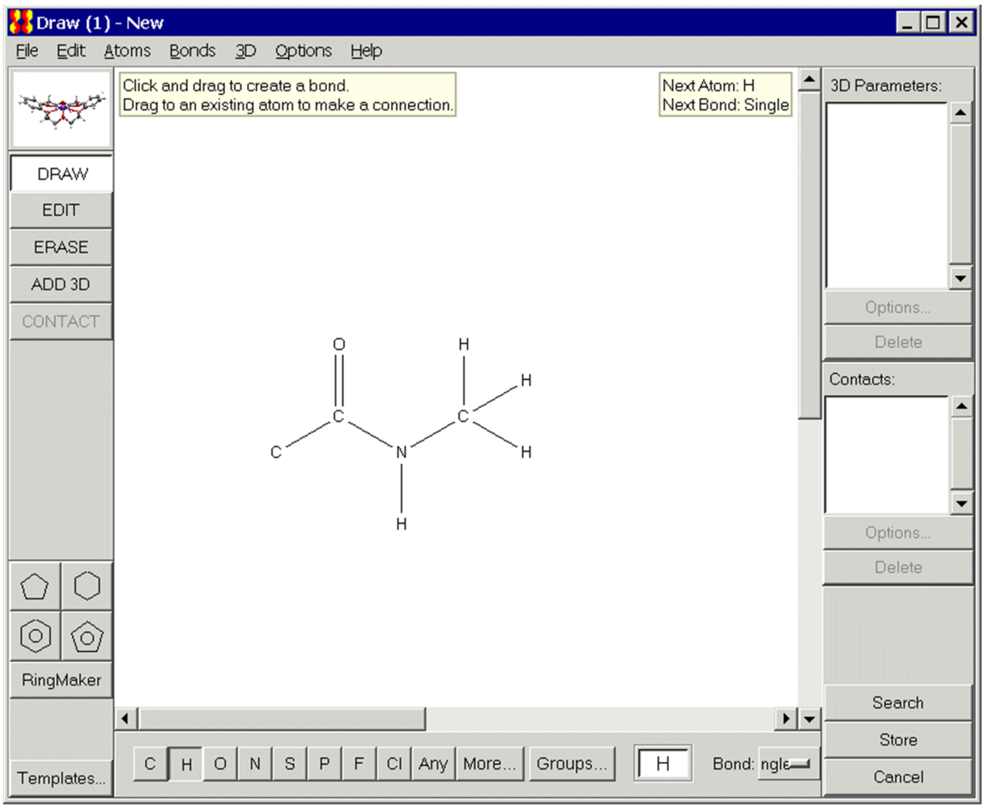
Although there is an extensive range of scatterplots in IsoStar, there will probably be occasions when you want to look at a plot that is not in the library. This tutorial will demonstrate how to perform a nonbonded search of the CSD, using ConQuest, and then convert the results into a scatterplot. We will take as an example the methylcarbamoyl group, -CONHMe, and look at how it forms hydrogen bonds.

It is assumed that you have basic familiarity with ConQuest, specifically how to build fragments and run simple substructure searches. If not, you should read the ConQuest documentation or use the ConQuest tutorials, available under the **Help** menu or at the end of the documentation.

## 20.4.2 Setting up a Nonbonded Search of the CSD

Open ConQuest and launch the sketcher window by clicking on **Draw** in the main interface.

Draw the methylcarbamoyl fragment, C-CONHMe. Include the H atoms and the carbon atom to which the methylcarbamoyl fragment is attached:

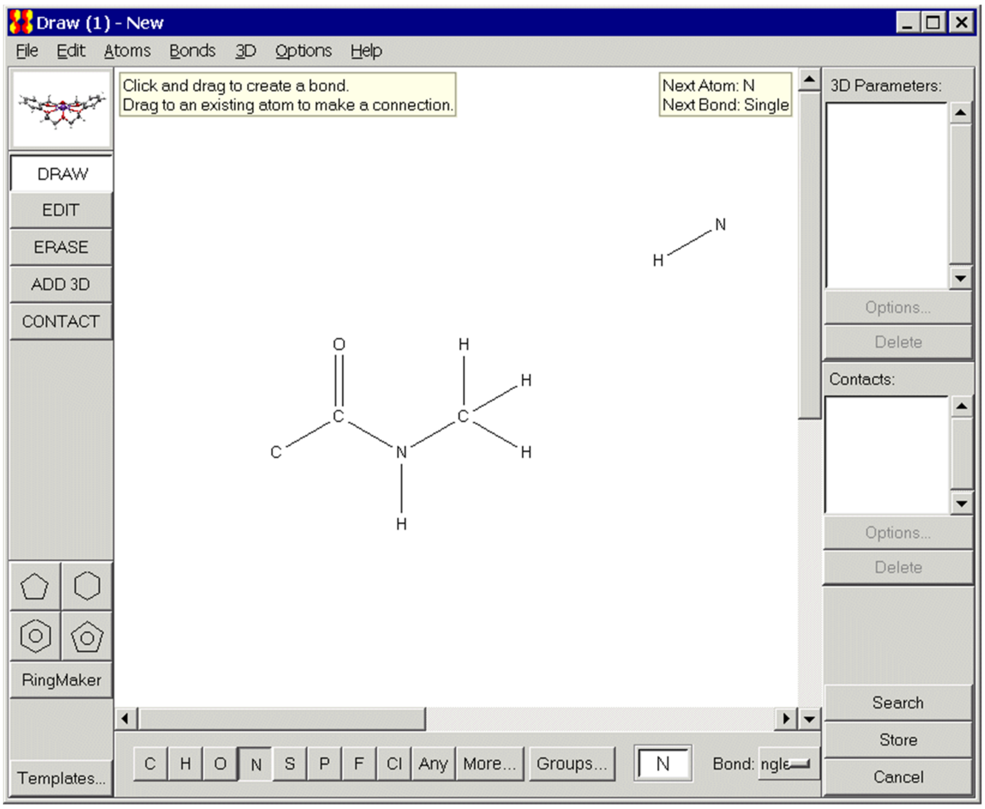


Click and drag to create a bond.
Drag to an existing atom to make a connection.

Next Atom: H
Next Bond: Single

Bond: ngle

Now draw a separate N-H group:



Click and drag to create a bond.
Drag to an existing atom to make a connection.

Next Atom: N
Next Bond: Single

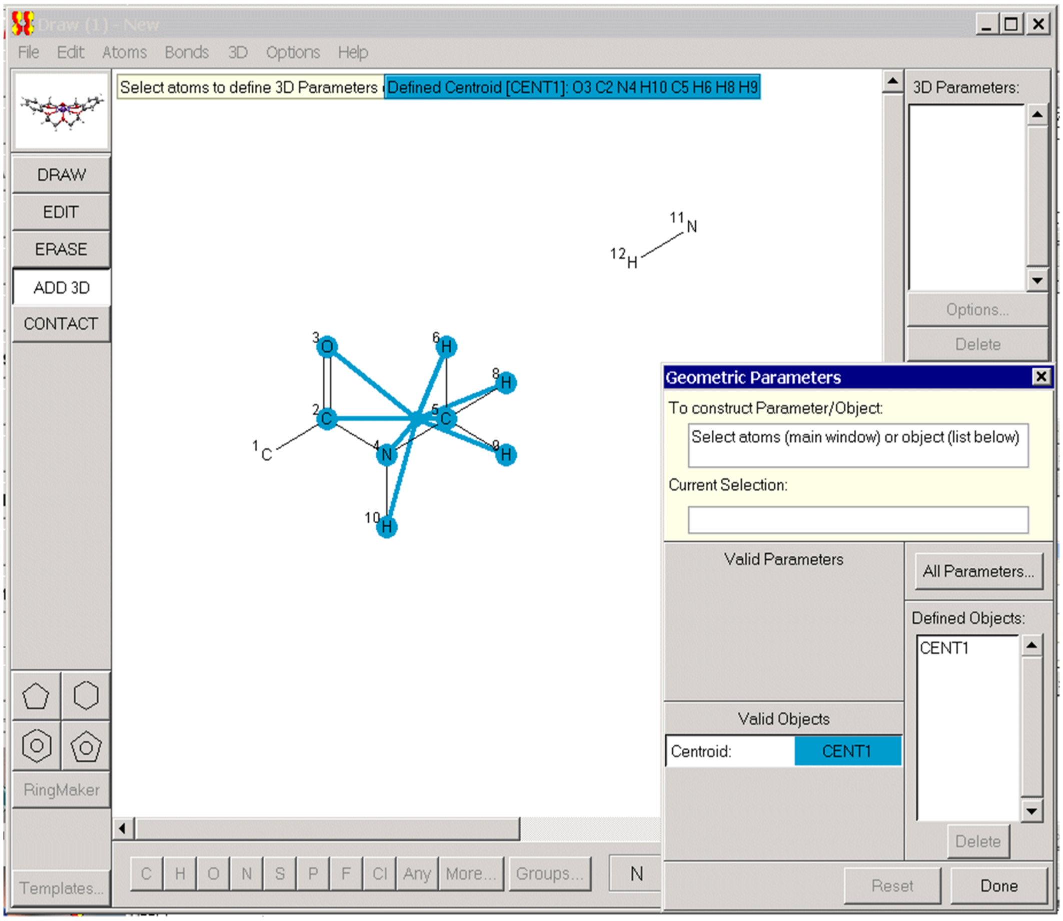
Bond: ngle

## 20.4.3 Setting Up 3D Constraints

Hit **ADD 3D**. This will generate a **Geometric Parameters** window.

Suppose that, in our non-bonded search, we want to find contacts between either the nitrogen or hydrogen atom of the N-H fragment, and any atom of the methylcarbamoyl fragment except for the point-of-attachment carbon. We can do this by specifying that the contact lies between two centroids, rather than specific atoms.

Hit each atom of the methylcarbamoyl fragment except for the point-of-attachment carbon (i.e. the carbon to which the -CONHMe is bonded) and pick **Define** next to **Centroid:** from the **Geometric Parameters** window. You have now defined a centroid, CENT1, comprising the desired atoms. The ConQuest window should look something like this:



Draw (1) - New

Select atoms to define 3D Parameters

Defined Centroid [CENT1]: O3 C2 N4 H10 C5 H6 H8 H9

Geometric Parameters

To construct Parameter/Object:

Select atoms (main window) or object (list below)

Current Selection:

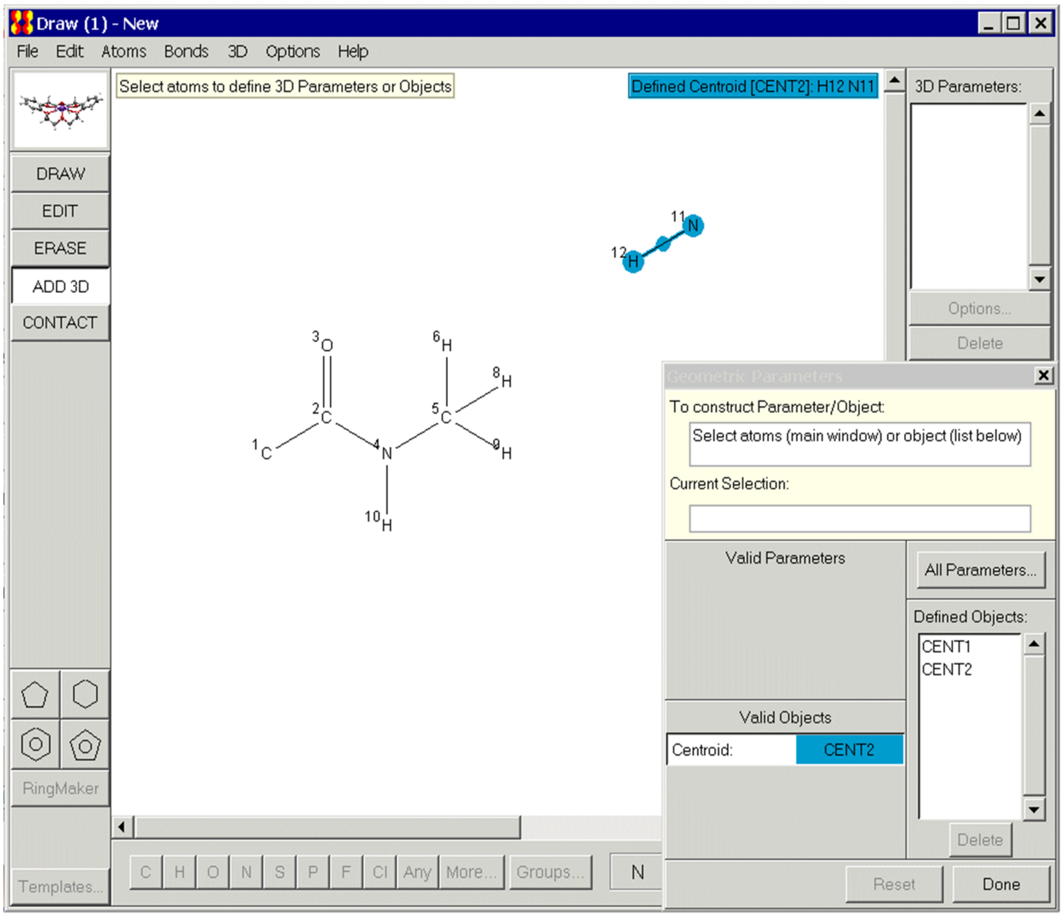
Valid Objects

Centroid: CENT1

Defined Objects:

CENT1

Now hit the **N** and **H** atoms of the N-H fragment, and then pick **Define** next to **Centroid:** again. A second centroid, CENT2, is now defined:



Select atoms to define 3D Parameters or Objects

Defined Centroid [CENT2]: H12 N11

To construct Parameter/Object:

Select atoms (main window) or object (list below)

Current Selection:

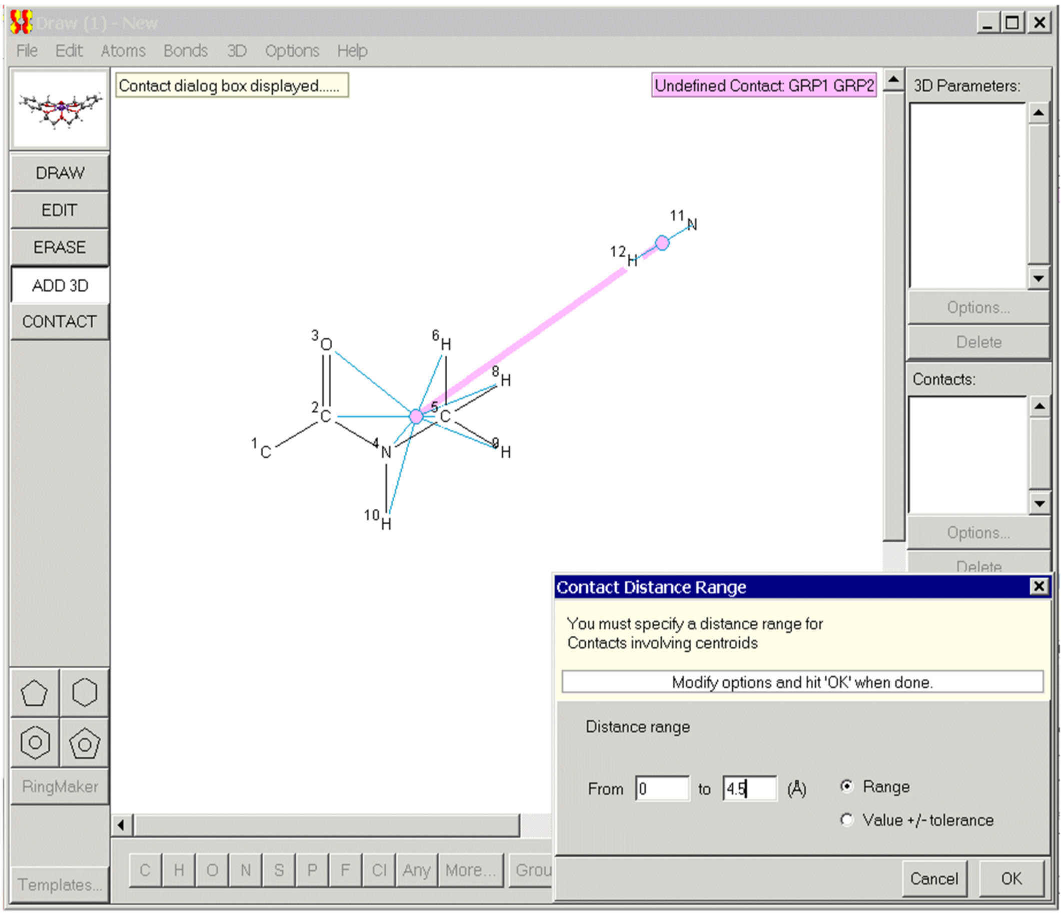
Defined Objects: CENT1 CENT2

Centroid: CENT2

Next, we need to define the actual search.

Select both **CENT1** and **CENT2** from the **Defined Objects** box, then click on the **Define...** box next to **Distance**. A **Distance Type** window will appear: select the **Contact** radio button then click **OK**. In the resultant window, ensure that the contact is defined between the two centroids and click **Define**.

We will now have to define a contact distance.



Contact dialog box displayed......

Undefined Contact: GRP1 GRP2

Contact Distance Range

You must specify a distance range for Contacts involving centroids

Modify options and hit 'OK' when done.

Distance range

From 0 to 4.5 (Å)

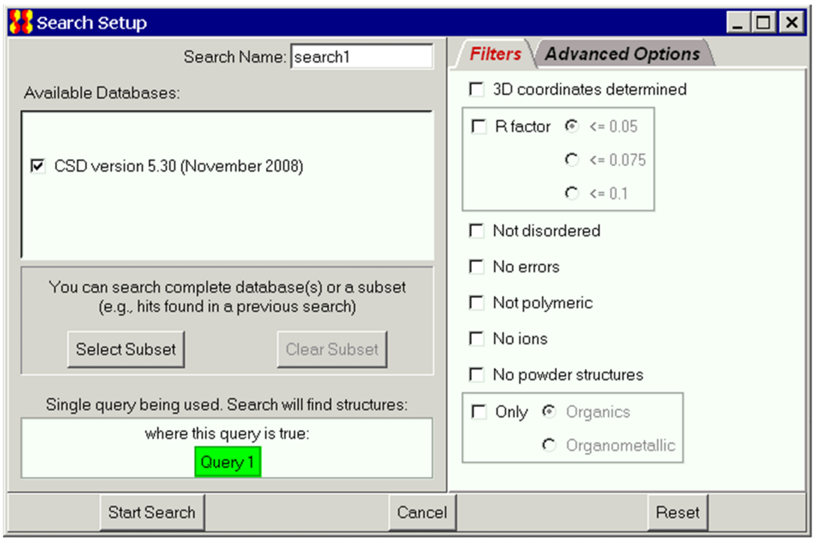
● Range

○ Value +/- tolerance

We will define the distance to be the sum of van der Waals radii, plus a tolerance of 1.0 Å. This gives a range of 0 to 4.5 Å which should be entered in the two distance range boxes. Click **OK** twice to close the **Distance Range** and **Non-bonded Contact Definition** windows, and then **Done** to close the **Geometric Parameters** window.

## 20.4.4 Running the Nonbonded Search

Hit **Search** in the sketcher window. A **Search Setup** window will pop up.



Search Setup

Search Name: search1

Available Databases:

☑ CSD version 5.30 (November 2008)

You can search complete database(s) or a subset (e.g., hits found in a previous search)

Single query being used. Search will find structures:

where this query is true:

Query 1

□ 3D coordinates determined

□ R factor ● <= 0.05 ○ <= 0.075 ○ <= 0.1

□ Not disordered

□ No errors

□ Not polymeric

□ No ions

□ No powder structures

□ Only ● Organics ○ Organometallic

Click on the **Advanced Options** tab and enable the **Normalise terminal H positions** tick box. This will instruct ConQuest to standardise all N-H bond lengths to 1.009 Å before searching for intermolecular contacts. This is an empirical way of correcting for the fact that N-H bond distances determined by X-ray diffraction are usually too short.

Now hit **Start Search** to set the search running. You will be taken to **View Results** window where a scrolling list of hits containing the substructure and contact is displayed.

## 20.4.5 Calculating a Scatterplot

To calculate a scatterplot from ConQuest data, you will need to save a number of files from the search you have just run. So create a new directory for these files in your home area.

In ConQuest, export the 3D parameters and data by clicking on **File** and selecting **Export Parameters and Data...** from the pull-down menu. Ensure that the **Include Defined Parameters** tick box is activated and the File Type option from the dropdown menu near the top left of the box is **Vista (.tab)**, then click **Save**. Browse to the directory you just created and save the files as e.g. `CONHMe.tab`.

The folder will contain three files with the extensions `.tab`, `.fgn` and `.fgd`.

Now export the fractional fragments by clicking on **File** and selecting **Export Entries As...**. Select **COORD: CSD Coordinate file** from the **Select file type** pull-down menu and ensure that the **All selected entries**, **Fractional** and **Hit Fragment Only** radio buttons are activated. This is particularly important as the Isostar plot will not be successfully generated if these buttons are not activated. Browse to the directory you created and save this file with the same name as before e.g. `CONHMe.cor`.

The new directory should now contain 4 files:

```
CONHMe.tab
```

```
CONHMe.fgn
```

```
CONHMe.fgd
```

```
CONHMe.cor
```

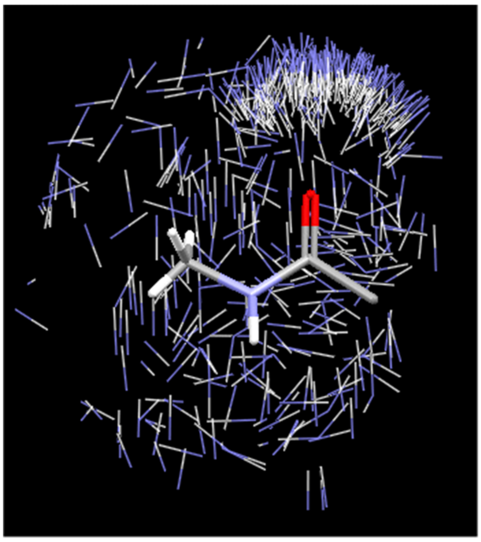
Browse to the directory containing all the files and open Isogen by typing the following at the command prompt:

```
isogen CONHMe &
```

You should now have a control window for a program called IsoGen, which reads the `.cor` and `.jnl` files and constructs an IsoStar-like scatterplot. In most cases (including this one), it is possible to calculate a reasonable scatterplot using the standard IsoGen defaults. The program will automatically detect topologically symmetrical atoms, overall fragment symmetry, conformational outliers, etc.

So, just hit the IsoGen top-level menu option **Run**. A new window will appear which will keep you updated on the progress of the scatterplot calculation. After some minutes, a pop-up window should appear, asking whether you want to view the scatterplot. Hit **Yes**.

The usual IsoStar graphical interface will be opened (if it is not already), showing the calculated scatterplot of N-H groups around methylcarbamoyl. You can view and manipulate the scatterplot in the usual way (see Tutorial 1 if you are not familiar with the operation of the IsoStar graphical interface).



Look at the files in your work directory. You will find one called `CONHMe.istr`. This is the file containing the scatterplot.

Close the IsoStar graphical interface and click **OK** in the IsoGen text window.

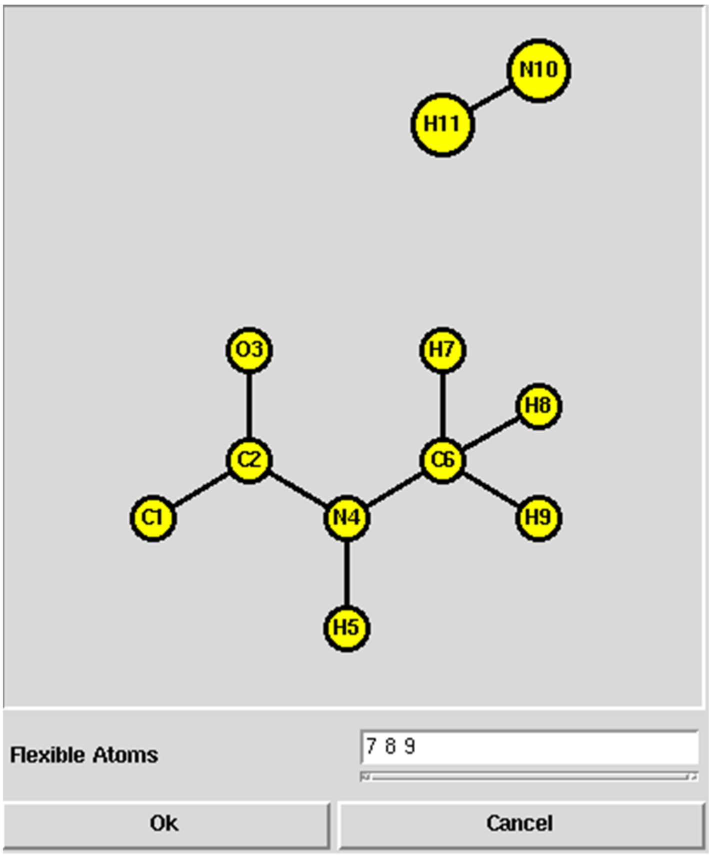
Most users can finish the tutorial at this point, since the key features of ConQuest and IsoGen have now been demonstrated. To exit the tutorial, close the IsoGen window with **File ... Exit IsoGen.**

However, if you want to see some Expert IsoGen options, continue with the tutorial.

## 20.4.6 Displaying Torsional Flexibility

We have produced a reasonably good scatterplot using the IsoGen defaults. However, the methyl group of the methylcarbamoyl fragment adopts a range of torsional conformations, and these have been averaged in the scatterplot. Suppose we want to show explicitly the existence of the methyl-group rotational flexibility. We can do this using IsoGen Expert options.

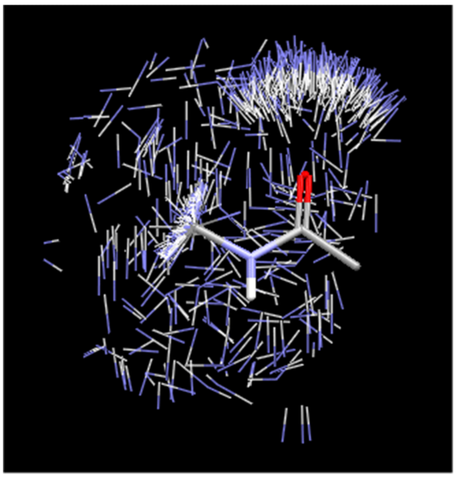
Select **Expert** from the IsoGen top-level menu, and then pick **Flexible atoms...** from the resulting sub-menu. A new window will appear, which shows the N-H and methylcarbamoyl fragments, as drawn in ConQuest. Pick the three methyl-group hydrogen atoms to define them as flexible. The window should look something like this (but your atom numbers will probably be different, as they depend on the order in which you drew the atoms and bonds in ConQuest):



Flexible Atoms 7 8 9

Hit **OK**.

Now hit the IsoGen top-level menu button **Run** again. A new scatterplot will be calculated. Don’t worry if you get some warning messages from the program - if necessary, just hit **OK** to remove them. When the pop-up window appears asking if you want to view the scatterplot, select **Yes**. You should now see the various orientations of the methyl group shown explicitly, something like this:



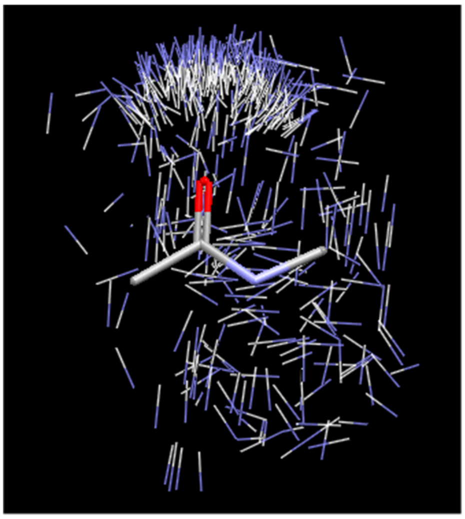
Close the IsoStar graphical interface and hit **OK** in the IsoGen text window.

### 20.4.7 Omitting Atoms

Another thing we can do is omit conformationally variable atoms altogether. First, we need to undo our previous **Flexible atoms** definition. Select **Expert** from the IsoGen top-level menu and **Flexible atoms...** from the sub-menu. Delete the numbers from the white box in the resulting pop-up window, and hit **OK.**

Now select the IsoGen option **Expert** and then **Omit atoms....** The search fragments are again displayed in a pop-up window. Pick the methyl-group H atoms and hit **OK**.

Hit **Run** in the IsoGen window to re-compute the scatterplot. This time, as you will see, the hydrogens of the methyl group are not displayed at all.



Finally, remove the windows that remain from the tutorial. The main IsoGen window can be removed by hitting **File** and then **Exit IsoGen.**

This finishes the tutorial, which has demonstrated how to produce scatterplots from the CSD by performing nonbonded searching in ConQuest and then post-processing the output with IsoGen. Scatterplots produced in this way may be made accessible from the IsoStar web page by putting them in the **Custom Plots** area.