
CTfile Formats

December 1999



December 1999

© Copyright 1999 by MDL Information Systems, Inc. All rights reserved. No part of this document may be reproduced by any means except as permitted in writing by MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

MDL is a registered trademark of MDL Information Systems, Inc.

ISIS is a trademark of MDL Information Systems, Inc.

All other product names are trademarks or registered trademarks of their respective holders.

U.S. GOVERNMENT RESTRICTED RIGHTS NOTICE

This software is provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subdivision (c)(1)(ii) of the Rights in Technical Data and Computer Software clause at FAR 252.227-7013. Contractor/Manufacturer is:

MDL Information Systems, Inc.
14600 Catalina Street
San Leandro, CA 94577

Table of Contents

Chapter 1	Introduction	
	Change Log1-1
	Standard CTfiles1-2
Chapter 2	The Connection Table [CTAB]	
	The Counts Line2-3
	The Atom Block2-4
	The Bond Block2-6
	The Atom List Block2-7
	The Stext Block.2-7
	The Properties Block.2-8
	The Properties Block for 3D Features.2-19
	3D features count line.2-21
	3D features detail lines2-21
	3D data constraints2-31
	Stereo Notes2-33
Chapter 3	Molfiles	
	The Header Block.3-2
Chapter 4	RGfiles	
Chapter 5	SDfiles	
	SDfile after a CFS search5-4
Chapter 6	Rxnfiles	
	Header Block6-1
	Reactants/Products6-3
	Molfile Blocks.6-3

Chapter 7	RDfiles	
	RDfile Header	7-2
	Molecule and Reaction Identifiers	7-2
	Data-field Identifier	7-3
	Data	7-3
Chapter 8	Atom Limit Enhancements	
	Phantom Extra Atom	8-1
	Superatom Attachment Point	8-2
	Superatom Class	8-3
	Large REGNO	8-3
	Sgroup Bracket Style	8-3
Chapter 9	Moving CTfiles On and Off the Clipboard in ISIS	
	Clipboard Objects	9-1
	Hints on Creating a Reader/Writer For CT	9-2
	Copying from the Clipboard	9-2
	Copying to the clipboard.	9-3
	Sample Code For Copying or Pasting a CTfile in MS Windows	9-4
Chapter 10	The Extended Molfile Format	
	Specifications For Atom and Bond Descriptions	10-3
	Conventions	10-4
	The Extended Connection Table	10-5
	CTAB block	10-5
	Counts line	10-5
	Atom block.	10-6
	Bond block.	10-9
	Link atom line	10-11
	Sgroup block	10-12
	3D block	10-20
	The Extended Rgroup Query Molfile	10-23
	Rgroup block	10-24
	Rgroup logic lines.	10-26

Introduction

MDL Information Systems supports a number of file formats for representation and communication of chemical information. This document describes the formats for MDL's CTfiles (chemical table files):

- Part I (Chapters 2 through 9) describes the standard CTfile formats.
- Part II (Chapter 10) describes the extended molfile format. All extended molfiles can be easily identified by the "V3000" version stamp in the header portion of the file. You are most likely to encounter the extended molfile format in CTfiles written from ISIS/Host or ISIS/Desktop version 2.0 or higher.

Change Log

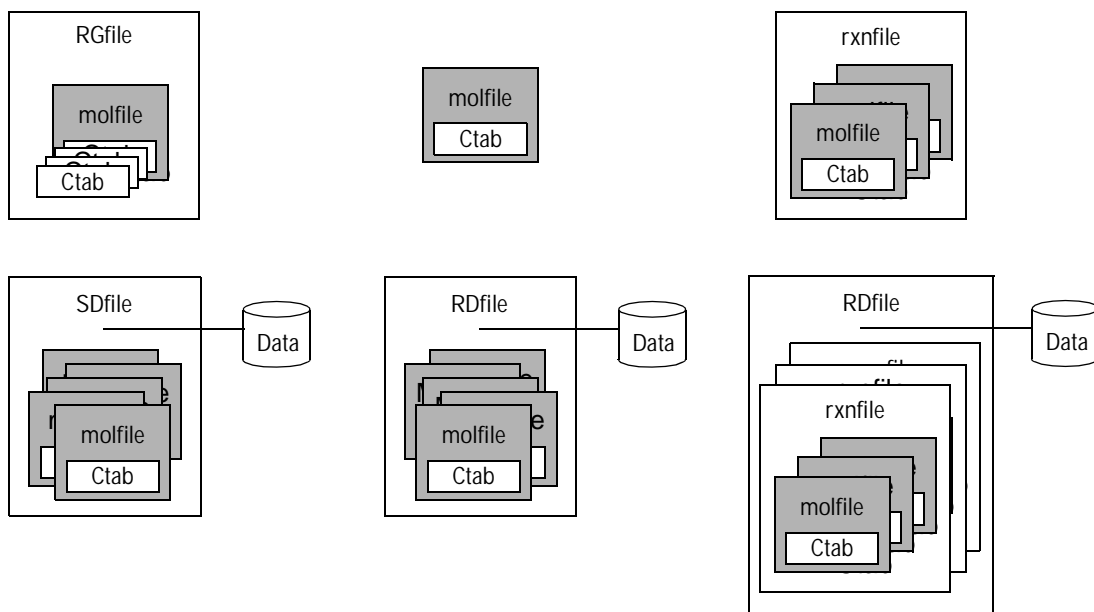
The following are the changes in this document:

Change	Page(s)
December, 1999	
Updated entries in "Atom List"	2-11
December, 1998	
Updated "Example of an SDfile"	5-3
August, 1998	
Added STBOX field	10-9
June, 1997	
Added Atom Attachment Order	2-11
Added new ATTCHORD field	10-6, 10-8

Change	Page(s)
October, 1996	
Minor corrections	2-3, 2-8
Enhanced description of connection table properties block	2-8
Added Sgroup bracket style	8-3, 10-14, 10-19,

Standard CTfiles

The following figure illustrates the relationship between the various file formats described below:



molfiles

Molecule files: Each molfile describes a single molecular structure which can contain disjoint fragments.

RGfiles

Rgroup files: An RGfile describes a single molecular query with Rgroups. Each RGfile is a combination of Ctabs defining the root molecule and each member of each Rgroup in the query.

rxnfiles	Reaction files: Each rxnfile contains the structural information for the reactants and products of a single reaction. MDL currently has two types of rxnfiles: the REACCS type and the CPSS-rxnfile written by CPSS programs (CPSS-rxnfiles are not described in this document.) CPSS programs cannot read a REACCS rxnfile; however, REACCS can read and write CPSS-rxnfiles for transfer to CPSS.
SDfiles	Structure-data files: An SDfile contains structures and data for any number of molecules. Together with RDfiles, SDfiles are the primary format for large-scale data transfer between MDL databases.
RDfiles	Reaction-data files: Similar to SDfiles in concept, the RDfile is a more general format that can include reactions as well as molecules, together with their associated data. Although RDfiles are used primarily by ISIS and REACCS, MACCS-II can also read and write RDfiles except for the reaction structure information (indicated by the square brackets in Table 1-1). CPSS reads and writes RDfiles with embedded molfiles and CPSS-rxnfiles (indicated by the curly brackets in Table 1-1).

Table 1-1 shows which CTfiles MDL programs can read and write.

Table 1-1 MDL Program

CTfile Type	MACCS-II	REACCS	ISIS	CPSS
molfiles	+	+	+	+
RGfiles	+		+	
rxnfiles		+	+	{+}
SDfiles	+		+	+
RDfiles	[+]	+	+	{+}

Some of the structural and query properties described in this document are generic in their applicability, while others are peculiar to certain CTfile types (see Table 1-2). The applicability of each property is identified in subsequent chapters by the icons shown in Table 1-2.

Table 1-2 Properties and identifying icons applicable to various CTfile types

Icon	Property	molfile	RGfile	SDfile	rxnfile	RDfile
G	Generic	+	+	+	+	+
Sg	Sgroup	+	+	+		
Rg	Rgroup	+	+	+		
3D	3D	+	+	+		
CP	CPSS	+		+	+	+
Rx	Reaction				+	+
Q	Query	+	+		+	

PART I
Standard File Formats

The Connection Table [CTAB]

CTab

A connection table (Ctab) contains information describing the structural relationships and properties of a collection of atoms. The atoms may be wholly or partially connected by bonds. Such collections may, for example, describe molecules, molecular fragments, substructures, substituent groups, polymers, alloys, formulations, mixtures, and unconnected atoms. The connection table is fundamental to all of MDL's file formats.

Figure 2-1 shows the connection table of a simple molecule (alanine) with the various data blocks identified. The connection table corresponds to the following alanine molecule. The atom numbers on the structure correspond to atom numbers in the Ctab. An atom number is assigned according to the order of the atom in the Atom Block.

- Stext (structural text descriptor) block: Used by ISIS and CPSS programs.
- Properties block: Provides for future expandability of Ctab features, while maintaining compatibility with earlier Ctab configurations.

The detailed format for each block outlined above follows:

Note: A blank *numerical* entry on any line should be read as “0” (zero). Spaces are significant and correspond to one or more of the following:

- Absence of an entry
- Empty character positions within an entry
- Spaces between entries; single unless specifically noted otherwise

The Counts Line

```
aaabbblllfffccssssxxrrpppiiimmvvvvv
```

Where:

aaa	= number of atoms (current max 255)*	G
bbb	= number of bonds (current max 255)*	G
lll	= number of atom lists (max 30)*	Q
fff	= (obsolete)	
ccc	= chiral flag: 0=not chiral, 1=chiral	G
sss	= number of stext entries	CP
xxx	= number of reaction components + 1	CP
rrr	= number of reactants	CP
ppp	= number of products	CP
iii	= number of intermediates	CP
mmm	= number of lines of additional properties, including the M END line. No longer supported and default set to 999	G
vvvvv	= Ctab version: 'V2000' or 'V3000'	G

* These limits apply to MACCS-II, REACCS, and the ISIS/Host Reaction Gateway, but *not* to the ISIS/Host Molecule Gateway or ISIS/Desktop.

For example, the counts line in the Ctab shown in Figure 2-1 shows six atoms, five bonds, the CHIRAL flag *on*, and three lines in the properties block:

```
6 5 0 0 1 0          3 V2000
```

The Atom Block

The Atom Block is made up of atom lines, one line per atom with the following format:

```
xxxxx. xxxxyyyyy. yyyyzzzzz. zzzz aaaddcccssshhhbbbvvhHHrrri i i mmmnnnee
```

where the values are described in Table 2-1.

Table 2-1 Meaning of values in the atom block

Field	Meaning	Values	Notes
x y z	atom coordinates		G
aaa	atom symbol	entry in periodic table or L for atom list, A, Q, * for unspecified atom, and LP for lone pair, or R# for Rgroup label	G Q G 3D Rg
dd	mass difference	-3, -2, -1, 0, 1, 2, 3, 4 (0 if value beyond these limits)	G Difference from mass in periodic table. Wider range of values allowed by M ISO line, below. Retained for compatibility with older Ctabs, M ISO takes precedence.
ccc	charge	0 = uncharged or value other than these, 1 = +3, 2 = +2, 3 = +1, 4 = doublet (^), 5 = -1, 6 = -2, 7 = -3	G Wider range of values in M CHG and M RAD lines below. Retained for compatibility with older Ctabs, M CHG and M RAD lines take precedence.
sss	atom stereo parity	0 = not stereo, 1 = odd, 2 = even, 3 = either or unmarked stereo center	G Ignored when read. See stereo notes on page 2-33.
hhh	hydrogen count + 1	1 = H0, 2 = H1, 3 = H2, 4 = H3, 5 = H4	Q H0 means no H atoms allowed unless explicitly drawn. Hn means atom must have <i>n</i> or more H's in excess of explicit H's.
bbb	stereo care box	0 = ignore stereo configuration of this double bond atom, 1 = stereo configuration of double bond atom must match	Q Double bond stereochemistry is considered during SSS only if both ends of the bond are marked with stereo care boxes.

Table 2-1 Meaning of values in the atom block (Continued)

Field	Meaning	Values	Notes
v v v	valence	0 = no marking (default) (1 to 14) = (1 to 14) 15 = zero valence	G Shows number of bonds to this atom, including bonds to implied H's.
H H H	H0 designator	0 = not specified, 1 = no H atoms allowed	CP Redundant with hydrogen count information. May be unsupported in future releases of MDL software.
r r r	reaction component type	reactant = 1, product = 2, intermediate = 3	CP
i i i	reaction component number	0 to (n-1)	CP
m m m	atom-atom mapping number	1 - number of atoms	Rx
n n n	inversion/retention flag	0 = property not applied 1 = configuration is inverted, 2 = configuration is retained,	Rx
e e e	exact change flag	0 = property not applied, 1 = change on atom must be exactly as shown	Rx Q

Note: With Ctab version V2000, the `dd` and `ccc` fields have been superseded by the M ISO, M CHG, and M RAD lines in the properties block, described below. For compatibility, all releases since MACCS-II 2.0, REACCS 8.1, and ISIS 1.0:

- Write appropriate values in both places if the values are in the old range.
- Use the atom block fields if there are no M ISO, M CHG, or M RAD lines in the properties block.

Support for these atom block fields may be removed in future releases of MDL software.

The Bond Block

The Bond Block is made up of bond lines, one line per bond, with the following format:

```
111222tttsssxxrrrccc
```

where the values are described in Table 2-2.

Table 2-2 Meaning of values in the bond block

Field	Meaning	Values	Notes
111	first atom number	1 - number of atoms	G
222	second atom number	1 - number of atoms	G
ttt	bond type	1 = Single, 2 = Double, 3 = Triple, 4 = Aromatic, 5 = Single or Double, 6 = Single or Aromatic, 7 = Double or Aromatic, 8 = Any	Q Values 4 through 8 are for SSS queries only.
sss	bond stereo	Single bonds: 0 = not stereo, 1 = Up, 4 = Either, 6 = Down, Double bonds: 0 = Use x-, y-, z-coords from atom block to determine cis or trans, 3 = Cis or trans (either) double bond	G The wedge (pointed) end of the stereo bond is at the first atom (Field 111 above)
xxx	not used		
rrr	bond topology	0 = Either, 1 = Ring, 2 = Chain	Q SSS queries only.
ccc	reacting center status	0 = unmarked, 1 = a center, -1 = not a center, Additional: 2 = no change, 4 = bond made/broken, 8 = bond order changes 12 = 4+8 (both made/broken and changes); 5 = (4 + 1), 9 = (8 + 1), and 13 = (12 + 1) are also possible	Rx (query only)

The Atom List Block **Q**

Note: Newer programs use the M ALS item in the properties block in place of the atom list block. The atom list block is retained for compatibility, but information in an M ALS item supersedes atom list block information.

Made up of atom list lines, one line per list, with the following format:

```
aaa kSSSSn 111 222 333 444 555
```

where:

aaa	= number of atom (L) where list is attached
k	= T = [NOT] list, F = normal list
n	= number of entries in list; maximum is 5
111. . . 555	= atomic number of each atom on the list
S	= space

The Stext Block **CP**

The Stext Block is made up of two-line entries with the following format:

```
xxxxx. xxxxyyyyy. yyyy  
TTTT. . .
```

where:

x y	= stext coordinate
T	= stext text

The Properties Block

The Properties Block is made up of `mmm` lines of additional properties, where `mmm` is the number in the counts line described above. If a version stamp is present, `mmm` is ignored and the file is read until an `M END` line is encountered. Currently `mmm` is no longer supported and set to 999 as the default.

Most lines in the properties block are identified by a prefix of the form `M XXX` where two spaces separate the `M` and `XXX`. Exceptions are:

- `A aaa`, `V aaa vvvvvv`, and `G aaappp`, which indicate ISIS and CPSS properties: atom alias, atom value, and group abbreviation (called residue in ISIS), respectively. **CP**
- `S SKPnnn` which causes the next `nnn` lines to be ignored.

The prefix: `M END` terminates the properties block.

Variables in the formats can change properties but keep the same letter designation. For example, on the Charge, Radical, or Isotope lines, the “uniformity” of the `vvv` designates a general property identifier. On Sgroup property lines, the `sss` uniformity is used as an Sgroup index identifier.

All lines that are not understood by the program are ignored.

The descriptions below use the following conventions for values in field widths of 3:

<code>n15</code>	number of entries on line; value = 1 to 15
<code>nn8</code>	number of entries on line; value = 1 to 8
<code>nn6</code>	number of entries on line; value = 1 to 6
<code>nn4</code>	number of entries on line; value = 1 to 4
<code>nn2</code>	number of entries on line; value = 1 or 2
<code>nn1</code>	number of entries on line; value = 1
<code>aaa</code>	atom number; value = (1 to number of atoms)

The format for the properties included in this block follows. The format shows one entry; ellipses (`. . .`) indicate additional entries.

Atom Alias **CP**

`A aaa`
`x. . .`

`aaa:` Atom number

`x. . .` Alias text

Atom Value **CP**

V aaa v...

aaa: Atom number

v... Value text

Group Abbreviation **CP**

G aaappp

x...

aaa: Atom number

ppp: Atom number

x... Abbreviation label.

Abbreviation is required for compatibility with CPSS. CPSS allowed abbreviations with only one attachment. The attachment is denoted by two atom numbers, *aaa* and *ppp*. All of the atoms on the *aaa* side of the bond formed by *aaa-ppp* are abbreviated. The coordinates of the abbreviation are the coordinates of *aaa*. The text of the abbreviation is on the following line (*x...*). In current versions of ISIS, abbreviations can have any number of attachments and are written out using the *Sgroup* appendixes. However, any ISIS abbreviations that do have one attachment are also written out in the CPSS-style, again for compatibility with CPSS, but this behavior might not be supported in future versions.

Charge **G**

M CHGnn8 aaa vvv ...

vvv: -15 to +15. Default of 0 = uncharged atom. When present, this property supersedes *all* charge and radical values in the atom block, forcing a 0 charge on all atoms not listed in an M CHG or M RAD line.

Radical **G**

M RADnn8 aaa vvv ...

vvv: Default of 0 = no radical, 1 = singlet (:), 2 = doublet (^), 3 = triplet (^ ^). When present, this property supersedes *all* charge and radical values in the atom block, forcing a 0 (zero) charge and radical on all atoms not listed in an M CHG or M RAD line.

Isotope **G**

M IS0nn8 aaa vvv ...

vvv: Absolute mass differing from natural abundance (as specified by PTABLE.DAT) within the range -18 to +12. When present, this property supersedes *all* isotope values in the atom block. Default (no entry) is natural abundance.

Ring Bond Count **Q**

M RBDnn8 aaa vvv ...

vvv: Number of ring bonds allowed: default of 0 = off, -1 = no ring bonds (r0), -2 = as drawn (r*); 2 = (r2), 3 = (r3), 4 or more = (r4).

Substitution Count **Q**

M SUBnn8 aaa vvv ...

vvv: Number of substitutions allowed: default of 0 = off, -1 = no substitution (s0), -2 = as drawn (s*); 1, 2, 3, 4, 5 = (s1) through (s5), 6 or more = (s6).

Unsaturated Atom **Q**

M UNSnn8 aaa vvv ...

vvv: At least one multiple bond: default of 0 = off, 1 = on.

Link Atom **Q**

M LI Nnn4 aaa vvv bbb ccc ...

vvv, bbb, ccc: Link atom (aaa) and its substituents, other than bbb and ccc, may be repeated 1 to vvv times, (vvv > = 2).

Atom List **Q**

M ALS aaannn e 11112222333344445555...

aaa: Atom number, value = (1 to #atoms).
 nnn: Number of entries on line (16 maximum).
 e: Exclusion, value is T if a 'NOT' list, F if a normal list.
 1111...: Atom symbol of list entry in field of width 4.

Note: This line contains the atom symbol rather than the atom number used in the atom list block. Any data found in this item supersedes data from the atom list block. The number of entries can exceed the fixed limit of *5* in the atom list block entry.

Attachment Point **Rg**

M AP0nn2 aaa vvv ...

vvv: Indicates whether atom aaa of the Rgroup member is the first attachment point (vvv = 1), second attachment point (vvv = 2), both attachment points (vvv = 3); default of 0 = no attachment.

Atom Attachment Order **Rg**

M AAL aaann2 111 v1v 222 v2v ...

aaa: Atom index of the Rgroup usage atom
 nn2: Number of pairs of entries that follow on the line
 111: Atom index of a neighbor of aaa
 v1v: Attachment order for the aaa-111 bond
 222: Atom index of a neighbor of aaa
 v2v: Attachment order for the aaa-222 bond

Note: v1v and v2v are either 1 or 2 for the simple

doubly attached Rgroup member.

This appendix provides explicit attachment list order information for R# atoms. The appendix contains atom neighbor index and atom neighbor value pairs. The atom neighbor value information identifies the atom neighbor index as the *ith* attachment. The implied ordering in V2000 molfiles is by atom index order for the neighbors of Rgroup usage atoms. If atom index order conflicts with the desired neighbor ordering at the R# atom, this appendix allows you to override to this default order.

If v1v=1 and v2v=2, ISIS/Host only writes this appendix if 111 is greater than 222. Note, however, that the attachment values can be written in any order.

Rgroup Label Location **Rg**

M RGPnn8 aaa rrr ...

rrr: Rgroup number, value from 1 to 32, labels position of Rgroup on root.

Rgroup Logic, Unsatisfied Sites, Range of Occurrence **Rg**

M LOGnn1 rrr iii hhh ooo

rrr: Rgroup number, value from 1 to 32.

iii: Number of another Rgroup which must only be satisfied if rrr is satisfied (IF rrr THEN iii).

hhh: RestH property of Rgroup rrr; default is 0 = off, 1 = on. If this property is applied (on), sites labeled with Rgroup rrr may only be substituted with a member of the Rgroup or with H.

ooo: Range of Rgroup occurrence required: n = exactly n, n - m = n through m, > n = greater than n, < n = fewer than n, default (blank) is > 0. Any non-contradictory combination of the preceding values is also allowed; for example: 1, 3-7, 9, >11.

Sgroup Type **Sg**

M STYnn8 sss ttt ...

sss: Sgroup number.

ttt: SUP = superatom, MUL = multiple group, SRU = SRU type, MON = monomer, MER = Mer type, COP = copolymer, CRO = crosslink, MOD = modification, GRA = graft, COM = component, MIX = mixture, FOR = formulation, DAT = data Sgroup, ANY = any polymer, GEN = generic.

Note: For a given Sgroup, an STY line giving its type must appear before any other line that supplies information about it. For a data Sgroup, an SDT line must describe the data field before the SCD and SED lines that contain the data (see Data Sgroup Data below). When a data Sgroup is linked to another Sgroup, the Sgroup must already have been defined.

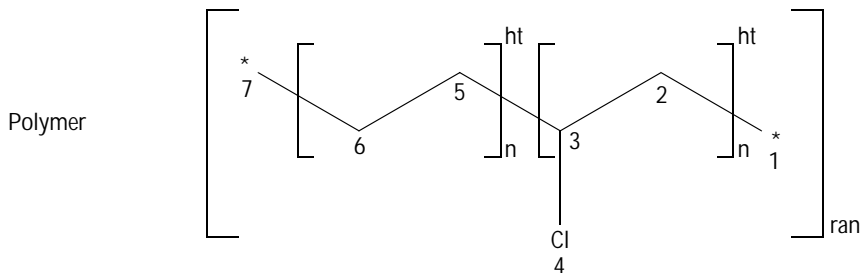
Sgroups can be in any order on the Sgroup Type line. Brackets are drawn around Sgroups with the M SDI lines defining the coordinates.

Sgroup Subtype **Sg**

M SSTnn8 sss ttt ...

ttt: Polymer Sgroup subtypes: ALT = alternating, RAN = random, BLO = block.

Figure 2-2 Ctab organization of an Sgroup structure



```
GSMACCS-1110179110412D 1 0.00374 0.00000 0
```

```
7 6 0 0 0 0          16 V2000
 2.9463 0.3489 0.0000 * 0 0 0 0 0 0
 1.6126 1.1189 0.0000 C 0 0 0 0 0 0
 0.2789 0.3489 0.0000 CI 0 0 3 0 0 0
 0.2789 -1.1911 0.0000 CI 0 0 0 0 0 0
-1.0548 1.1190 0.0000 C 0 0 0 0 0 0
-2.3885 0.3490 0.0000 C 0 0 0 0 0 0
-3.9246 1.1470 0.0000 * 0 0 0 0 0 0
```

```
1 2 1 0 0 0
2 3 1 0 0 0
3 4 1 0 0 0
5 6 1 0 0 0
5 3 1 0 0 0
7 6 1 0 0 0
```

Number of entries on line

```
M STY 3 1 SRU 2 SRU 3 COP
M SST 1 3 RAN
M SLB 3 1 5 2 6 3 7
M SCN 2 1 HT 2 HT
M SAL 1 2 5 6
M SBL 1 2 5 6
M SDI 1 4 -0.6103 1.2969 -0.6103 0.1710
M SDI 1 4 -3.1565 0.1850 -3.1565 1.3110
M SAL 2 3 2 3 4
M SBL 2 2 1 5
M SDI 2 4 2.2794 1.2969 2.2794 0.1709
M SDI 2 4 -0.1657 0.1710 -0.1657 1.2969
M SAL 3 7 1 2 3 4 5 6 7
M SDI 3 4 3.6382 1.6391 3.6382 -1.7685
M SDI 3 4 -4.7070 -1.7685 -4.7070 1.6391
M END
```

Type
Subtype
Label
Connectivity

General
Sgroup
Info

Sgroup 1

Sgroup 2

Sgroup 3

Header block (see Chapter 3)

Counts line

Atom block

Bond block

Atom list block
Stext block

Ctab
block

Sgroup
properties

Sgroup Labels **Sg**

M SLBnn8 sss vvv ...

vvv: Unique Sgroup identifier (for MACCS-II only, the integer label is from 1-512).

Sgroup Connectivity **Sg**

M SCNnn8 sss ttt ...

ttt: HH = head-to-head, HT = head-to-tail, EU = either unknown. Left justified.

Sgroup Expansion **Sg**

M SDS EXPn15 sss ...

sss: Sgroup index of expanded superatoms.

Sgroup Atom List **Sg**

M SAL sssn15 aaa ...

aaa: Atoms in Sgroup sss.

Sgroup Bond List **Sg**

M SBL sssn15 bbb ...

bbb: Bonds in Sgroup sss. (For data Sgroups, bbb's are the containment bonds, for all other Sgroup types, bbb's are crossing bonds.)

Multiple Group Parent Atom List **Sg**

M SPA sssn15 aaa ...

aaa: Atoms in paradigmatic repeating unit of multiple group sss.

Note: To ensure that all current molfile readers consistently interpret chemical structures, multiple groups are written in their fully expanded state to the molfile. The M SPA atom list is a subset of the full atom list that is defined by the Sgroup Atom List M SAL entry.

Sgroup Subscript **Sg**

M SMT sss m. . .

m. . . : Text of subscript Sgroup sss. (For multiple groups, m. . . is the text representation of the multiple group multiplier. For superatoms, m. . . is the text of the superatom label.)

Sgroup Correspondence **Sg**

M CRS sssnn6 bb1 bb2 bb3

bb1, bb2: Crossing bonds that share a common bracket.

bb3: Crossing bond in repeating unit that connect to bond bb1.

Sgroup Display Information **Sg**

M SDI sssnn4 x1 y1 x2 y2

x1, y1, x2, y2: Coordinates of bracket endpoints (FORTRAN format 4F10.4).

Superatom Bond and Vector Information **Sg**

M SBV sss bb1 x1 y1

bb1: Bond connecting to contracted superatom.

x1, y1: Vector for bond bb1 connecting to contracted superatom sss (FORTRAN format 2F10.4).

Data Sgroup Field Description **Sg**

M SDT sss fff. . . fffgghhh. . . hhh i j j j . . .

sss: Index of data Sgroup.

fff. . . fff: 30 character field name (in MACCS-II no blanks, commas, or hyphens).

gg: Field type (in MACCS-II F = formatted, N = numeric, T = text).

hhh. . . hhh: 20-character field units or format.

- ii: Nonblank if data line is a query rather than Sgroup data, MQ = MACCS-II query, IQ = ISIS query, PQ = *program name code* query.
- jjj...: Data query operator (blank for MACCS-II).

Data Sgroup Display Information **Sg**

M SDD sss xxxxx. xxxxyyyyy. yyyy eeefgh i jjj kkk ll m noo

- sss: Index of data Sgroup.
- x, y: Coordinates (2F10.4).
- eee: (Reserved for future use.)
- f: Data display, A = attached, D = detached.
- g: Absolute, relative placement, A = absolute, R = relative.
- h: Display units, blank = no units displayed, U = display units.
- i: (Reserved for future use.)
- jjj: Number of characters to display (1...999 or ALL).
- kkk: Number of lines to display (unused, always 1).
- ll: (Reserved for future use.)
- m: Tag character for tagged detached display (if non-blank).
- n: Data display DASP position (1...9). (MACCS-II only)
- oo: (Reserved for future use.)

Data Sgroup Data **Sg**

M SCD sss d...
M SED sss d...

- d...: Line of data for data Sgroup sss (69 chars per line, columns 12-80)

Note: A line of data is entered as text in 69-character substrings. Each SCD line adds 69 characters to a text buffer (starting with successive SCDs at character positions 1, 70, and 139). Following zero or more

SCDs must be an SED, which may supply a final 69 characters. The SED initiates processing of the buffered line of text: trailing blanks are removed and right truncation to 200 characters is performed, numeric and formatted data are validated, and the line of data is added to data Sgroup *sss*. Left justification is not performed.

A data Sgroup may have more than one line of data, so more than one set of SCD and SED lines can be present for the same data Sgroup. The lines are added in the same order that they are encountered.

If 69 or fewer characters are to be entered on a line, they may be entered with a single SED not preceded by an SCD. On the other hand, if desired a line may be entered to a maximum of 3 SCDs followed by a blank SED that terminates the line. The set of SCD and SED lines describing one line of data for a given data Sgroup must appear together, with no intervening lines for other data Sgroups' data.

Sgroup Hierarchy Information **Sg**

M SPLnn8 ccc ppp . . .

ccc: Sgroup index of the child Sgroup.

ppp: Sgroup index of the parent Sgroup (*ccc* and *ppp* must already be defined via an STY line prior to encountering this line).

Sgroup Component Numbers **Sg**

M SNCnn8 sss ooo . . .

sss: Index of component Sgroup.

ooo: Integer component order (1...256). This limit applies only to MACCS-II.

3D Feature Properties **3D**

M \$3Dnnn

M \$3D. . . See below for information on the properties block of a 3D molfile. These lines must all be contiguous.

End of Block

M END

This entry goes at the end of the properties block and is required for molfiles which contain a version stamp in the counts line.

The Properties Block for 3D Features **3D**

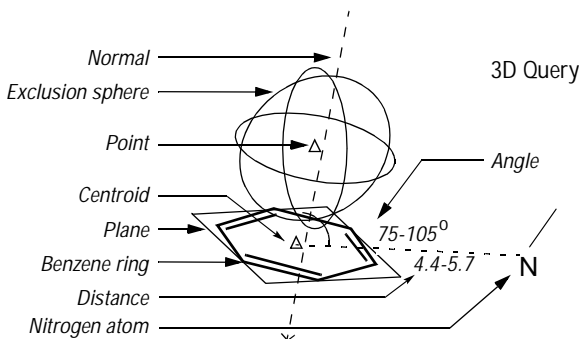
For each 3D feature, the properties block includes:

- One 3D features count line
- One or more 3D features detail lines

The characters M \$3D appear at the beginning of each line describing a 3D feature. The information for 3D features starts in column 7.

Figure 2-3 illustrates the molfile corresponding to the following 3D query:

Figure 2-3 Ctab organization of a 3D query



```

3D Query
MACCS-1110179109553D 1 1.00000 0.00000 0

 8 7 0 0 0 0 18 V2000
 1.0252 0.2892 1.1122 C 0 0 0 0 0 0
-0.4562 0.6578 1.3156 C 0 0 0 0 0 0
-1.4813 0.3687 0.2033 C 0 0 0 0 0 0
-1.0252 -0.2892 -1.1122 C 0 0 0 0 0 0
 0.4562 -0.6578 -1.3156 C 0 0 0 0 0 0
 1.4813 -0.3687 -0.2033 C 0 0 0 0 0 0
 4.1401 -0.1989 1.3456 N 0 0 0 0 0 0
 4.6453 0.5081 1.7417 C 0 0 0 0 0 0

 1 2 1 0 0 0
 2 3 2 0 0 0
 3 4 1 0 0 0
 4 5 2 0 0 0
 5 6 1 0 0 0
 6 1 2 0 0 0
 7 8 1 0 0 0

M $3D 7
M $3D -7 6
M $3D 3
M $3D 6 4 2
M $3D -5 13
M $3D 6 0.0000
M $3D 1 2 3 4 5 6
M $3D -8 7
M $3D 9 10
M $3D -3 6
M $3D 9 11 -2.0000
M $3D-16 12
M $3D 12 1 0 1.5000
M $3D-12 10
M $3D 12 9 7 75.0000 105.0000
M $3D -9 3
M $3D 7 9 4.4000 5.7000
M END
    
```

Header block (see Chapter 3)

Counts line

Atom block

Bond block

Atom list block

Stext block

Ctab block

3D Features Count

Centroid

Plane

Normal to Plane

Point

Exclusion sphere

Angle

Distance

Properties block

3D features count line

The first line in the properties block is the 3D features count line and has the following format:

```
M $3Dnnn
```

where `nnn` is the number of 3D features on a model.

3D features detail lines

The lines following the 3D features count line describe each 3D feature on a model. Each 3D feature description consists of an identification line and one or more data lines:

- The identification line is the first line and contains the 3D feature's type identifier, color, and name.
- Each data line describes the construction of the 3D feature.

Identification line

The 3D feature identification line has the following format:

```
M $3Dfffccc aaa...aaa ttt...ttt
```

where the variables represent:

<code>fff</code>	3D feature type
<code>ccc</code>	Color number (an internal MDL number which is terminal dependent)
<code>aaa...aaa</code>	3D feature name (up to 32 characters)
<code>ttt...ttt</code>	Text comments (up to 32 characters) used by MDL programs (see 3D data constraints on page 2-31)

Table 2-3 lists the 3D feature type identifiers.

Table 2-3 3D feature type identifiers

Identifier	Meaning
-1	Point defined by two points and a distance (in Angstroms)
-2	Point defined by two points and a percentage
-3	Point defined by a point, a normal line, and a distance
-4	Line defined by two or more points (A best fit line if more than two points)
-5	Plane defined by three or more points (A best fit plane if more than three points)
-6	Plane defined by a point and a line
-7	Centroid defined by points
-8	Normal line defined by a point and a plane
-9	Distance defined by two points and a range (in Angstroms)
-10	Distance defined by a point, line, and a range (in Angstroms)
-11	Distance defined by a point, plane, and a range (in Angstroms)
-12	Angle defined by three points and a range (in degrees)
-13	Angle defined by two intersecting lines and a range (in degrees)
-14	Angle defined by two intersecting planes and a range (in degrees)
-15	Dihedral angle defined by 4 points and a range (in degrees)
-16	Exclusion sphere defined by a point and a distance (in Angstroms)
-17	Fixed atoms in the model
<i>nnn</i>	A positive integer indicates atom or atom-pair data constraints

Data line

The 3D feature defines the data line format. Each 3D object is treated as a pseudoatom and identified in the connection table by a number. The 3D object numbers are assigned sequentially, starting with the next number greater than the number of atoms. The data line formats for the 3D feature types are:

Type	Description of Data Line
-1	The data line for a point defined by two points and a distance (Å) has the following format:

```
M $3Di i i j j j d d d d d . d d d d
```

where the variables represent:

i i i	ID number of a point
j j j	ID number of a second point
d d d d d . d d d d	Distance from first point in direction of second point (Å), 0 if not used

The following example shows POINT_1 created from the atoms 1 and 3 with a constraint distance of 2Å.

The first line is the identification line. The second line is the data line.

```
M $3D -1 4 POINT_1
M $3D 1 3 2.0000
```

-2	The data line for a point defined by two points and a percentage has the format:
----	--

```
M $3Di i i j j j d d d d d . d d d d d
```

where the variables represent:

i i i	ID number of a point
j j j	ID number of a second point
d d d d d . d d d d	Distance (fractional) relative to distance between first and second points, 0 if not used

Type **Description of Data Line**

-3 The data line for a point defined by a point, a normal line, and a distance (Å) has the format:

```
M $3Di i i l l l d d d d d . d d d d
```

where the variables represent:

 i i i ID number of a point

 l l l ID number of a normal line

 d d d d d . d d d d Distance (Å), 0 if not used

Note: For chiral models, the distance value is signed to specify the same or opposite direction of the normal.

-4 The data lines for a best fit line defined by two or more points have the following format:

```
M $3Dpppt t t t t t . t t t t
```

```
M $3Di i i j j j . . . z z z . . .
```

where the variables represent:

 p p p Number of points defining the line

 t t t t t . t t t t Deviation (Å), 0 if not used.

 i i i Each i i i , j j j , and z z z is the ID number
 j j j of an item in the model that defines the
 line

 j j j

 . . .

 z z z (to maximum of 20 items per data line)

The following line is defined by the four points 1, 14, 15, and 19 and has a deviation of 1.2Å. The first line is the identification line. The second and third lines are the data lines.

```
M $3D -4 2 N_TO_AROM
```

```
M $3D 4 1.2000
```

```
M $3D 1 14 15 19
```

Type **Description of Data Line**

-5 The data lines for a plane defined by three or more points (a best fit plane if more than three points) have the following format:

```
M $3Dpppttttt. tttt
M $3Di i i j j j . . . zzz
. . .
```

where the variables represent:

ppp	Number of points defining the line
ttttt. tttt	Deviation (Å), 0 if not used.
i i i	Each i i i, j j j, and zzz is the ID number j j j of an item in the model that defines the line
j j j	
. . .	
zzz	(to maximum of 20 items per data line)

The following line is defined by the four points 1, 14, 15, and 19 and has a deviation of 1.2Å. The first line is the identification line. The second and third lines are the data lines.

```
M $3D -5 4 PLANE_2
M $3D 3
M $3D 1 5 14
```

-6 The data line for a plane defined by a point and a line has the following format:

```
M $3Di i i i i i
```

where the variables represent:

i i i	ID number of a point
i i i	ID number of a line

The following plane is defined by the point 1 and the plane 16. The first line is the identification line. The second line is the data line.

```
M $3D -6 3 PLANE_1
M $3D 1 16
```

Type **Description of Data Line**

-7 The data lines of a centroid defined by points have the following format:

```
M $3Dppp
M $3Di i i j j j . . . z z z . . .
```

where the variables represent:

ppp	Number of points defining the centroid
i i i	Each i i i, j j j, and z z z is the ID number j j j of an item in the model that defines the centroid
j j j	
. . .	
z z z	(maximum of 20 items per data line).

The following centroid, ARO_CENTER, is defined by 3 items: 6, 8, and 10. The first line is the identification line. The second and third lines are the data lines.

```
M $3D -7 1 ARO_CENTER
M $3D 3
M $3D 6 8 10
```

-8 The data line for a normal line defined by a point and a plane has the following format:

```
M $3Di i i j j j
```

where the variables represent:

i i i	ID number of a point
j j j	ID number of a plane

The following normal line, ARO_NORMAL, is defined by the point 14 and the plane 15. The first line is the identification line. The second line is the data line.

```
M $3D -8 1 ARO_NORMAL
M $3D 14 15
```

Type	Description of Data Line								
-9	<p>The data line for a distance defined by two points and a range (Å) has the following format:</p> <pre>M \$3Di i i j j j d d d d d . d d d d z z z z z . z z z z</pre> <p>where the variables represent:</p> <table border="0"> <tr> <td style="padding-right: 20px;">i i i</td> <td>ID number of a point</td> </tr> <tr> <td>j j j</td> <td>ID number of a second point</td> </tr> <tr> <td>d d d d d . d d d d</td> <td>Minimum distance (Å)</td> </tr> <tr> <td>z z z z z . z z z z</td> <td>Maximum distance (Å)</td> </tr> </table> <p>The following distance, L, is between items 1 and 14 and has a minimum distance of 4.9Å and a maximum distance of 6.0Å. The first line is the identification line. The second line is the data line.</p> <pre>M \$3D -9 6 L M \$3D 1 14 4.9000 6.0000</pre>	i i i	ID number of a point	j j j	ID number of a second point	d d d d d . d d d d	Minimum distance (Å)	z z z z z . z z z z	Maximum distance (Å)
i i i	ID number of a point								
j j j	ID number of a second point								
d d d d d . d d d d	Minimum distance (Å)								
z z z z z . z z z z	Maximum distance (Å)								
-10	<p>The data line for a distance defined by a point, line, and a range (Å) has the format:</p> <pre>M \$3Di i i l l l d d d d d . d d d d z z z z z . z z z z</pre> <p>where the variables represent:</p> <table border="0"> <tr> <td style="padding-right: 20px;">i i i</td> <td>ID number of a point</td> </tr> <tr> <td>l l l</td> <td>ID number of a line</td> </tr> <tr> <td>d d d d d . d d d d</td> <td>Minimum distance (Å)</td> </tr> <tr> <td>z z z z z . z z z z</td> <td>Maximum distance (Å)</td> </tr> </table>	i i i	ID number of a point	l l l	ID number of a line	d d d d d . d d d d	Minimum distance (Å)	z z z z z . z z z z	Maximum distance (Å)
i i i	ID number of a point								
l l l	ID number of a line								
d d d d d . d d d d	Minimum distance (Å)								
z z z z z . z z z z	Maximum distance (Å)								
-11	<p>The data line for a distance defined by a point, plane, and a range (Å) has the format:</p> <pre>M \$3Di i i j j j d d d d d . d d d d z z z z z . z z z z</pre> <p>where the variables represent:</p> <table border="0"> <tr> <td style="padding-right: 20px;">i i i</td> <td>ID number of a point</td> </tr> <tr> <td>j j j</td> <td>ID number of a plane</td> </tr> <tr> <td>d d d d d . d d d d</td> <td>Minimum distance (Å)</td> </tr> <tr> <td>z z z z z . z z z z</td> <td>Maximum distance (Å)</td> </tr> </table>	i i i	ID number of a point	j j j	ID number of a plane	d d d d d . d d d d	Minimum distance (Å)	z z z z z . z z z z	Maximum distance (Å)
i i i	ID number of a point								
j j j	ID number of a plane								
d d d d d . d d d d	Minimum distance (Å)								
z z z z z . z z z z	Maximum distance (Å)								

Type **Description of Data Line**
 -12 The data line for an angle defined by three points and a range (in degrees) has the following format:

M \$3Di i i j j j kkkdddd. ddddzzzz. zzzz

where the variables represent:

i i i	ID number of a point
j j j	ID number of a second point
kkk	ID number of a third point
dddd. dddd	Minimum degrees
zzzz. zzzz	Maximum degrees

The following angle, THETA1, is defined by the three points: 5, 17, and 16. The minimum angle is 80° and the maximum is 105°. The first line is the identification line. The second line is the data line.

M \$3D-12 5 THETA1
 M \$3D 5 17 16 80.0000 105.0000

-13 The data line for an angle defined by two lines and a range (in degrees) has the following format:

M \$3DI l l mmmdddd. ddddzzzz. zzzz

where the variables represent:

l l l	ID number of a line, mmm ID number of a second line
dddd. dddd	Minimum degrees
zzzz. zzzz	Maximum degrees

THETA2 is defined by the lines 27 and 26 with maximum and minimum angles of 45° and 80°. The first line is the identification line. The second line is the data line.

M \$3D-13 5 THETA2
 M \$3D 27 26 45.0000 80.0000

Type **Description of Data Line**

-14 The data line for an angle defined by two planes and a range (in degrees) has the following format:

```
M $3Di i i j j j d d d d d . d d d d z z z z z . z z z z
```

where the variables represent:

i i i	ID number of a plane
j j j	ID numbers of a second plane
d d d d d . d d d d	Minimum degrees
z z z z z . z z z z	Maximum degrees

-15 The data line for a dihedral angle defined by four points and a range (in degrees) has the following format:

```
M $3Di i i j j j k k k l l l d d d d d . d d d d z z z z z . z z z z
```

where the variables represent:

i i i	ID number of a point
j j j	ID number of a second point
k k k	ID number of a third point
l l l	ID number of a fourth point
d d d d d . d d d d	Minimum degrees
z z z z z . z z z z	Maximum degrees

DIHED1 is defined by the items 7, 6, 4, and 8 with minimum and maximum angles of 45° and 80°, respectively. The first line is the identification line. The second line is the data line.

```
M $3D-15 5 DI HED1
M $3D 7 6 4 8 45.0000 80.0000
```

Type **Description of Data Line**
-16 The data lines for an exclusion sphere defined by a point and a distance (Å) have the following format:

```
M $3Di i i uuuaadddd. dddd  
M $3Dbbbccc... zzz ...
```

where the variables represent:

i i i	ID number of the center of the sphere
uuu	1 or 0. 1 means unconnected atoms are ignored within the exclusion sphere during a search; 0 otherwise
aaa	Number of allowed atoms
dddd. dddd	Radius of sphere (Å)
bbb	Each bbb, ccc, and zzz
ccc	is an ID number of an allowed atom.
...	
zzz	(to maximum of 20 items per data line)

The following exclusion sphere is centered on point 24, has a radius of 5, and allows atom 9 within the sphere. The first line is the identification line. The second and third lines are the data lines.

```
M $3D-16 7 EXCL_SPHERE  
M $3D 24 0 1 5.0000  
M $3D 9
```

Type	Description of Data Line										
-17	<p>The data lines of the fixed atoms have the following format:</p> <pre>M \$3Dppp M \$3Di i i j j j . . . z z z . . .</pre> <p>where the variables represent:</p> <table border="0" style="margin-left: 40px;"> <tr> <td style="padding-right: 20px;">ppp</td> <td>Number of fixed points</td> </tr> <tr> <td>iii</td> <td>Each i i i , j j j , and z z z is an ID number of a fixed atom</td> </tr> <tr> <td>jjj</td> <td></td> </tr> <tr> <td>...</td> <td></td> </tr> <tr> <td>zzz</td> <td>(to maximum of 20 items per data line)</td> </tr> </table> <p>The following examples shows 4 fixed atoms. The first line is the identification line. The second and third lines are the data lines.</p> <pre>M \$3D-17 M \$3D 4 M \$3D 3 7 12 29</pre>	ppp	Number of fixed points	iii	Each i i i , j j j , and z z z is an ID number of a fixed atom	jjj		...		zzz	(to maximum of 20 items per data line)
ppp	Number of fixed points										
iii	Each i i i , j j j , and z z z is an ID number of a fixed atom										
jjj											
...											
zzz	(to maximum of 20 items per data line)										

3D data constraints 3D Q

A positive integer is used as a type identifier to indicate an atom or atom-pair data constraint. Two lines are used to describe a data constraint. The lines have the following format:

```
M $3Dnncccaaa. . . aaabbbbbbbppppppppss. . . sss
M $3Di i i j j j ddd. . . ddd
```

where the variables represent:

nnn	Database-field number
ccc	Color
aaa. . . aaa	Database-field name (up to 30 characters)
bbbbbbbb	/BOX = box-number (source of data) (up to 8 characters)
pppppppp	/DASP = n1, n2 where n1 and n2 are digits from 1-9 (data size and position) (up to 9 characters)

sss. . . sss	/DISP = 3DN (name), 3DV (value), 3DQ (query), NOT (no text)
	First three in any combination to maximum total of 15 characters
iii	ID number of an atom
jjj	ID number of a second atom for atom-pair data, 0 if data is atom data
ddd. . . ddd	Data constraint (based on format from database) (up to 64 characters)
	ISIS 3D data query syntax and MACCS-II 3D data query syntax are not identical. The ISIS data query requires a search operator, a blank space, then one or more operands. For more information on ISIS data query syntax, see the ISIS Help system entries on SBF (Search By Form) or QB (Query Builder) for entering text in a query. For information on MACCS-II data searches, see the <i>MACCS-II Command Language Reference</i> .

Note: For MACCS-II, the atom number 999 stands for all atoms. The MACCS-II wild card character (@) can be used in the data constraints.

The following example shows a numeric data constraint for the field CND0.CHARGE on atom 12. The first line is the identification line. The second line is the data line.

```
M $3D 7 0 CND0. CHARGE
M $3D 12 0 -0.3300 -0.1300
```

The following example shows a numeric data constraint for the field BOND.LENGTH on the atom pair 1 and 4. The first line is the identification line. The second line is the data line.

```
M $3D 9 0 BOND. LENGTH
M $3D 1 4 2.0500 1.8200
```

The following example shows a data constraint allowing any charge value for the field CHARGE on all the atoms. The first line is the identification line. The second line is the data line.

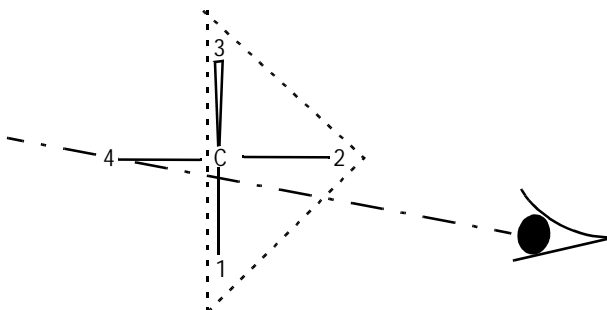
```
M $3D 12 0 CHARGE
M $3D999 0 @
```

Stereo Notes

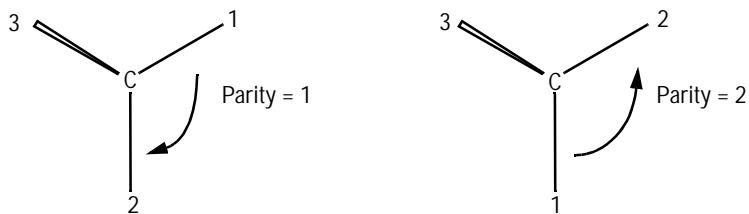
Parity can be illustrated as follows:

Mark a bond attached at a stereo center Up or Down to define the configuration. Number the atoms surrounding the stereo center with 1, 2, 3, and 4 in order of increasing atom number (position in the atom block) (a hydrogen atom should be considered the highest numbered atom, in this case atom 4). View the center from a position such that the bond connecting the highest-numbered atom (4) projects behind the plane formed by atoms 1, 2, and 3.

Note: In the figure, atoms 1, 2, and 4 are all in the plane of the paper, and atom 3 is above the plane.

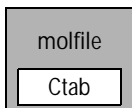


Sighting towards atom number 4 through the plane (123), you see that the three remaining atoms can be arranged in either a clockwise or counterclockwise direction in ascending numerical order.



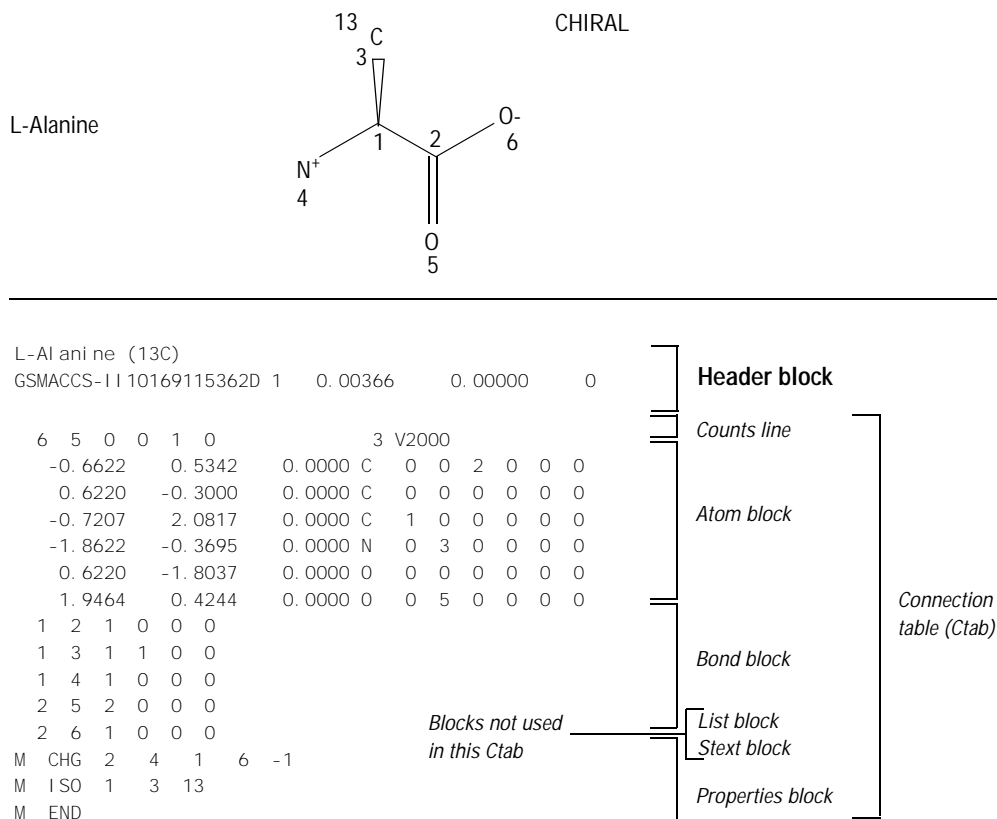
The Ctab lists a parity value of 1 for a clockwise arrangement at the stereo center and 2 for counterclockwise. A center with an Either bond has a parity value of 3. An unmarked stereo center is also assigned a value of 3. The first example above has a parity value of 2.

Molfiles



A molfile consists of a header block and a connection table. Figure 3-1 shows a molfile for alanine corresponding to the following structure:

Figure 3-1 Molfile organization illustrated using alanine



The format for a molfile is:

- Header block: This identifies the molfile with the molecule name, user's name, program, date, and other miscellaneous information and comments
- Ctab block (described in Chapter 2)

The detailed format for the header block follows.

The Header Block

Line 1: Molecule name. This line is unformatted, but like all other lines in a molfile may not extend beyond column 80.

Caution: This line must not contain any of the reserved tags that identify any of the other CTAB file types such as \$MDL (RGfile), \$\$\$\$ (SDfile record separator), \$RXN (rxnfile), or \$RDFILE (RDfile headers).

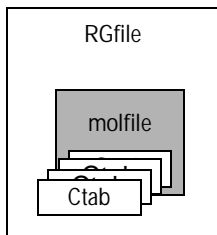
Line 2: User's first and last initials (I), program name (P), date/time (M/D/Y,H:m), dimensional codes (d), scaling factors (S, s), energy (E) if modeling program input, internal registry number (R) if input through MDL form. This line has the format:

```
(FORTRAN:  I I P P P P P P P M D D Y Y H H m d d S S s s s s s s s s s s E E E E E E E E E E R R R R R R R R  
A2  A8    <--A10--->A2I 2   F10.5     F12.5     I 6   )
```

A blank line can be substituted for line 2.

Line 3: A line for comments. If no comment is entered, a blank line must be present.

RGfiles



The format of an RGfile (Rgroup query file) is shown below. Lines beginning with \$ define the overall structure of the Rgroup query; the molfile header block is embedded in the Rgroup header block.

In addition to the primary connection table (Ctab block) for the root structure, a Ctab block defines each member (*m) within each Rgroup (*r).

```

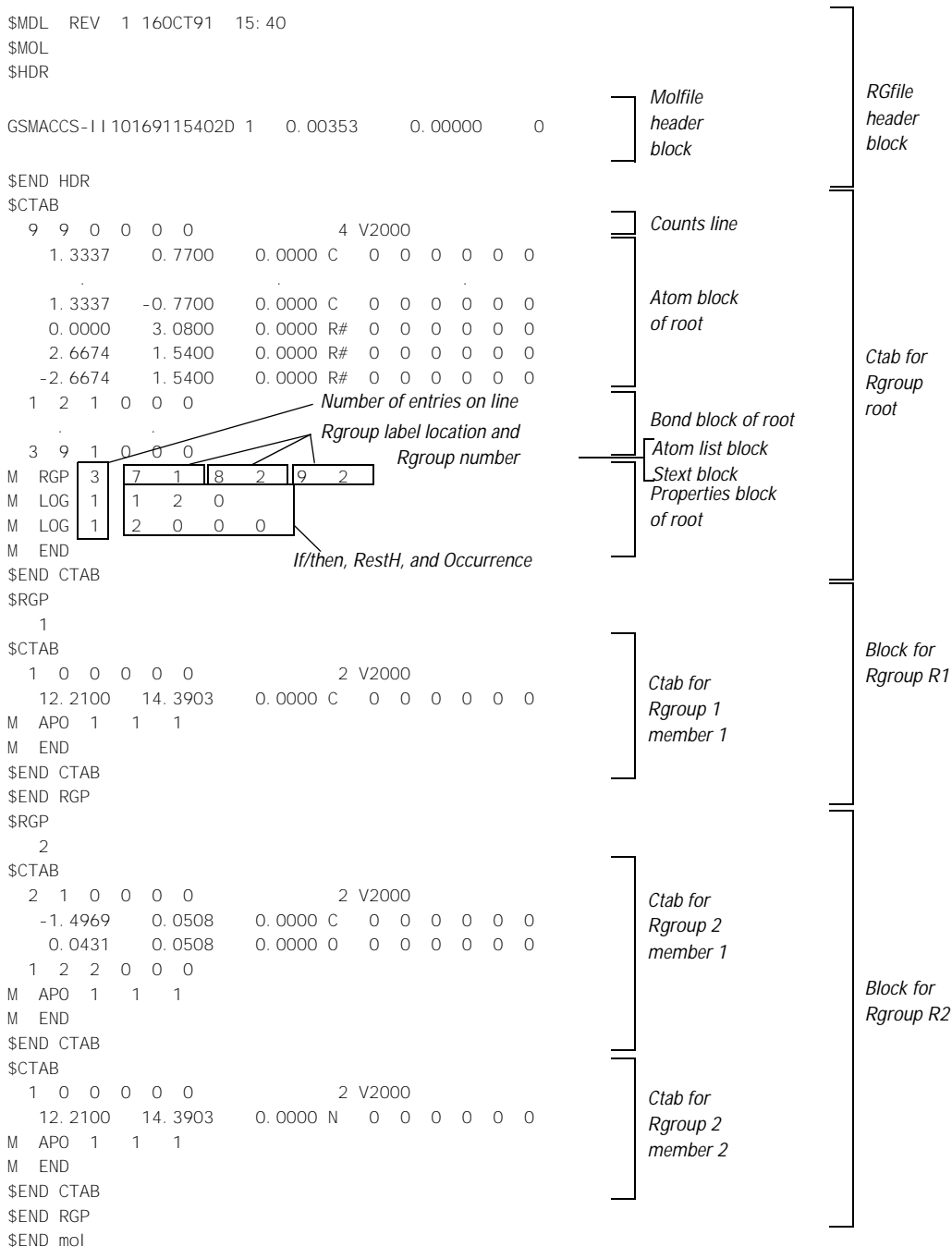
$MDL REV 1 date/time
$MOL
$HDR
[Mol file Header Block (see Chapter 3) = name, pgm info, comment]
$END HDR
$CTAB
[Ctab Block (see Chapter 2) = count + atoms + bonds + lists + props]
$END CTAB
$RGP
  rrr [where rrr = Rgroup number]
    $CTAB
      *m [Ctab Block]
    $END CTAB
  $END RGP
$END mol
  
```

where:

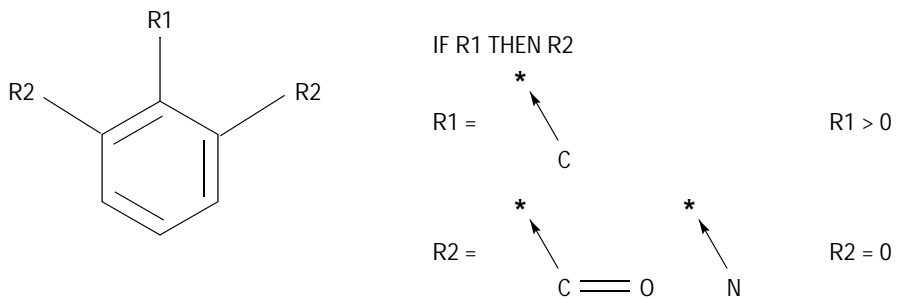
*r (Rgroup) is repeated to a maximum of 32

*m (member) is repeated to a maximum of 255 total atoms and bonds per Rgroup

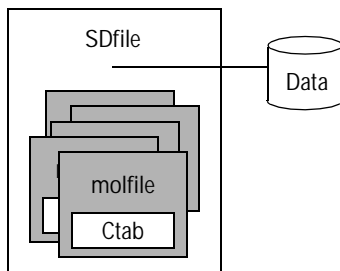
Figure 4-1 Example of an RGfile (Rgroup query file)



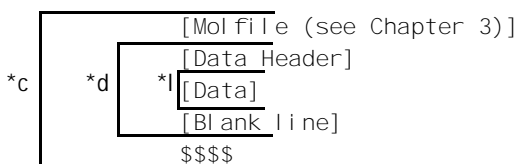
The RGfile shown in Figure 4-1 corresponds to the following Rgroup query:



SDfiles



An SDfile (structure-data file) contains the structural information and associated data items for one or more compounds. An example of an SDfile is shown in Figure 5-1. The format is:



where:

- *l is repeated for each line of data
- *d is repeated for each data item
- *c is repeated for each compound

A *[Molfile]* block has the molfile format described in Chapter 3 or Chapter 10.

A *[Data Header]* (one line) precedes each item of data, starts with a *greater than* (>) sign, and contains at least one of the following:

- The field name enclosed in angle brackets. For example: <melting point>
- The field number, DT*n*, where *n* represents the number assigned to the field in a MACCS database

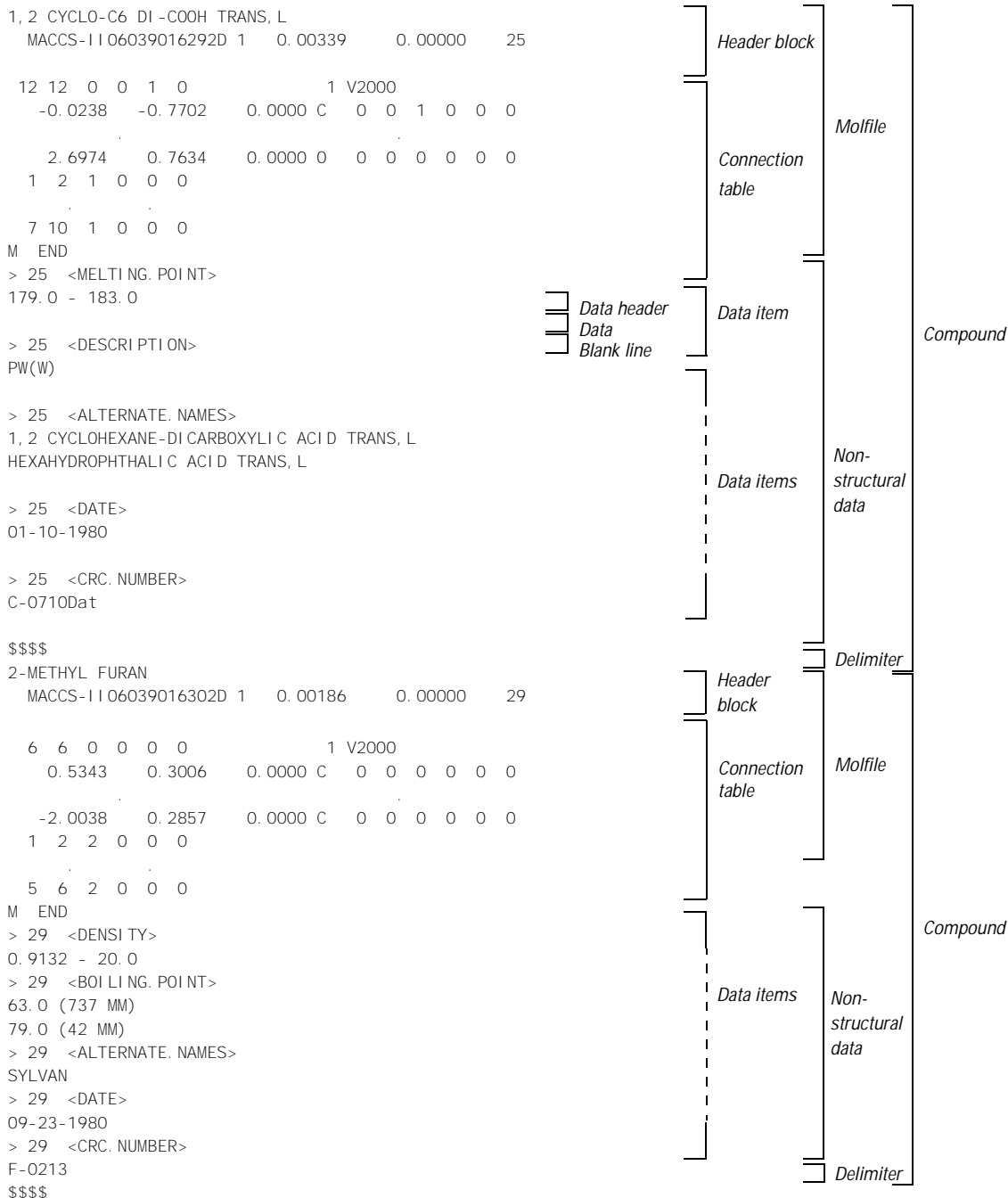
Optional information for the data header includes:

- The compound's external and internal registry numbers. External registry numbers must be enclosed in parentheses.
- Any combination of information

The following are examples of valid data headers:

```
> <MELTI NG. POI NT>  
> 55      (MD-08974)      <BOI LI NG. POI NT>   DT12  
> DT12   55  
> (MD-0894)  <BOI LI NG. POI NT>   FROM ARCHI VES
```

Figure 5-1 Example of an SDfile



A *[Data]* value may extend over multiple lines containing up to 200 characters each. A blank line terminates each data item.

A line containing four dollar signs (\$\$\$\$) terminates each complete data block describing a compound.

A datfile (data file) is effectively an SDfile with no *[Molfile]* descriptions or \$\$\$ delimiters. The *[Data Header]* in a datfile must include either an external or internal registry number in addition to a field name or number.

SDfile after a CFS search

After a conformationally flexible substructure (CFS) search, the following format information is appended by ISIS/Base PL to your SDfile after the connection table:

- Query information (M \$3D appendix lines added to embedded molfile)
- CFS generated data (*DATA)
- MAPPED ATOMS and BONDS

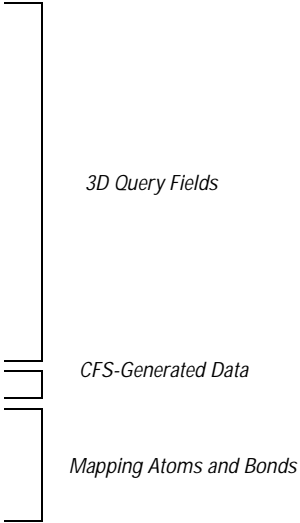
This information describes, for example, how query atoms are mapped, the atom coordinates in models, and what is fitted during a CFS search.

Figure 5-2 Example of SDfile with appended CFS query information

```
M CHG 2 14 -1 16 1
M $3D 5
M $3D -9 3
M $3D 13 18 6.3000 8.3000
M $3D -9 3
M $3D 18 9 3.1000 5.1000
M $3D -9 3
M $3D 18 4 2.4000 4.4000
M $3D -9 3
M $3D 13 9 2.8000 4.8000
M $3D -9 3
M $3D 13 4 3.1000 5.1000
M END
> 31 <*DATA>
Method = Derivative

> 31 <MAPPED ATOMS AND BONDS>
(8 13 14 3 9 4 18) (12 13 7 8)

$$$$
```

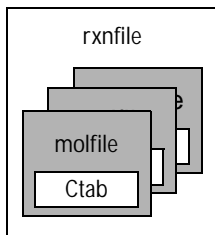


3D Query Fields

CFS-Generated Data

Mapping Atoms and Bonds

Rxnfiles



Rxnfiles contain structural data for the reactants and products of a reaction. An example rxnfile for a simple reaction is shown in Figure 6-1. The format is:

```

      [Rxnfile Header]
      rrrppp
*r --- $MOL
      [Molfile of reactant]
*p --- $MOL
      [Molfile of product]
  
```

where:

*r is repeated for each reactant
 *p is repeated for each product

Header Block

Line 1: \$RXN in the first position on this line identifies the file as a reaction file.

Line 2: A line which is always blank.

Line 3: The program name and version (P), date/time (M/D/Y,H:m), and reaction registry number (R). This line has the format:

```
PPPPPPPPMDDYYHHmmRRRRRRRR
```

(FORTRAN: A14 <--A12--> 18)

A blank line can be substituted for line 3.

Line 4: A line for comments. If no comment is entered, a blank line must be present.

Reactants/Products

A line identifying the number of reactants and products, in that order. The format is:

rrrppp

where the variables represent:

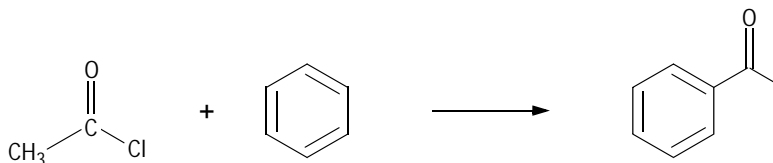
rrr Number of reactants

ppp Number of products

Molfile Blocks

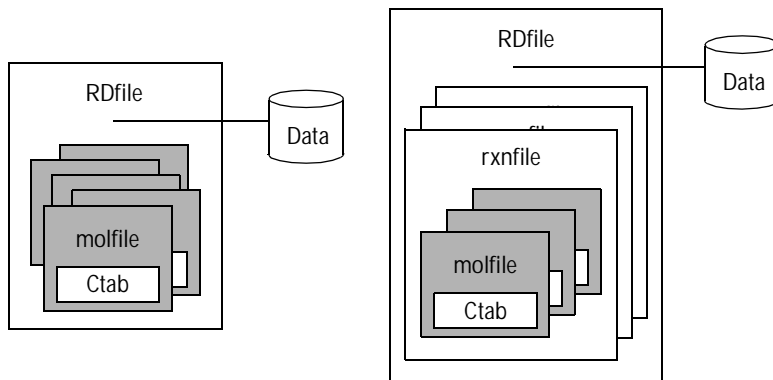
A series of blocks, each starting with \$MOL as a delimiter, giving the molfile for each reactant and product in turn. The molfile blocks are always in the same order as the molecules in the reaction; reactants first and products second.

The rxnfile in Figure 6-1 corresponds to the following reaction:



Note: MACCS-II cannot read or write connection tables for reactions.

RDfiles



An RDfile (reaction-data file) consists of a set of editable “records.” Each record defines a molecule or reaction, and its associated data. An example RDfile incorporating the rxnfile described in Chapter 6 is shown in Figure 7-1. The format for an RDfile is:

```

[RDfile Header]
*r [Molecule or Reaction Identifier]
  *d [Data-field Identifier]
    [Data]

```

where:

- *d is repeated for each data item
- *r is repeated for each reaction or molecule

Each logical line in an RDfile starts with a keyword in column 1 of a physical line. One or more blanks separate the first argument (if any) from the keyword. The blanks are ignored when the line is read. After the first argument, blanks are significant.

An argument longer than 80 characters breaks at column 80 and continues in column 1 of the next line. (The argument may continue on additional lines up to the physical limits on text length imposed by the database.)

The RDfile must not contain any blank lines except as part of embedded molfiles, rxnfiles, or data. An identifier separates records.

RDfile Header

- Line 1:* \$RDFILE 1: The *[RDfile Header]* must occur at the beginning of the physical file and identifies the file as an RDfile. The version stamp "1" is intended for future expansion of the format.
- Line 2:* \$DATM: Date/time (M/D/Y, H:m) stamp. This line is treated as a comment and ignored when the program is read.

Molecule and Reaction Identifiers

A *[Molecule or Reaction Identifier]* defines the start of each complete record in an RDfile. The form of a *molecule* identifier must be one of the following:

```
$MFMT [$MI REG internal-regno [$MEREG external-regno]] embedded mol file
$MI REG internal-regno
$MERE G external-regno
```

where:

- \$MFMT defines a molecule by specifying its connection table as a molfile
- \$MIREG *internal-regno* is the internal registry number (sequence number in the database) of the molecule
- \$MERE G *external-regno* is the external registry number of the molecule (any uniquely identifying character string known to the database, for example, CAS number)
- Square brackets (`[]`) enclose optional parameters
- An embedded molfile (see Chapter 3) follows immediately after the \$MFMT line

The forms of a *reaction* identifier closely parallel that of a molecule:

```
$RFMT [$RI REG internal -regno [$REREG external -regno]] embedded rxnfile
$PCRXN [$RI REG internal -regno [$REREG external -regno]] embedded CPSS rxnfile CP
$RI REG internal -regno
$REREG external -regno
```

where:

- \$RFMT defines a reaction by specifying its descriptor as a rxnfile and \$PCRXN **CP** defines a reaction by specifying its descriptor as a CPSS-style rxnfile
- \$RIREG *internal-regno* is the internal registry number (sequence number in the database) of the reaction
- \$REREG *external-regno* is the external registry number of the reaction (any uniquely identifying character string known to the database)
- Square brackets ([]) enclose optional parameters
- An embedded rxnfile (see Chapter 6) follows immediately after the \$RFMT line, and an embedded CPSS-style rxnfile follows immediately after the \$PCRXN **CP** line

Data-field Identifier

The *[Data-field Identifier]* specifies the name of a data field in the database. The format is:

```
$DTYPE field name
```

Data

Data associated with a field follows the field name on the next line and has the form:

```
$DATUM datum
```

The format of *datum* depends upon the data type of the field as defined in the database. For example: integer, real number, real range, text, molecule regno.

For fields whose data type is “molecule regno,” the *datum* must specify a molecule and, with the exception noted below, use one of the formats defined above for a molecular identifier. For example:

\$DATUM \$MFMT embedded mol file

\$DATUM \$MREG external -regno

\$DATUM \$MI REG internal -regno

In addition, the following special format is accepted:

\$DATUM molecule-identifier

Here, *molecule-identifier* acts in the same way as *external-regno* in that it can be any text string known to the database that uniquely identifies a molecule. (It is usually associated with a data field different from the *external-regno*.)

Figure 7-1 Example of a reaction RDfile

```

$RDFILE 1
$DATM 10/17/91 10:41
$RFMT $RI REG 7439
$RXN

  REACCS81 1017911041 7439

  2 1
$MOL

  REACCS8110179110412D 1 0.00380 0.00000 315

  4 3 0 0 0 0 0 0 0 0 0
  1 4 1 0 0 0 4
$MOL

  REACCS8110179110412D 1 0.00371 0.00000 8

  6 6 0 0 0 0 0 0 0 0 0
  5 6 2 0 0 0 2
$MOL

  REACCS8110179110412D 1 0.00374 0.00000 255

  9 9 0 0 0 0 0 0 0 0 0
  6 9 2 0 0 0 2
$DTYPE rxn: VARIATION(1): rxnTEXT(1)
$DATUM CrCl3
$DTYPE rxn: VARIATION(1): LI TTEXT(1)
$DATUM A G Repin, Y Y Makarov-Zemlanski i, Zur Russ Fiz-Chim, 44,
p. 2360, 1974

$DTYPE rxn: VARIATION(1): CATALYST(1): REGNO
$DATUM $MFMT $MI REG 688

  REACCS8110179110412D 1 0.00371 0.00000 0

  4 3 0 0 0 0 0 0 0 0 0
  1 4 1 0 0 0 0
$DTYPE rxn: VARIATION(1): PRODUCT(1): YIELD
$DATUM 70.0

$RFMT $RI REG 8410
$RXN

  REACCS81 1017911041 8410

  2 1
$MOL

```

Rxnfile header
 #Reactants
 #Products
 Molfile for first reactant
 Molfile for second reactant
 Molfile for product
 RDfile header
 Mol/Rxn identifier
 Data block for reaction
 Start of next record

Atom Limit Enhancements

The formats presented in this chapter were added to support the chemical representation enhancements of ISIS 2.0 Desktop.

Phantom Extra Atom

The format for phantom extra atom information is as follows:

```
M PXA aaaxxxx. xxxxyyyy. yyyyzzzz. zzzz H e...
```

where:

aaa	= Index of (real) atom for attachment
xyz	= Coordinates for the added atom
H	= Atom symbol
e...	= Additional text string (for example, the superatom label)

The FORTRAN format for the phantom extra atom entry is as follows:

```
(I 4, 4F10. 4, 1X, A3, 1X, A)
```

The bond to the added phantom atom is added as a crossing bond to the outermost Sgroup that contains atom *aaa*. Note this appendix supplies coordinates and up to 35 characters of 'label' that can be used for the ISIS/Desktop superatom conversion mechanism. The ISIS/Desktop uses this appendix to register hydrogen-only superatoms, which are often used as superatom leaving groups on the desktop, but which cannot be directly registered into host database. The hydrogen-only leaving groups are converted to PXA appendices for registration, and converted back when ISIS/Desktop reads the structure.

The following are limitations on phantom extra atom:

- Superatom nesting cases
- No bonded phantom atom-phantom atom support

Superatom Attachment Point

The format for superatom attachment point is as follows:

```
M  SAP sssnn6 iii ooo cc
```

where:

sss	= Index of superatom Sgroup
nn6	= Number of <code>iii,ooo,c</code> entries on the line (6 maximum)
iii	= Index of the attachment point atom
ooo	= Index of atom in <code>sss</code> that leaves when <code>iii</code> is substituted
cc	= 2 character attachment identifier (for example, 'H' or 'T' for head/tail). No validation of any kind is performed, and ' ' is allowed. ISIS/Desktop uses the first character as the ID of the leaving group to attach if the bond between <code>ooo</code> and <code>iii</code> is deleted, and uses the second character to indicate the sequence polarity: <code>l</code> for left, <code>r</code> for right, and <code>x</code> for none (a crosslink).

The bond `iii-ooo` is either a sequence bond, a sequence crosslink bond, or a bond to a leaving group that terminates a sequence or caps a crosslink bond. In some cases, this bond may have been deleted by the user, probably to perform a substructure search. In this case, `ooo` will be 0. If the leaving group attached to `iii` consists of only a hydrogen, the leaving group will be replaced by a Phantom Extra Atom, as previously described. In this case, `iii` is set equal to `ooo` as a signal to ISIS/Desktop that a hydrogen-only leaving group must be reattached to `iii`.

The FORTRAN format for the superatom attachment point entry is as follows:

```
(I4, I4, 1X, A2)
```

An attachment point entry is one `iii,ooo,cc` triad.

Multiple M SAP lines are permitted for each superatom Sgroup to the maximum of the atom attachment limit. The order of the attachment entries is significant because the first `iiii,ooo,c` becomes the first connection made when drawing to the collapsed superatom, and so forth.

Superatom Class

The format for superatom class is as follows:

```
M SCL sss d...
```

where:

sss = Index of superatom Sgroup

d... = Text string (for example, PEPTIDE, ...) 69 characters maximum

This appendix identifies the class of the superatom Sgroup. It distinguishes, for example, peptide groups from nucleotides.

Large REGNO

The format for the regno appendix is as follows:

```
M REG r...
```

where:

r... =Free format integer regno

This appendix supports overflow of the I6 regno field in the molfile header. If this appendix is present, the value of the regno in the molfile header is superceded.

Sgroup Bracket Style

The format for the Sgroup bracket style is as follows:

```
M SBTnn8 sss ttt ...
```

where:

sss = Index of Sgroup

ttt = Bracket display style: 0 = default, 1 = curved
(parenthetic) brackets

This appendix supports altering the display style of the Sgroup brackets.

Moving CTfiles On and Off the Clipboard in ISIS

Clipboard Objects

The two objects named here as SK and mSK are used to move MDL sketches on and off the clipboard in ISIS. The names and contents of these with respect to the PC (MS Windows), Macintosh, and SGI (Motif) platforms are summarized in Table 9-1 and described in the *ISIS Sketch File Formats* document. The additional object, CT, is also introduced. This contains structural information in CTfile format to facilitate structure exchanges between ISIS and non-MDL applications. The object, mSK, is not meaningful to platforms such as SGI, because Motif lacks a metafile format like the Macintosh or MS Windows metafile for storing drawing commands.

Table 9-1 ISIS clipboard objects-names and content

Clipboard Object	MS Windows Clipboard Format	Macintosh Scrap Type	SGI Motif Clip-board Format	Contents	Available in ISIS version
SK	MDLSK	swsD	MDL_SKETCH	Buffered MDL sketch file	1.0 and up
CT	MDLCT	swsC	MDL_MOL	Buffered MDL CTfile (molfile, RGfile or rxnfile)	1.01 and up
mSK	CF_METAFIL EPICT	PICT		Picture with MDL sketch embedded	1.0 and up

ISIS will look for the objects listed in Table 9-1 in the order SK, CT, mSK and will take the first available for the image. The metafile, mSK, cannot be distinguished until after it is read from the clipboard, because the embedded file is not identified.

Note: CT has variable length lines. Each line is prefixed with one byte containing the length of the line. Thus, a blank line contains one byte of zero.

Hints on Creating a Reader/Writer For CT

Separate input/output routines from the CTfile interpreter.

Use open/read/close routines to read the contents of the buffer from the clipboard line by line.

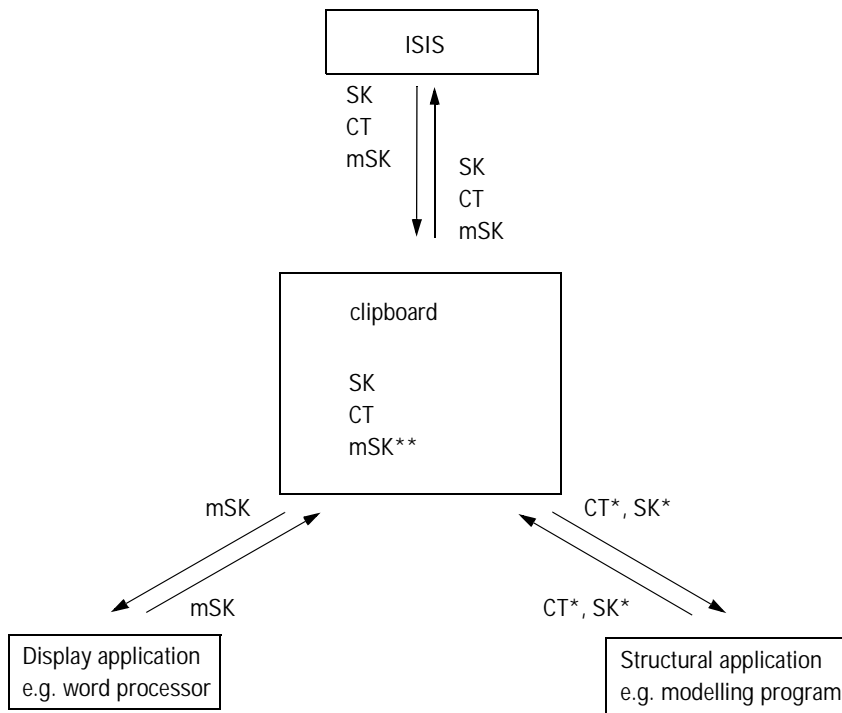
Copying *from* the Clipboard

Look for CT on the clipboard. If present and the first line contains:

- "\$RXN", the file is a rxnfile
- "\$MDL", the file is an RGfile
- Otherwise, the file is a molfile

Alternatively, you can develop your own procedure for reading a sketch file (SK* in Figure 9-1).

Figure 9-1 Transfer options



* Employing user-supplied file reader or writer

** Except SGI

Copying *to* the clipboard

Clear the clipboard of any existing data.

You may choose from among the following options recognizable by ISIS:

- Post a CT containing a buffered CTfile (rxnfile, RGfile or molfile) (with Version 1.01 or later of ISIS).
- Post an SK containing a buffered sketch file.
- Post your own rendering as a metafile or PICT image (PC and Macintosh, respectively) recognizable only by ISIS/Draw. However, this does not preserve the chemistry.

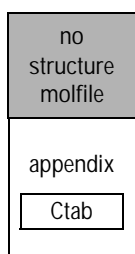
Sample Code For Copying or Pasting a CTfile in MS Windows

```
/* cutpaste.c */
extern HWND hwnd;          /* handle to application's main window */
static int ctFormat;
/*-----*/
InitClipboard()
{
    ctFormat = RegisterClipboardFormat("MDLCT");
} /*-----*/
CopyToClipboard(HANDLE ghCTBuffer)
/*ghCTBuffer is a global handle to a buffer containing the ASCII ctfile. Do not
delete it because it becomes the property of the clipboard after the
SetClipboardData() call. */
{
    if (OpenClipboard(hwnd)) {
        EmptyClipboard();
        SetClipboardData(ctFormat, ghCTBuffer);
        CloseClipboard();
    }
}
/*-----*/ HANDLE
PasteFromClipboard()
{
    HANDLE ghCTBuffer = NULL;
    if (IsClipboardFormatAvailable(ctFormat)) {
        if (OpenClipboard(hwnd)) {
            ghCTBuffer = GetClipboardData(ctFormat);
            CloseClipboard();
        }
    }
    /*ghCTBuffer is a global handle to a buffer containing the ASCII ctfile. It is
still the property of the clipboard so do not delete or alter it. */
    return(ghCTBuffer);
} /*-----*/
```

PART II

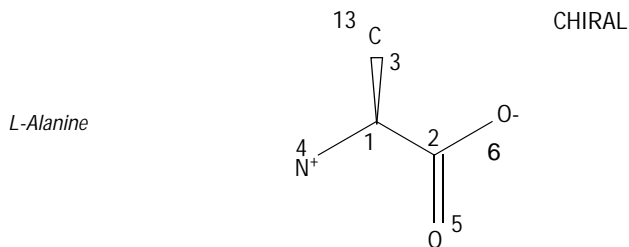
Extended File Formats

The Extended Molfile Format



The extended (V3000) molfile consists of a regular molfile “no structure” followed by a single molfile appendix that contains the body of the connection table (Ctab). Figure 10-1 shows both an alanine structure and the extended molfile corresponding to it. See Figure 2-1 for the V2000 version of this same structure.

Figure 10-1 Extended molfile organization illustrated using alanine



```

L-Alanine
GSMACCS-II07129516502D 1 0.00366 0.00000 0
Figure 1, J. Chem. Inf. Comput. Sci., Vol 32, No. 3., 1992
 0 0 0 0 0 999 V3000
M V30 BEGIN CTAB
M V30 COUNTS 6 5 0 0 1
M V30 BEGIN ATOM
M V30 1 C -0.6622 0.5342 0 0 CFG=2
M V30 2 C 0.6622 -0.3 0 0
M V30 3 C -0.7207 2.0817 0 0 MASS=13
M V30 4 N -1.8622 -0.3695 0 0 CHG=1
M V30 5 O 0.622 -1.8037 0 0
M V30 6 O 1.9464 0.4244 0 0 CHG=-1
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 1 1 2
M V30 2 1 1 3 CFG=1
M V30 3 1 1 4
M V30 4 2 2 5
M V30 5 1 2 6
M V30 END BOND
M V30 END CTAB
M END

```

Header block

Comments line

Counts line

Atom block

Bond block

Sgroup block

Rgroup block

3D block

Connection table Ctab

Blocks not used
in this Ctab

Note that the "no structure" is flagged with the "V3000" instead of the "V2000" version stamp.

There are two other changes to the header in addition to the version:

- The number of appendix lines is always written as 999, regardless of how many there actually are. (All current readers will disregard the count and stop at “M END”.)
- The “dimensional code” is maintained more explicitly. Thus “3D” really means 3D, although “2D” will be interpreted as 3D if any non-zero Z coordinates are found.

Unlike the V2000 molfile, the V3000 extended Rgroup molfile has the same header format as a non-Rgroup molfile.

Note: Do not create a molfile with a pre-V3000 Rgroup header (“\$MDL”, and so forth) but with V3000 Ctab blocks. This is not allowed. A pre-V2000 Rgroup molfile can only have embedded molfiles that are also pre-V3000 versions, for example, the version is either “V2000” or “ ”.

Specifications For Atom and Bond Descriptions

The general syntax of an entry is:

```
M V30 key posval posval ... [keyword=value] [keyword=value] ...
```

or

```
M V30 BEGIN key [blockname]
```

```
M V30 posval posval ... keyword=value keyword=value ...
```

```
...
```

```
M V30 END key
```

Each line must begin with “M V30 ” with the two blank spaces after M and one blank space after 30. Following this is a list of zero or more required positional values (posval). Optional values may follow which use a ‘KEYWORD=value’ format. Items are separated by white space. There can also be white space preceding the first item. Trailing white space is ignored.

The value of a keyword can be a list containing two or more values:

```
KEYWORD=(N val 1 val 2 ... val N)
```

where N specifies the number of items that follow.

Values (posval, value, or val 1, and so forth) can be strings. Strings that contain blank spaces or start with left parenthesis or double quote, must be surrounded by double quotes. A double quote may be entered literally by doubling it.

Each entry is one line of no more than 80 characters. To allow continuation when the 80-character line is too short, use a dash (-) as the last character. When read, the line is concatenated with the next line by removing the dash and stripping the initial "M V30 " from the following line. For example:

```
M V30 10 20 30 " abc-  
M V30   def"
```

is read as:

```
M V30 10 20 30 " abc  def"
```

Generally, each section of the molfile is enclosed in a *block* that consists of lines such as:

```
M V30 BEGIN key [blockname]  
...  
M V30 END key
```

The 'key' value defines the kind of block, for example, CTAB, ATOM, or BOND. Depending upon the type of block, there may or may not be values on the BEGIN line.

Conventions

The new format conventions used in this chapter are as follows:

UPPERCASE	Literal text, to be entered as shown. Only the position of "M V30 " is significant; white space may be added anywhere else to improve readability. Note that <i>both</i> lower- and uppercase characters, or any combination of them, are acceptable for literals. They are shown here in uppercase only for readability.
lowercase	A token, which is defined elsewhere.
[]	An optional item. Do not include the brackets.
[]*	An optional item, where there may be zero, one, two, or more of the item.
	Separates two or more options, only one of which is valid.
/	Separates two or more items. Either or both may appear in any order.

`}` Braces are used for grouping. They indicate indefinite or definite repeat.

The Extended Connection Table

The features of the extended connection table are described in this section.

CTAB block

A Ctab block defines the basic connection table, which is defined as:

```
M V30 BEGIN CTAB [ctabname]
counts-line
atom-block
[bond-block]
[sgroup-block]
[3d-block]
[link-line]*
M V30 END CTAB
```

The atom block, like the counts line, is required. The Sgroup block, 3D block, and link lines may occur in any order after the atom and bond blocks. The counts line, atom block, and bond block must appear in the order indicated.

Counts line

A counts line is required, and must be first. It specifies the number of atoms, bonds, 3D objects, and Sgroups. It also specifies whether or not the CHIRAL flag is set. Optionally, the counts line can specify molregno. This is only used when the regno exceeds 999999 (the limit of the format in the molfile header line). The format of the counts line is:

```
M V30 COUNTS na nb nsg n3d chiral [REGNO=regno]
```

where:

```
na      = number of atoms
nb      = number of bonds
nsg     = number of Sgroups
n3d     = number of 3D constraints
chiral  = 1 if molecule is chiral, 0 if not
regno   = molecule or model regno
```

Atom block

An atom block specifies all node information for the connection table. It must precede the bond block. It has the following format:

```
M V30 BEGIN ATOM
M V30 i n d e x  t y p e  x  y  z  a a m a p  -
M V30 [CHG=val ] [RAD=val ] [CFG=val ] [MASS=val ] -
M V30 [VAL=val ] -
M V30 [HCOUNT=val ] [STBOX=val ] [I NVRET=val ] [EXACHG=val ] -
M V30 [SUBST=val ] [UNSAT=val ] [RBCNT=val ] -
M V30 [ATTCHPT=val ] -
M V30 [RGROUPS=(nval s val [val ...])] -
M V30 [ATTCHORD=(nval s nbr1 val 1 [nbr2 val 2 ...])] -
...
M V30 END ATOM
```

The values are described in Table 10-1.

Table 10-1 Meaning of values in the atom block

Field	Meaning	Values	Notes
i n d e x	Atom index	Integer > 0	Identifies atoms. The actual value of the index does not matter as long as each index is unique to each atom. However, extremely large numbers used as indexes can cause the program to fail to allocate memory for the correspondence array.
t y p e	Atom type	Type = reserved atom <i>or</i> atom <i>or</i> [NOT] ['atom, atom,...'] where reserved atom = R# = Rgroup A = "any" atom Q = any atom but C or H * = "star" atom	A character string. If the string contains white space, it must be quoted. It can be a single atom or an atom list enclosed in square brackets with an optional preceding NOT.
x y z	Atom coordinates	Atom = character string Angstroms	For example, 'C' or 'Cl'

Table 10-1 Meaning of values in the atom block (Continued)

Field	Meaning	Values	Notes
aamap	Atom-atom mapping	0 = no mapping > 0 = mapped atom	Reaction property
CHG	Atom charge	Integer 0 = none (default)	Same range as V2000.
RAD	Atom radical	0 = none (default) 1 = singlet 2 = doublet 3 = triplet	
CFG	Stereo configuration	0 = none (default) 1 = odd parity 2 = even parity 3 = either parity	
MASS	Atomic weight	Integer > 0	Default = natural abundance
VAL	Valence	Integer > 0 <i>or</i> 0 = none (default) -1 = zero	Abnormal valence
HCOUNT	Query hydrogen count	Integer > 0 <i>or</i> 0 = not specified (default) -1 = zero	Same maximum as V2000.
STBOX	Stereo box	0 = ignore the configuration of this double bond atom (default) 1 = consider the stereo configuration of this double bond atom	Both atoms of a double bond must be marked to search double bond stereochemistry
I NVRET	Configuration inversion	0 = none (default) 1 = configuration inverts 2 = configuration retained	Reaction property
EXACHG	Exact change	0 = property not applied (default) 1 = exact change as displayed in the reaction	Reaction property

Table 10-1 Meaning of values in the atom block (Continued)

Field	Meaning	Values	Notes
SUBST	Query substitution count	Integer > 0 <i>or</i> 0 = not specified (default) -1 = none	Same maximum as V2000.
UNSAT	Query unsaturation flag	0 = not specified (default) 1 = unsaturated	
RBCNT	Query ring bond count	Integer > 0 <i>or</i> 0 = not specified (default) -1 = none	Same maximum as V2000.
ATTCHPT	Rgroup member attachment points	Attachment points on member: -1 = first and second site 1 = first site only 2 = second site only	When the Rgroup member atom has two attachment points, the atom with the lowest index number attaches to the first attachment point
RGROUPS	nval s is the number of Rgroups that comprise this R# atom. val is the Rgroup number.	Integer > 0	
ATTCHORD	nval s is the number of values that follow on the ATTCHORD line nbr1 is atom neighbor index #1, nbr2 is index #2... val 1 is the attachment order for the nbr1 attachment...	Integer > 0	A list of atom neighbor index and atom neighbor value pairs that identify the attachment order information at the R# atom

Bond block

A bond block specifies all edge information for the connection table. It must precede the Sgroup or 3D blocks. Its format is:

```
M V30 BEGIN BOND
M V30 index type atom1 atom2 [CFG=val] [TOPO=val] [RXCTR=val] [STBOX=val]
...
M V30 END BOND
```

where the values are described in Table 10-2.

Table 10-2 Meaning of values in the bond block

Field	Meaning	Values	Notes
index	Bond index	Integer > 0	The actual value of the index does not matter as long as all are unique. However, extremely large numbers used as indexes can cause the program to fail to allocate memory for the correspondence array.
type	Bond type	Integer: 1 = single 2 = double 3 = triple 4 = aromatic 5 = single or double 6 = single or aromatic 7 = double or aromatic 8 = any	Types 4 through 8 are for queries only.
atom1, atom2	Atom indexes	Integer > 0	Atom1 and Atom2 are bond end points.
CFG	Bond configuration	0 = none (default) 1 = up 2 = either 3 = down	
TOPO	Query property	0 = not specified (default) 1 = ring	

Table 10-2 Meaning of values in the bond block (Continued)

Field	Meaning	Values	Notes
RXCTR	Reacting center status	2 = chain 0 = unmarked (default) -1 = not a reacting center 1 = generic reacting center Additional: 2 = no change 4 = bond made or broken 8 = bond order changes 12 =(4 + 8) bond made or broken and changes 5 = (4 + 1), 9 = (8 + 1), and 13 =(12 + 1) are also possible	
STBOX	Stereo box	0 = ignore the configuration of this double bond (default) 1 = consider the stereo configuration of this double bond	A double bond must be marked to search double bond stereochemistry

Link atom line

There is one link atom line for each link atom in the Ctab. A link atom line has the format:

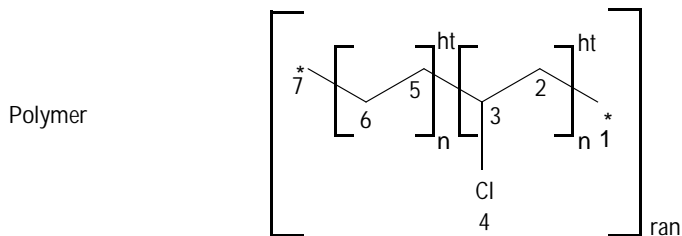
```
M V30 LINKNODE mi nrep maxrep nbonds i natom outatom [i natom outatom...]
```

Table 10-3 Meaning of values in link lines

Field	Meaning	Values	Notes
mi nrep	Minimum number of group repetitions.	1	For future expansion. Not currently used.
maxrep	Maximum number of group repetitions.	Integer > 0	
nbonds	Number of directed bonds defining the group.	nbonds = # of pairs of inatom-outatom tuples	Number of tuples is usually two but may be one for link nodes with an attachment point.
i natom	Atom index of atom in the repeating group.	Integer > 0	
outatom	Atom index of atom bonded to i natom, but outside of repeating group.	Integer > 0	

Sgroup block

The Sgroup block contains general Sgroup information and information on each Sgroup structure as shown in Figure 10-2. For the V2000 version of this Sgroup structure and connection table, see Figure 2-2.

Figure 10-2 Connection table organization of an Sgroup structure

Pol ymer

```
GSMACCS-II 07129516502D 1 0.00374 0.00000 0
Figure 5, J. Chem. Inf. Comput. Sci., Vol 32, No. 3., 1992
0 0 0 0 0 999 V3000
M V30 BEGIN CTAB
M V30 COUNTS 7 6 3 0 0
M V30 BEGIN ATOM
M V30 1 * 2.9463 0.3489 0 0
M V30 2 C 1.6126 1.1189 0 0
M V30 3 C 0.2789 0.3489 0 0 CFG=3
M V30 4 Cl 0.2789 -1.1911 0 0
M V30 5 C -1.0548 1.119 0 0
M V30 6 C -2.3885 0.349 0 0
M V30 7 * -3.9246 1.147 0 0
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 1 1 2
M V30 2 1 2 3
M V30 3 1 3 4
M V30 4 1 5 6
M V30 5 1 5 3
M V30 6 1 7 6
M V30 END BOND
M V30 BEGIN SGROUP
M V30 1 SRU 5 ATOMS=(2 5 6) XBONDS=(2 5 6) BRKXYZ=(9 -0.6103 1.2969 0 -0.6103 -
M V30 0.171 0 0 0 0) BRKXYZ=(9 -3.1565 0.185 0 -3.1565 1.311 0 0 0 0) -
M V30 CONNECT=HT
M V30 2 SRU 6 ATOMS=(3 2 3 4) XBONDS=(2 1 5) BRKXYZ=(9 2.2794 1.2969 0 2.2794 -
M V30 0.1709 0 0 0 0) BRKXYZ=(9 -0.1657 0.171 0 -0.1657 1.2969 0 0 0 0) -
M V30 CONNECT=HT
M V30 3 COP 7 ATOMS=(7 1 2 3 4 5 6 7) BRKXYZ=(9 3.6382 1.6391 0 3.6382 -
M V30 -1.7685 0 0 0 0) BRKXYZ=(9 -4.707 -1.7685 0 -4.707 1.6391 0 0 0 0) -
M V30 SUBTYPE=RAN
M V30 END SGROUP
M V30 END CTAB
M END
```

Header block

Comments line

Counts line

Atom block

Bond block

Rgroup block

3D block

Group Properties

Group 1 Group 2 Group 3

Clab block

Blocks not used in this Ctab

An Sgroup block defines all Sgroups in the molecule, including superatoms. The format is as follows:

```

M V30 BEGIN SGROUP
[M V30 DEFAULT [CLASS=class] -]
M V30 index type extindex -
M V30 [ATOMS=(natoms atom [atom ...])] -
M V30 [XBONDS=(nxbonds xbond [xbond ...])] -
M V30 [CBONDS=(ncbonds cbond [cbond ...])] -
M V30 [PATOMS=(npatoms patom [patom ...])] -
M V30 [SUBTYPE=subtype] [MULT=mult] -
M V30 [CONNECT=connect] [PARENT=parent] [COMPNO=compno] -
M V30 [XBHEAD=(nxbonds xbond [xbond ...])] -
M V30 [XBCORR=(nxbpairs xb1 xb2 [xb1 xb2 ...])] -
M V30 [LABEL=label] -
M V30 [BRKXYZ=(9 bx1 by1 bz1 bx2 by2 bz2 bx3 by3 bz3)]* -
M V30 [ESTATE=estate] [CSTATE=(4 xbond cbvx cbvy cbvz)]* -
M V30 [FIELDNAME=fieldname] [FIELDINFO=fieldinfo] -
M V30 [FIELDDISP=fielddisp] -
M V30 [QUERYTYPE=querytype] [QUERYOP=queryop] -
M V30 [FIELDDATA=fielddata] ... -
M V30 [CLASS=class] -
M V30 [SAP=(3 aix lvid id)]* -
M V30 [BRKTYP=bracketType] -
...
M V30 END SGROUP

```

The DEFAULT field provides a way to specify default values for keyword options. The same keyword options and values as defined in Table 10-4.

Table 10-4 Meaning of values in the Sgroup block

Field	Meaning	Values	Notes
index	Sgroup index	integer > 0	The actual value of the index does not matter as long as all indexes are unique. However, extremely large numbers used as indexes can cause the program to fail to allocate memory for the correspondence array.
type	Sgroup type	String. Only first 3 letters are significant: SUPERatom MULTiple	

Table 10-4 Meaning of values in the Sgroup block (Continued)

Field	Meaning	Values	Notes
		SRU	
		MONomer	
		COPolymer	
		CROsslink	
		MODification	
		GRAft	
		COMponent	
		MIXture	
		FORmulation	
		DATA	
		ANY	
		GENeric	
exti ndex	External index value	Integer => 0: If 0, positive integer assigned	Use 0 to autogenerate a number. This is the V2000 Sgroup label.
ATOMS	natoms is the number of atoms that define the Sgroup.	Integer > 0	
	atom is the atom index.	Integer > 0	
XBONDS	nbonds is the number of crossing bonds.	Integer > 0	
	xbond is the crossing-bond index.	Integer > 0	
CBONDS	ncbonds is the number of containment bonds.	Integer > 0	Only used for Data Sgroups.
	cbond is the containment-bond index.	Integer > 0	

Table 10-4 Meaning of values in the Sgroup block (Continued)

Field	Meaning	Values	Notes
PATOMS	<p>npatom is the number of paradigmatic repeating unit atoms.</p> <p>patom is the atom index of an atom in the paradigmatic repeating unit for a multiple group.</p>	Integer > 0	This field is expected to become obsolete and is retained for compatibility with MACCS-II. The field is only used for multiple groups.
SUBTYPE	subtype is the Sgroup subtype.	<p>String. Only the first 3 letters are significant:</p> <p>ALternate</p> <p>RANdom</p> <p>BLOck</p>	
MULT	mult is the multiple group multiplier.	Integer > 0	
CONNECT	connect is the connectivity.	<p>String values are as follows:</p> <p>EU (default)</p> <p>HH</p> <p>HT</p>	The default, if missing, is EU. The MDL V2000 writer never writes an EU entry.
PARENT	parent is the parent Sgroup index.	Integer > 0	
COMPNO	compno is the component order number.	Integer > 0	
XBHEAD	nxbonds is the number of crossing bonds that cross the "head" bracket.	Integer > 0	
	xbond is the crossing-bond index.	Integer > 0	If XBHEAD is missing, no bonds are paired as the head or tail of the repeating unit.

Table 10-4 Meaning of values in the Sgroup block (Continued)

Field	Meaning	Values	Notes
XBCORR	<p>nxbpairs</p> <p>xb1 - xb2 is the pairs of crossing-bond correspondence, that is, xb1 connects to xb2.</p>	<p>2 x the number of pairs of crossing-bond correspondence, that is, the number of values in list.</p> <p>Integer > 0</p>	
LABEL	label is the display label for this Sgroup.	String	For example, superatom name
BRKXYZ	bx1 - bz3 are the double (X,Y,Z) display coordinates in each bracket.	Angstroms	<p>By specifying 3 triples, the format allows a 3D display.</p> <p>However, only the first two (X, Y) coordinates are currently used. The Z value and last (X, Y) coordinates are currently ignored and should be set to zero.</p>
ESTATE	estate is the expanded display state information for superatoms.	<p>String</p> <p>E = expanded superatom or multiple group</p>	Only superatoms and multiple groups (shortcuts) in an expanded internal state are supported. This field defines whether a superatom or multiple group is displayed as expanded or contracted. This field is expected to become obsolete.
CSTATE	<p>xbond is the crossing bond of the expanded superatom.</p> <p>cbvx - cvbz is the vector to contracted superatom.</p>	<p>Integer > 0</p> <p>Angstroms</p>	<p>Display vector information for the contracted superatom.</p> <p>Only present for expanded superatoms. One CSTATE entry per crossing bond.</p>
FIELDNAME	fieldname is the name of data field for Data Sgroup.	String	

Table 10-4 Meaning of values in the Sgroup block (Continued)

Field	Meaning	Values	Notes
FI ELDI NFO	fi el di nfo is the program-specific field information.	Free-format string	Example: In MACCS-II this is: "<type> <units/format>"
FI ELDDI SP	fi el ddi sp is the Data Sgroup field display information.	Free-format string	This string is interpreted by V3000 as identical to V2000 appendix for Data Sgroup display ('M SDD') except for the index value.
QUERYTYPE	querytype is the type of query or no query if missing.	String ' ' = not a query (default) 'MQ' = MACCS-II query 'IQ' = ISIS query '<p>Q' = <program> query	
QUERYOP	queryop is the query operator.	String. ISIS: query operator MACCS-II: blank or missing	Example: "=" or "LIKE" in ISIS
FI ELDDATA	fi el ddata is the query or field data.	Free-format string	Only one entry per query, but can be more than one for actual data. The order of the entries is important.
CLASS	cl ass is the character string for superatom class.	String	Example: PEPTIDE
SAP	ai dx is the index of attachment point or potential attachment point atom. lvi dx is the index of leaving atom.	Integer > 0 Allowed integers are:	

Table 10-4 Meaning of values in the Sgroup block (Continued)

Field	Meaning	Values	Notes
		0 = none or implied H 'aidx' = atom index number of attachment point atom # = atom index number of atom bonded to 'aidx'	
	i d is the attachment identifier.	String (two chars in V2000)	There must be multiple entries if superatom has more than one attachment point. The order of the entries defines the order of the attachment points. Note that SAP entries may or may not include the actual attachment points, depending on the particular superatom and its representation on the ISIS/Desktop.
BRKTYP	bracketType is the displayed bracket style.	Allowed values for this string are: BRACKET (default) PAREN	This information supports Sgroup enhancements on the ISIS/Desktop.

Correspondence with existing V2000 appendices:

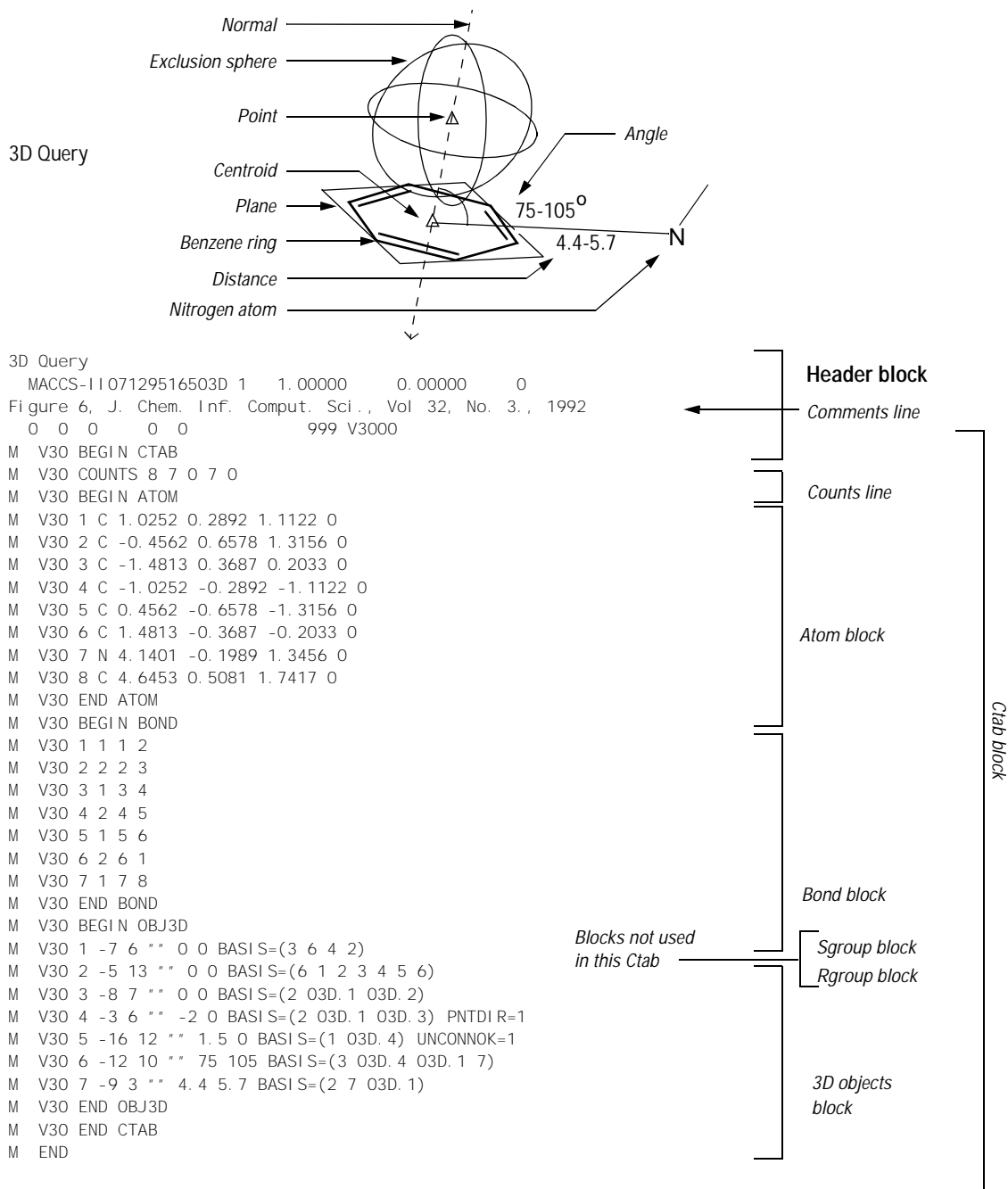
M STY = type
M SST = SUBTYPE
M SLB = ext i ndex
M SCN = CONNECT
M SDS = ESTATE
M SAL = ATOMS
M SBL = XBONDS or CBONDS
M SPA = PATOMS
M SMT = LABEL and MULT
M CRS = XBHEAD, XBCORR
M SDI = BRKXYZ
M SBV = CSTATE
M SDT = FI EL DNAME, FI EL DI NFO, QUERYTYPE, QUERYOP
M SDD = FI EL DDI SP
M SCD = (not requi red)
M SED = FI EL DDATA
M SPL = PARENT

M SNC = COMPNO
M SAP = SAP
M SCL = CLASS
M SBT = BRKTYP

3D block

The 3D block contains 3D information as shown in Figure 10-3. For the V2000 version of this 3D query and its connection table, see Figure 2-3.

Figure 10-3 Connection table organization of a 3D query



A 3D block specifies information for all 3D objects in the connection table. It must follow the atom and bond blocks. As in V2000 molfiles, there can be only one fixed-atom constraint.

The format of the 3D block is as follows:

```
M V30 BEGIN OBJ3D
M V30 index type color name value1 value2 -
M V30 BASIS=(nbvals bval [bval ...]) -
M V30 [ALLOW=(nbvals val [val ...])] [PNTDIR=val] [ANGDIR=val] -
M V30 [UNCONNOK=val] [DATA=strval] -
M V30 [COMMENT=comment]
...
M V30 END OBJ3D
```

Table 10-5 Meaning of values in the 3D block

Field	Meaning	Values	Notes
index	3D object index	Integer > 0	The actual value of the index does not matter as long as all indexes are unique. However, extremely large numbers used as indexes can cause the program to fail to allocate memory for the correspondence array.
type	Object type	Integer < 0 for geometric constraints for data constraints Integer > 0 are field IDs	This format is the same as V2000.
color	Color value	Integer > 0	
name	Object name or, for data query, the field name.	String	
value1	Distance, radius, deviation, or minimum value.	Floating point, value1 = 0 if constraint has no floating values	
value2	Maximum value for range constraints.	Floating point, value2 = 0 if not a range constraint	
BASIS	nbvals is the number of objects in basis.	Integer > 0	

Table 10-5 Meaning of values in the 3D block (Continued)

Field	Meaning	Values	Notes
	bval is the atom number or 3D object index.	Integer <i>or</i> O3D.integer	For objects where order is important, for example, in an angle constructed from three points, the order must be the same as in V2000 molfiles.
ALLOW	nval s is the number of atoms allowed in an exclusion sphere. val is the atom number.	Integer > 0 Integer > 0	
PNTDI R		0 = point has no direction 1 = point has direction	
ANGDI R		0 = dihedral angle has no direction 1 = dihedral angle has direction	MACCS-II uses 'Chiral'.
UNCONNOK		0 = unconnected atoms are not OK 1 = unconnected atoms are OK	
DATA	strval is the data query string	String	
COMMENT	string comment	String. Normally uses the MACCS-II DASP, DISP, and BOX values	Same as V2000 molfile

The Extended Rgroup Query Molfile

A single molecule or Rgroup molecule connection table. The header is contained in the normal header location, that is, in the first three lines of the file. The body of the new molecule is contained in new appendixes, organized as follows:

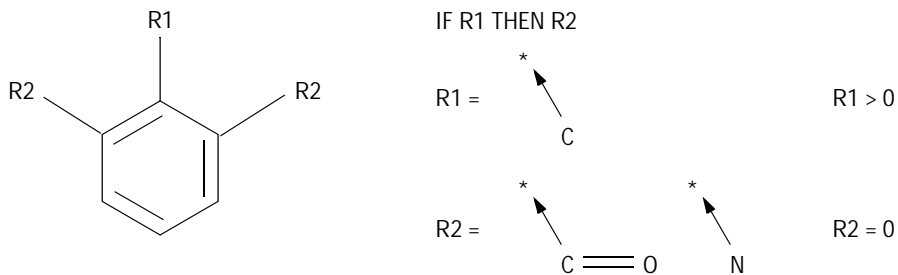
A molecule block consists of a main Ctab, plus optionally one or more Rgroup definitions.

```
ctab-block
[rgroup-block] *
```

Rgroup block

The Rgroup file shown in Figure 10-4 corresponds to the following Rgroup query. For the V2000 version of the Rgroup query and its connection table, see Figure 4-1.

Figure 10-4 Connection table organization of an Rgroup query (Continued on next page)



```

GSMACCS-I | 07139508292D 1 0.00353 0.00000 0
  0 0 0 0 0 999 V3000
M V30 BEGIN CTAB
M V30 COUNTS 9 9 0 0 0
M V30 BEGIN ATOM
M V30 1 C 1.3337 0.77 0 0
M V30 2 C 0 1.54 0 0
M V30 3 C -1.3337 0.77 0 0
M V30 4 C -1.3337 -0.77 0 0
M V30 5 C 0 -1.54 0 0
M V30 6 C 1.3337 -0.77 0 0
M V30 7 R# 0 3.08 0 0 RGROUPS=(1 1)
M V30 8 R# 2.6674 1.54 0 0 RGROUPS=(1 2)
M V30 9 R# -2.6674 1.54 0 0 RGROUPS=(1 2)
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 1 1 2
M V30 2 2 2 3
M V30 3 1 3 4
M V30 4 2 4 5
M V30 5 1 5 6
M V30 6 2 6 1
M V30 7 1 1 8
M V30 8 1 2 7
M V30 9 1 3 9
M V30 END BOND
M V30 END CTAB
M V30 BEGIN RGROUP 1
M V30 RLOGIC 2 0 ""
M V30 BEGIN CTAB
M V30 COUNTS 1 0 0 0 0
M V30 BEGIN ATOM
M V30 1 C 12.21 14.3903 0 0 ATTCHPT=1
M V30 END ATOM
M V30 END CTAB
M V30 END RGROUP
M V30 BEGIN RGROUP 2
M V30 RLOGIC 0 0 0
M V30 BEGIN CTAB
M V30 COUNTS 2 1 0 0 0
M V30 BEGIN ATOM
M V30 1 C -1.4969 0.0508 0 0 ATTCHPT=1
M V30 2 O 0.0431 0.0508 0 0
M V30 END ATOM
M V30 BEGIN BOND
M V30 1 2 1 2
M V30 END BOND
M V30 END CTAB
M V30 BEGIN CTAB
M V30 COUNTS 1 0 0 0 0
M V30 BEGIN ATOM
M V30 1 N 12.21 14.3903 0 0 ATTCHPT=1
M V30 END ATOM
M V30 END CTAB
M V30 END RGROUP
M FND

```

Diagram illustrating the structure of the Molfile format, showing the organization of data into blocks and sub-blocks:

- Header block**: Contains the initial header information (GSMACCS-I | 07139508292D 1 0.00353 0.00000 0).
- Counts line**: A line indicating the number of atoms and bonds (0 0 0 0 0).
- Atom block of root**: A block containing atom coordinates and properties for the root molecule (M V30 1 C 1.3337 0.77 0 0 to M V30 9 R# -2.6674 1.54 0 0 RGROUPS=(1 2)).
- Bond block of root**: A block containing bond information for the root molecule (M V30 1 1 1 2 to M V30 9 1 3 9).
- Block for Rgroup R1**: A block containing data for Rgroup 1, including a counts line (M V30 COUNTS 1 0 0 0 0), an atom block (M V30 1 C 12.21 14.3903 0 0 ATTCHPT=1), and an end of Rgroup marker (M V30 END RGROUP).
- Block for Rgroup R2**: A block containing data for Rgroup 2, including a counts line (M V30 COUNTS 2 1 0 0 0), an atom block (M V30 1 C -1.4969 0.0508 0 0 ATTCHPT=1 to M V30 2 O 0.0431 0.0508 0 0), a bond block (M V30 1 2 1 2), and an end of Rgroup marker (M V30 END RGROUP).

An Rgroup block defines one Rgroup. Each Ctab block specifies one member.

```
M V30 BEGIN RGROUP rgroup-number
[rgroup-logic-line]
ctab-block
[ctab-block]*
M V30 END RGROUP
```

Table 10-6 Meaning of values in the Rgroup block

Field	Meaning	Values	Notes
rgroup-number	Index of this rgroup	Integer > 0	

Rgroup logic lines

There is zero or one Rgroup logic line for each Rgroup in the molecule. If present, the Rgroup logic line specifies if-then logic between Rgroups, the convention about unfilled valence sites, and the Rgroup occurrence information. Its format is:

```
M V30 RLOGIC thenR RestH Occur
```

Table 10-7 Meaning of values in Rgroup logic line

Field	Meaning	Values	Notes
thenR	Number of a "then" Rgroup	0 = none (default)	
RestH	Attachment(s) at Rgroup position	0 = off, that is, any molecule fragment at any unsatisfied Rgroup location (default) 1 = only hydrogen or a member of Rgroup is allowed	
Occur	String specifying number (range) of Rgroup occurrence sites that need to be satisfied.	String '> 0' = default	Similar to MACCS-II and ISIS: [N[,N[,...]]]