# Supporting Information 

# Adapting CHMTRN (CHeMistry TRaNslator) 

## for a New Use

# Philip N. Judson*, ${ }^{a}$ Wolf-Dietrich Ihlenfeldt, ${ }^{b}$ Hitesh Patel, ${ }^{c}$ Victorien Delannée, ${ }^{c}$ Nadya Tarasova, ${ }^{d}$ Marc C. Nicklaus ${ }^{\text {c }}$ 

a) Heather Lea, Bland Hill, Norwood, Harrogate HG3 1TE, England.
b) Xemistry GmbH, Schwalbenweg 5, D-61479 Glashütten, Germany.
c) Computer-Aided Drug Design Group, Chemical Biology Laboratory, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD 21702, USA.
d) Synthetic Biologics and Drug Discovery Group, Laboratory of Cancer Immunometabolism, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Frederick, MD 21702, USA.

## INTRODUCTION

The terms used in our current implementation of CHMTRN are listed in Tables S1 to S19. Additional details for some of them are presented below. The original CHMTRN language
contains many further terms which we have not so far needed in the current project and for which we have not implemented supporting functions - e.g. to call external subroutines that calculate a variety of physico-chemical properties such as HOMO*COEFFICENT, ELECTROPHILIC*LOCALIZATION*ENERGY.

The original CHMTRN language used only upper case characters, that being the normal convention for computer code at the time the language was developed. Our CHMTRN interpreter is not case sensitive but, for consistency and backward compatibility, we have continued to write CHMTRN code in upper case.

## NOTES ON TOPICS REQUIRING MORE DETAIL THAN IS GIVEN IN THE TABLES

## Hydrogen atoms

Hydrogen atoms are not normally treated as atoms in their own right in CHMTRN but as a property of the atom to which they are attached. Thus methane is not perceived as four hydrogen atoms each attached by a single bond to the same carbon atom, but as a carbon atom with a hydrogen count of 4. The transform writer can, however, include explicit hydrogen atoms by using the term HYDROGEN*ATOM(S).

The distinction between explicit and implicit hydrogen atoms has implications for how CHMTRN statements should be phrased. Consider the substructure in Figure S1(a), in which the atoms are numbered as shown, the two R groups are non-hydrogen and there are no additional non-hydrogen substituents on atom 2. Provided that the usual carbon valency is satisfied by attached hydrogen atoms, it might be more clearly represented as in Figure S1(b), in which "H2" represents two implicit hydrogen atoms on atom 2. In Figure S 1 (c) one explicit hydrogen atom has been attached to atom 2 and so, to make up unsatisfied valency, there is one implicit
hydrogen. In CHMTRN, the value returned for "THE NUMBER OF HYDROGEN ATOMS ON ATOM*2" is two in case (b) and case (c). However, the value returned for "IF THERE IS A HYDROGEN*ATOM ALPHA TO ATOM*2 ..." is false in case (b) and true in case (c). By default, explicit and implicit hydrogen atoms are all ignored when a statement does not refer directly to hydrogen: e.g. there are two non-hydrogen atoms alpha to atom $2\left(\mathrm{R}^{1}\right.$ and $\left.\mathrm{R}^{2}\right)$ and so the value returned for "THE NUMBER OF ATOMS ALPHA TO ATOM*2" is two in case (c) as well as in case (b). The word EXPLICIT*HYDROGEN can be used in order to include explicit hydrogen atoms: "THE NUMBER OF ATOMS ALPHA TO ATOM*2 INCLUDING EXPLICIT*HYDROGEN" would return the value two for case (b) and the value three for case (c).


Figure S1. Representation of implicit and explicit hydrogen atoms

## PATTERNS

## 1D*PATTERNs

For convenience, the following description of PATRAN 1D patterns, which is included in the paper for which this supporting information is provided, is repeated here.

Atoms are represented by their elemental symbols in capital letters (e.g. chlorine is "CL"), with the letter X representing any element. Branching is indicated using parentheses. Bonds are represented by "-", "=", "\#", "\%", "\&", and "+", for single, double, triple, aromatic, any bond
type, and no bond (the separator between fragments in a disconnected graph), respectively. Alternative atoms and alternative bonds are separated by commas. Rings are indicated by an @ sign followed by the sequence number of an earlier atom in the pattern string to which a bond connects to complete the ring (atoms and bonds are automatically numbered sequentially from the left).

Atom and bond properties are inserted between square brackets. Alternative values for a property are separated by commas; alternative properties are separated by semicolons (e.g. [HS=0;ARYL=YES]). The atom properties currently supported are listed in Table S1.

For example, the following pattern represents the substructures in Figure S 2 (where no substituent is shown in Figure S2, any substituent is allowed).
$\mathrm{N}[\mathrm{HS}=1] \% \mathrm{C}(-\mathrm{C}[\mathrm{FGS}=\mathrm{ESTER}, \mathrm{AMIDE}]) \% \mathrm{C}, \mathrm{N} \% \mathrm{C} \% \mathrm{C} @ 1$


Figure S2. A pair of related generic structures

Long patterns can be written on multiple lines by breaking after a bond symbol, comma, or semicolon.

## 2D*PATTERNs

Many LHASA transforms include code under the header, 2D*PATTERN. When a CHMTRN file is used with LHASA, the purpose of the 2D*PATTERN is to provide keying information about the precursors generated by transforms, in order to support the prediction of reactions (as
distinct from retro-reactions). Figure S3 is an example of a 2D pattern from transform 2875 for a Copper[I]-catalyzed azide-alkyne cycloaddition, written recently by Martin Ott.

| H C | @1 $0, \mathrm{~N}$ | H C | MINUS | $0, \mathrm{~N}$ |
| :---: | :---: | :---: | :---: | :---: |
| 1 \| | \% " | \| | | I | " |
| $\mathrm{C} \% \mathrm{C} \% \mathrm{~N} \% \mathrm{~N} \% \mathrm{~N}--\mathrm{S}-\mathrm{C}=\mathrm{C} \# \mathrm{C} \quad \mathrm{N}=\mathrm{N}=\mathrm{N} \quad \mathrm{S}-\mathrm{C}$ |  |  |  |  |
|  | " |  | , | " |
|  | O, N |  | PLUS | O, N |

Figure S3. An example of a 2D pattern

2D patterns are difficult to write and maintain, because of the requirement for multiple lines of text to be correctly aligned, and they carry less information about atom and bond properties than 1D patterns. They are included in transforms written for our current project, for backward compatibility with LHASA, but not used by our interpreter, which uses NEW*1D*PATTERNs.

## CHMTRN STATEMENTS

## Names of Transforms and Other Textual Values

The first line of a CHMTRN transform contains its unique, integer identifier. The second line contains its name, entered as free text, which is intended to tell a chemist what reaction the transform describes. Examples are:

NAME Aldol Reaction
and
NAME Ester or Amide or Thiolester Formation.
Within the body of a transform, the name to be displayed to an end-user can be changed using TITLE. Actual conditions to be displayed to the user can be specified using ACTUAL*CONDITIONS (see also the section entitled "Conditions") and warning messages can
be communicated to the end user using WARNING. The statements can be conditional. For example:

TITLE 3178: Lactone Formation
IF BIT 1 IS ON THEN ACTUAL*CONDITIONS 877: Pd/Ferrocene complex, Cs2CO3.
For reasons of history, in all three cases, although the free text is included in the CHMTRN statement, the integer preceding it must be unique to that text. In a future implementation it might be better for the text to be maintained in separate lookup files and for only the integer identifiers to be included in the transform statements.

## Subroutines

Subroutines can be called from within transforms using
CALL aaa
where aaa is the name of the subroutine and the header of the subroutine in the CHMTRN file is

SUBROUTINE aaa
Up to five atoms and bonds can be passed as parameters to the subroutine. For example,
CALL MYSUB AT ATOM*2 AND ATOM*4 AND BOND*3
Within the subroutine, the atoms and/or bonds are identified as SPECIFIED*ATOM n or SPECIFIED*BOND $n$, where n is determined by the position of the atom or bond in the parameter list. So in the above example, SPECIFIED*ATOM 1 is ATOM*2, SPECIFIED*ATOM 2 is ATOM*4, and SPECIFIED*BOND 1 is BOND*3.

A subroutine terminates when a RETURN statement is encountered, and it returns SUCCESS (true) or FAIL (false) to the calling transform.

## Mechanism Statement

Transforms include a block of code setting out the changes needed to convert the product (target) into the precursors, described as the transform mechanism and preceded and followed by a set of four dots (....). For example,
$\qquad$
BREAK BOND*2
ATTACH A BROMIDE TO ATOM*3
....
In a transform that is described retrosynthetically, statements in the body of the transform that precede the mechanism block relate to the product (target); if there are statements after the mechanism block they relate to the precursors (reactants).

## Appendages

Branches in a structure can be compared using statements such as
KILL IF THE APPENDAGES FROM ATOM*2 TOWARDS ATOM*3 \& AND FROM ATOM*5 TOWARDS ATOM*7 ARE NOT IDENTICAL

The word separators TOWARDS and AND must always be used with the term APPENDAGES.

Alternatively, where appropriate, a statement such as

## RAISE*RATING SLIGHTLY IF ATOM*2 IS EQUIVALENT TO ATOM*5

tests whether the two atoms belong to the same symmetry class.

## Conditions

There are two kinds of statements in CHMTRN about reaction conditions. LHASA uses a CONDITIONS statement to check for potential reactions at sites other than the intended one. It highlights the problem sites automatically to the user and may propose appropriate protection. Incompatibilities between functional groups listed in Table S7 and reaction conditions are stored in a lookup matrix. For this reason, CONDITIONS can only be selected from a predefined set. If the actual conditions are not included in the predefined set, the transform writer selects the nearest one(s), in order to trigger automatic recognition of incompatibilities in a query structure. Multiple conditions can be listed, separated by AND. Sequences of conditions can be included, using FOLLOWED BY, to take account of changes of conditions during the course of a reaction or during workup. For example,

## CONDITIONS DCCD AND R3N FOLLOWED BY pH2:4

Actual conditions are recorded in a transform using ACTUAL*CONDITIONS so that they can be reported to an end-user. They are not used by the program. Each actual conditions phrase, which is free text, must be preceded by a unique identifying number and a colon. For example,

IF BIT 1 IS ON THEN ACTUAL*CONDITIONS 878: Pd/BINAP, Cs2CO3

## Rating

The knowledge base for the LHASA program originally included a numerical rating in the range 0 to 100, in increments of 5, for each transform, which was the transform writer's assessment of the utility of the retro-reaction it described. The synonymous terms, RATING and INITIAL*RATING are supported.

For example,
RATING 65
or

## INITIAL*RATING 65

The rating value is not the likely percentage yield of the reaction: it takes into account factors such as how well-documented a reaction is and how consistent yields are (one reaction might deliver only $40 \%$ yields but do so consistently, while another might deliver yields ranging from a few percent to almost one hundred percent with no clear reason for the variability). Within the body of a transform the initial rating can be increased or decreased by multiples of five using ADD (or INCREMENT) or SUBTRACT (or DECREMENT). For example,

## ADD 25 IF THERE IS A DONATING GROUP ON ATOM*3

It was difficult to set out guidelines for consistent initial rating of transforms by different writers. So an alternative approach was introduced in the 1990s, in which the different factors were assessed independently. In addition, instead of using numerical values, the individual factors were assigned categorical values. An initial rating was calculated automatically by the program, allowing the weighting of the different factors to be adjusted. The factors contributing to rating were as follows. In each case guidelines can be drawn up on how to decide upon the appropriate category (e.g. what average yield should be considered GOOD and what FAIR). As well as potentially improving consistency, separating the factors contributing to the rating makes it possible for an end-user to customise priorities if, say, orientational selectivity is a primary consideration.

## CONDITION*FLEXIBILITY

An assessment of how many different types of reagents can be used for the reaction.

## HETEROSELECTIVITY

An assessment of how selective the reaction is towards the functional groups in the reaction centre compared with other functional groups.

## HOMOSELECTIVITY

An assessment of how selective the reaction is towards the functional group(s) in the intended site of reaction compared with instances of the same functional group(s) in other environments in the same structure.

## ORIENTATIONAL*SELECTIVITY

An assessment of the selectivity of the reaction towards one location in the site of reaction compared with another (e.g. leading to Markovnikov or anti-Markovnikov addition).

## RELIABILITY

An assessment of how predictable a reaction is, both in terms of yield and, where relevant, stereochemical outcome.

## REPUTATION

How widely reported, and reproduced, the reaction is (or, to be more exact, was at the time that the transform was written or last updated).

THERMODYNAMICS

How easily the reaction proceeds spontaneously.
TYPICAL*YIELD

A category based on the percentage yield normally to be expected.

Categories that can be assigned to rating factors are: BAD, POOR, FAIR, GOOD, EXCELLENT, and NOT*APPLICABLE. In addition, for TYPICAL*YIELD only, the category VERY*GOOD can be used.

## Stereochemistry

The terms SYN and ANTI, with the meanings understood by a chemist, are used to describe double bond geometry and also diastereoisomerism. They can be used in relation to atoms and/or bonds.

In the case of double bonds, the usage is straightforward. For example, IF ATOM* 1 IS ANTI TO ATOM*4 ...

IF ATOM*1 IS ANTI TO BOND*3 ...
Defining the relationship between atoms or bonds on a ring or saturated chain is more complicated and uses the approach described by Masamune et al, ${ }^{1}$ using an additional atom attached to the same atom as each of those of interest to define the layout of the path between them. For example,

## IF ATOM*1 IS SYN TO ATOM*5 ABOUT THE PATH FROM ATOM*2 TO ATOM*6 ...

might be used to test the stereochemical relationship between atom A1 and atom A5 in Figure
S4. The chain between atom A2 and atom A6 is laid out zigzag fashion as shown in the figure. If A1 and A5 are both above, or both below, the plane of the paper, then they are recognised as SYN to each other. Otherwise they are ANTI.


Figure S4. Layout of a structure to define a stereochemical relationship

## Properties Relating to Rings

The term BRIDGE(D) supports conditional statements such as IF THE CURRENT*RING IS BRIDGED THEN ... IF THE CURRENT*RING IS BRIDGED AT ATOM*2 THEN ...

The original CHMTRN includes other terms relating to the properties of atoms and bonds in rings which are not currently used in SAVI and have therefore not yet been re-implemented: 3*OR*MORE*RINGS, CIS*BRIDGE(D), CORE, LOCKED*BOAT, LOCKED*CHAIR, OFF*THE*BRIDGE, ON*THE*BRIDGE, SPIRO, TRANS*BRIDGE(D).

## Strained Bonds

A bond is considered strained if it is:
in a three- or four-membered ring;
or in a five-membered ring that is part of a [2.1.1], [2.2.1], or [3.2.1] system;
or a trans double bond in a ring of size 9 or less;
or a double bond exocyclic to a three- or four-membered ring.
The atom(s) attached to the bond and in the ring are also classified as strained and so statements like either of the following can be used:

IF BOND*3 IS STRAINED ...
IF ATOM*5 IS STRAINED ...
The term BREDT*STRAINED refers to bridgehead double bonds that are in a ring of size less than 8 and to the associated bridgehead atom: IF ATOM* 1 IS BREDT*STRAINED ...

## REFERENCE

1 Masamune, S.; Ali, Sk. A.; Snitman, D. L.; Garvey, D. S. Highly Stereoselective Aldol Condensation Using an Enantioselective Chiral Enolate. Angew. Chem. Int. Ed., 1980, 19, 557.

TABLES

Table S1. Atom and Bond Properties Supported in 1D Patterns

| Code | Property | Allowed values |
| :---: | :---: | :---: |
| ARYL | Whether an atom is aromatic | YES, NO, EITHER |
| CHARGE | Whether an atom is charged or a radical | ANION, CATION, NEUTRALa, RADICAL, YES, NO ${ }^{\text {a }}$, EITHER |
| EPS ${ }^{\text {b }}$ | The number of electron pairs on an atom | 0, 1, 2, 3, 4 |
| FGNOT | Functional groups that are not allowed | Any of the groups listed in Table S7 |
| FGS | Functional groups one of which must be present | Any of the groups listed in Table S7 |
| HETS ${ }^{\text {b }}$ | The number of hetero atoms alpha to an atom | 0, 1, 2, 3, 4 |
| $\mathrm{HS}^{\text {b }}$ | The number of hydrogens (explicit and implicit) on an atom | 0, 1, 2, 3, 4 |
| RING ${ }^{\text {b }}$ | The size of a ring containing an atom | 0 or an integer of 3 or more, YES, $\mathrm{NO}^{\mathrm{c}}$, EITHER ${ }^{\text {d }}$ |
| FUSION ${ }^{\text {e }}$ | Whether a bond is a fusion bond between rings | DIARYL, ALKYLARYL, DIALKYL, TRIALKYL, YES, NO, EITHER |

a) NEUTRAL and NO are synonymous.
b) In addition to $=$, the operators $>$ and $<$ are supported for numerical values.
c) $\mathrm{RING}=0$ means the same as $\mathrm{RING}=\mathrm{NO}$ (i.e. the atom is acyclic).
d) EITHER means the atom is allowed to be cyclic or acyclic.
e) In addition to =, the operator \# is supported, representing "not equals".

Table S2. Transform Properties

| Reserved word ${ }^{\text {a }}$ | Notes |
| :---: | :---: |
| 1D*PATTERN | See the section entitled 'Patterns' in the paper to which this document is a supplement |
| 2D*PATTERN | See the section entitled ' 2 D *PATTERN' |
| ACTUAL*CONDITIONS n:aaa | See the section entitled 'Conditions' |
| CONDITION*FLEXIBILITY aaa | See the section entitled 'Rating' |
| CONDITIONS aaa | The most similar reaction conditions in a lookup matrix to the actual ones used for the reaction that a transform describes. The matrix contains information on the sensitivity of functional groups to reaction conditions and supports automatic attachment of protecting groups in LHASA. See the section entitled 'Conditions'. |
| COPYRIGHT aaa | Copyright claim (e.g. "COPYRIGHT Smith and Co., 2019") |
| COST $\mathrm{aaa}^{\text {b }}$ | See the section entitled 'Rating' in the paper to which this document is a supplement |
| END*PATTERN(S) | See the section entitled 'Patterns' in the paper to which this document is a supplement |
| END*REFERENCES | Marks the end of the section in a transform containing bibliographic references and similar information (see REFERENCES). |
| HETEROSELECTIVITY aaa | See the section entitled 'Rating' |
| HOMOSELECTIVITY aaa | See the section entitled 'Rating' |
| INITIAL*RATING n | See the section entitled 'Rating' |
| NAME aaa | The name of the transform (e.g. "NAME Aldol Reaction"). Transform names do not need to be unique (the unique identifier is the transform number). |
| NEW* ${ }^{\text {d }}$ * PATTERN ${ }^{\text {b }}$ | See the section entitled 'Patterns' in the paper to which this document is a supplement |
| ORIENTATIONAL*SELECTIVITY aaa | See the section entitled 'Rating' |

OPT*REFERENCES aaa ${ }^{\text {b }}$
OPTIONAL*REFERENCES

RATING n
REFERENCES

RELIABILITY aaa
REPUTATION aaa
SAFETY aaa
THERMODYNAMICS aaa
TITLE n:aaa

TRANSFORM n
TYPICAL*YIELD aaa

Identifies optional reference(s), separated by "AND", that are to be displayed to the user (e.g. "IF THERE IS A KETONE GROUP ON ATOM* 1 THEN OPT*REFERENCES 3 AND 5".

Marks the start of an optional subsection at the end of the REFERENCES section, containing bibliographic references that are only displayed to the user if they are selected by a statement in the main body of the transform. Each reference in the optional references subsection is preceded by "nnn: ", where $n n n$ is an integer that is unique in the optional references list of the transform.

See the section entitled 'Rating'
Marks the beginning of the section in a transform containing bibliographic references and similar information such as a brief textual or diagramatic description of the reaction/retroreaction.

See the section entitled 'Rating'
See the section entitled 'Rating'
See the section entitled 'Rating'
See the section entitled 'Rating'
An alternative transform name, specified in the main body of the transform, to be displayed to the user in place of the one set by NAME.

The unique integer identifier for the transform.
See the section entitled 'Rating'
a) aaa represents a string value; n represents an integer.
b) Terms that are not part of the original CHTMRN language. They have been added in the current project.

Table S3. Rating Variables

## MODERATELY

SEVERELY
SLIGHTLY
STRONGLY
BAD
POOR
FAIR
GOOD
VERY*GOOD
EXCELLENT
NOT*APPLICABLE

Table S4. Atom Types

ALUMINIUM ${ }^{\text {a }}$<br>BORON<br>BROMINE(S)<br>CARBON(S)<br>CHLORINE(S)<br>COPPER<br>FLUORINE(S)<br>HYDROGEN*ATOM(S) ${ }^{\text {b }}$<br>IODINE(S)<br>NITROGEN(S)<br>OXYGEN(S)<br>PHOSPHORUS<br>SELENIUM<br>SILICON(S)<br>SULFUR<br>SULPHUR<br>STANNUM<br>TIN<br>TITANIUM

a) The use of "aluminum" as an alternative IUPAC name was not formally adopted until 1993, a long time after CHMTRN was created. It could be added, in the same way as SULFUR and SULPHUR are supported, but we have not so far needed to do so.
b) See the section entitled 'Hydrogen atoms'

Table S5. Other Atom Properties

| Reserved word ${ }^{\text {a }}$ | Notes ${ }^{\text {b }}$ |
| :---: | :---: |
| ALKALI*METAL ${ }^{\text {c }}$ | The atom is $\mathrm{Li}, \mathrm{Na}, \mathrm{K}, \mathrm{Rb}$, or Cs |
| ALKYL(S) | The number of ALKYL*ATOMs (see Table S8) attached to the atom. The return value is an integer. |
| ALLYLIC | The atom is a saturated carbon atom alpha to a carboncarbon double bond |
| ANTI | (1) Tests the relationship between a pair of atoms and/or bonds about a double bond. (2) Tests the diastereomeric relationship between a pair of atoms and/or bonds. See the section entitled 'Stereochemistry'. |
| AROMATIC | The atom is aromatic. |
| ARYL | Synonym for AROMATIC. |
| BENZYLIC | A saturated carbon atom that is alpha to an aromatic carbon atom. |
| BREDT*STRAINED | The atom is on a double bond which is at a bridgehead and in a ring of size less than eight. See the section entitled 'Strained Bonds'. |
| BRIDGEHEAD(S) | Bridgehead atoms in bridged cyclic systems. |
| BRIDGED*TO*CURRENT*RING | The atom is in a chain that bridges the CURRENT*RING. |
| CIS | Tests whether two atoms on the same ring are cis to each other. The two atoms and the ring must all be specified (e.g. "IF ATOM*1 IS CIS TO ATOM*2 ABOUT THE RING CONTAINING BOND*4 THEN...") |
| CIS*OLEFIN | A carbon atom joined to another carbon atom by a double bond where both carbon atoms carry one (and only one) hydrogen atom and the substituents are in a cis relationship to each other. |
| COMMON | Tests whether atoms and/or bonds are in the same ring (e.g. "IF ATOM*1 AND BOND*3 ARE IN A COMMON RING ..."). |


| CONJUGATED | An atom that is alpha to a multiply-bonded atom which <br> is connected to, but not part of, a structural group that <br> keys the transform. |
| :--- | :--- |
| DOUBLY | There is a double bond on the atom (normally used <br> with the buzz word BONDED - i.e. "DOUBLY <br> BONDED"). <br> ELECTRON*DENSITY <br> The electron density calculated for the atom (the return <br> value is decimal). <br> ENOLIZABLE |
| The atom is saturated, not a bridgehead atom, has at |  |
| least one hydrogen atom on it, and is activated by a |  |
| withdrawing bond. |  |
| EXPLICIT*HYDROGEN | Tests whether two atoms belong to the same symmetry <br> class (e.g "RAISE*RATING MODERATELY IF <br> ATOM*3 IS EQUIVALENT TO ATOM*4"). See the |
| section entitled 'Appendages'. |  |
| FUSED | See the section entitled 'Hydrogen atoms'. |
| FUSED*TO*CURRENT*RING | The atom is at a ring fusion. <br> The atom is in a ring that is fused to the <br> CURRENT*RING. <br> FUSION |
| Synonym for FUSED. |  |


|  | ATOM*3 THEN ..."). |
| :---: | :---: |
| MORE*HINDERED THAN | Synonym for MORE*HINDERED. |
| MORE*HINDERED*THAN | Synonym for MORE*HINDERED. |
| MULTIPLY | A bond on the atom is double, triple, or aromatic. |
| MULTIPLY*BONDED | Synonym for MULTIPLY. |
| NOT*ANTI | Negates the value returned for ANTI (q.v.) |
| NOT*SYN | Negates the value returned for SYN (q.v.) |
| OFF*CURRENT*RING | The atom is not in the ring that has been designated as the CURRENT*RING. |
| OFF*THE*RING | Synonym for OFF*CURRENT*RING. |
| OFFPATH | The atom is not included in the keying pattern for the transform. |
| OFF*PATH | Synonym for OFFPATH. |
| OFFRING | Synonym for OFF*CURRENT*RING. |
| OFF*RING | Synonym for OFF*CURRENT*RING. |
| ON*THE*RING | Synonym for ON*CURRENT*RING. |
| ONPATH | The atom is included in the keying pattern for the transform. |
| ON*PATH | Synonym for ONPATH. |
| ON*CURRENT*RING | The atom is in the ring that has been designated as the CURRENT*RING. |
| ON*RING | Synonym for ON*CURRENT*RING. |
| ONRING | Synonym for ON*CURRENT*RING. |
| POSITIVE*CHARGE | The atom carries a positive charge. |
| PRIMARY*CENTER(S) | There is one (and only one) carbon atom attached to the atom. |
| PRIMARY*CENTRE(S) | Synonym for PRIMARY*CENTER(S). |
| QUATERNARY*CENTER(S) | There are four carbon atoms attached to the atom. |
| QUATERNARY*CENTRE(S) | Synonym for QUATERNARY*CENTER(S). |


| SECONDARY*CENTER(S) | There are two (and only two) carbon atoms attached to <br> the atom. <br> SECONDARY*CENTRE(S) |
| :--- | :--- |
| Synonym for SECONDARY*CENTER(S). |  |
| STRAINED | A bond on the atom is STRAINED. See the section <br> entitled 'Strained Bonds'. <br> SYN <br> (1) Tests the relationship between a pair of atoms <br> and/or bonds about a double bond. (2) Tests the <br> diastereomeric relationship between a pair of atoms <br> and/or bonds. See the section entitled <br> 'Stereochemistry'. |
| TERTIARY*CENTER(S) | There are three (and only three) carbon atoms attached <br> to the atom. |
| TERTIARY*CENTRE(S) | Synonym for TERTIARY*CENTER(S). <br> TRANS <br> Tests whether two atoms on the same ring are trans to <br> each other. The two atoms and the ring must all be <br> specified (e.g. "IF ATOM*1 IS TRANS TO ATOM*2 <br> ABOUT THE RING CONTAINING BOND*4 <br> THEN...") <br> TRANS*OLEFIN <br> A carbon atom joined to another carbon atom by a <br> double bond where both carbon atoms carry one (and <br> only one) hydrogen atom and the substituents are in a |
| trans relationship to each other. |  |

a) Where an optional letter S is shown, either word can be used synonymously. E.g. "ALKYL" and "ALKYLS" are both supported (the combined term, "ALKYL(S)", is used for convenience in the table and is not part of the language).
b) The values of properties are Boolean (true or false) unless otherwise indicated.
c) Term that is not part of the original CHTMRN language. It has been added in the current project.

Table S6. Bond Properties

| Reserved word | Notes |
| :---: | :---: |
| ALLYLIC | The bond is on an allylic carbon atom and is not the bond linking that atom to the doubly bonded one. |
| ANTI | (1) Tests the relationship between a pair of atoms and/or bonds about a double bond. (2) Tests the diastereomeric relationship between a pair of atoms and/or bonds. See the section entitled 'Stereochemistry'. |
| AROMATIC | The bond is aromatic. |
| ARYL | Synonym for AROMATIC. |
| BREDT*STRAINED | The bond is a double bond at a bridgehead and in a ring of size less than eight. See the section entitled 'Strained Bonds'. |
| BRIDGEHEAD(S) | The bonds attached to bridgehead atoms in bridged cyclic systems. |
| BRIDGED*TO*CURRENT*RING | The bond is in a chain that bridges the CURRENT*RING. |
| CIS*OLEFIN | A double bond between two carbon atoms where both carbon atoms carry one (and only one) hydrogen atom and the substituents are in a cis relationship to each other. |
| COMMON | Tests whether atoms and/or bonds are in the same ring (e.g. "IF BOND*1 AND ATOM*3 ARE IN A COMMON RING ..."). |
| CONJUGATED | A bond on an atom that is alpha to a multiply-bonded atom in a functional group and is not the one connecting the atom to the functional group. |
| DIARYL*FUSION | A fusion bond between two aromatic rings. |
| DONATING | The bond is activated/deactivated by a DONATING group. |
| DOUBLE*BOND(S) | A double bond. |
| FUSED*TO*CURRENT*RING | The bond is in a ring that is fused to the CURRENT*RING. |


| FUSION | A fusion bond between two rings. |
| :---: | :---: |
| MULTIPLE | A double, triple, or aromatic bond. |
| MULTIPLE*BOND(S) | Synonym for MULTIPLE. |
| NOT*ANTI | Negates the value returned for ANTI (q.v.). |
| NOT*SYN | Negates the value returned for SYN (q.v.). |
| OFF*CURRENT*RING | The bond is not in the ring that has been designated as the CURRENT*RING. |
| OFFPATH | The bond is not included in the keying pattern for the transform. |
| OFFRING | Synonym for OFF*CURRENT*RING. |
| OFF*SPECIFIED*PATH | Used only in subroutines. The bond is not included in the set passed to the subroutine as part of the specified path. |
| ON*CURRENT*RING | The bond is in the ring that has been designated as the CURRENT*RING. |
| ONPATH | The bond is included in the keying pattern for the transform. |
| ON*SPECIFIED*PATH | Used only in subroutines. The bond is included in the set passed to the subroutine as part of the specified path. |
| ONRING | Synonym for ON*CURRENT*RING. |
| SINGLE*BOND(S) | A single bond. |
| SMALL*FUSION | A fusion bond between a pair of rings of sizes $3 / 3$, $3 / 4,3 / 5,3 / 6,4 / 4.4 / 5,4 / 6$, or $5 / 5$. |
| SMALL*FUSIONS*BOND(S) | Synonym for SMALL*FUSION. |
| STRAINED | See the section entitled 'Strained Bonds'. |
| SYN | (1) Tests the relationship between a pair of atoms and/or bonds about a double bond. (2) Tests the diastereomeric relationship between a pair of atoms and/or bonds. See the section entitled 'Stereochemistry'. |
| TRANS*OLEFIN | A double bond between two carbon atoms where both carbon atoms carry one (and only one) hydrogen |

## TRIPLE*BOND(S)

WITHDRAWING
atom and the substituents are in a trans relationship to each other.

A triple bond.
The bond is activated/deactivated by a WITHDRAWING group. Note that the withdrawing bonds of a non-expandable withdrawing group (e.g. NITRO) are on the same atom as the withdrawing group; the withdrawing bonds of an expandable withdrawing group (e.g. KETONE) are on the atom(s) alpha to it.

Table S7. Functional Groups

| Reserved word $^{\mathrm{a}}$ | Substructure |
| :--- | :--- |
| ACETAL $^{\mathrm{b}}$ | $\mathrm{C}-\mathrm{O}-\mathrm{C}^{*}-\mathrm{O}-\mathrm{C}$ |
| ACETYLENE | $\mathrm{C}^{*=\mathrm{C}}$ |
| ACID | $\mathrm{C}^{*}(=\mathrm{O})-\mathrm{OH}$ |
| ACID*HALIDE | $\mathrm{C}^{*}(=\mathrm{O}) \mathrm{X}[\mathrm{X}=\mathrm{Cl}, \mathrm{Br}$, or I] |
| ALCOHOL | $\mathrm{C}^{*}-\mathrm{OH}$ |
| ALDEHYDE | $\mathrm{C}^{*}(=\mathrm{O})-\mathrm{H}$ |
| ALLENE | $\mathrm{C}^{*}=\mathrm{C}=\mathrm{C}$ |
| AMIDE | AMIDE*1, AMIDE*2, or AMIDE*3 |
| AMIDE*1 | $\mathrm{C}^{*}(=\mathrm{O})-\mathrm{NH} 2$ |
| AMIDE*2 | $\mathrm{C}^{*}(=\mathrm{O})-\mathrm{NH}-\mathrm{C}$ |
| AMIDE*3 | $\mathrm{C}^{*}(=\mathrm{O})-\mathrm{N}(-\mathrm{C})-\mathrm{C}$ |
| AMIDZ | $\mathrm{C}^{*}-\mathrm{N}-\mathrm{C}=\mathrm{O}$ |
| AMINE | AMINE*1, AMINE*2, or AMINE*3 |
| AMINE*1 | $\mathrm{C}^{*}-\mathrm{NH} \mathrm{H}_{2}$ |
| AMINE*2 | $\mathrm{C}^{*}-\mathrm{NH}-\mathrm{C}$ |


| AMINE*3 | C*-N(-C)-C |
| :---: | :---: |
| AMINE*OXIDE | $\mathrm{C}^{*}-\mathrm{N}^{+}\left(-\mathrm{O}^{-}\right)(-\mathrm{C})-\mathrm{C}$ |
| ANHYDRIDE | $\mathrm{C} *(=\mathrm{O})-\mathrm{O}-\mathrm{C}=\mathrm{O}$ |
| AZIDE | $\mathrm{C}^{*}-\mathrm{N}=\mathrm{N}^{+}=\mathrm{N}^{-}$ |
| AZIRIDINE | $\mathrm{C}^{*}-\mathrm{N}$ |
|  | \/ |
|  | C |
| AZO | C*-N=N-C |
| BROMIDE | C*-Br |
| C*SULFONATE | $\mathrm{C}^{*}-\mathrm{SO}_{2}-\mathrm{O}$ |
| C*SULPHONATE | Synonym for C*SULFONATE |
| CARBAMATE*C | $\mathrm{N}(-\mathrm{C})(-\mathrm{C})-\mathrm{C} *(=\mathrm{O})-\mathrm{O}-\mathrm{C}$ |
| CARBAMATE*H | NH-C* $=0$ )-O |
| CARBONIUM | $\mathrm{C}^{+}$ |
| CARBONYL | ALDEHYDE or KETONE |
| CARBOXYL | ACID, ESTER, AMIDE, ACID*HALIDE, THIOESTER, or ANHYDRIDE |
| CHLORIDE | $\mathrm{C}^{*}-\mathrm{Cl}$ |
| CYANO | C*\#N |
| DIAZO | $\mathrm{C}^{*}=\mathrm{N}^{+}=\mathrm{N}^{-}$ |
| DISULFIDE | C*-S-S-C |
| DISULPHIDE | Synonym for DISULFIDE |
| DITHIOACETAL | RS-C*H-SR ${ }^{\prime}$ |
| DITHIOKETAL | RS-C* (-C)(-C)-SR ${ }^{\prime}$ |
| ENAMINE | $\mathrm{C}=\mathrm{C}^{*}-\mathrm{N}(-\mathrm{C})-\mathrm{C}$ |
| ENOL*ETHER | $\mathrm{C}=\mathrm{C}^{*}$-O-C |


| EPISULFIDE | $\mathrm{C}^{*}-\mathrm{S}$ |
| :---: | :---: |
|  | \/ |
|  | C |
| EPISULPHIDE | Synonym for EPISULFIDE |
| EPOXIDE | C*-O |
|  | \/ |
|  | C |
| ESTER | $\mathrm{C}^{*}(=\mathrm{O})-\mathrm{O}-\mathrm{C}$ |
| ESTERX | C*-O-C=O |
| ETHER | C*-O-C |
| FLUORIDE | C*-F |
| FUNCTIONAL | Any of the groups defined in this table |
| GEM*DIHALIDE | $\mathrm{C}^{*} \mathrm{XY}[\mathrm{X}, \mathrm{Y}=\mathrm{F}, \mathrm{Cl}, \mathrm{Br}$, or I$]$ |
| GLYCOL | R-O-C*-C-O-R' ${ }^{\text {R, }} \mathrm{R}^{\prime}=\mathrm{C}$ or H$]$ |
| HALIDE | BROMIDE, CHLORIDE, FLUORIDE, or IODIDE |
| HALOAMINE | C*-N-X [ $\mathrm{X}=\mathrm{F}, \mathrm{Cl}, \mathrm{Br}$, or I] |
| HALOHYDRIN | X-C*-O-R [ $\mathrm{X}=\mathrm{F}, \mathrm{Cl}, \mathrm{Br}$, or $\mathrm{I} ; \mathrm{R}=\mathrm{C}$ or H$]$ |
| HEMIACETAL | C-O-C* $\mathrm{H}-\mathrm{OH}$ |
| HYDRATE | HO-C*-OH |
| HYDRAZONE | $\mathrm{C} *=\mathrm{N}-\mathrm{N}$ |
| HYDROXYL | Synonym for ALCOHOL |
| HYDROXYLAMINE | C*-N-OH |
| IMINE | $\mathrm{C}^{*}=\mathrm{N}-\mathrm{R}[\mathrm{R}=\mathrm{C}$ or H$]$ |
| IODIDE | C*-I |
| ISOCYANATE | $\mathrm{C}^{*}-\mathrm{N}=\mathrm{C}=\mathrm{O}$ |
| ISOCYANIDE | $\mathrm{C}^{*}-\mathrm{N}^{+} \equiv \mathrm{C}^{-}$ |


| KETONE | $\mathrm{C}-\mathrm{C}^{*}(=\mathrm{O})-\mathrm{C}$ |
| :---: | :---: |
| LACTAM | $\mathrm{O}=\mathrm{C}^{*}-\mathrm{NH}$ |
|  | \/ |
|  | $\mathrm{C}_{\mathrm{n}}$ |
| LACTONE | $\mathrm{O}=\mathrm{C}^{*}-\mathrm{O}$ |
|  | \/ |
|  | $\mathrm{C}_{\mathrm{n}}$ |
| METHYLENE | $\mathrm{C} *=\mathrm{CH}_{2}$ |
| N*CARBAMATE | $\mathrm{C}^{*}-\mathrm{N}-\mathrm{C}(=\mathrm{O})-\mathrm{O}-\mathrm{C}$ |
| N*UREA*C | $\mathrm{C}^{*}-\mathrm{N}(-\mathrm{C})-\mathrm{C}(=\mathrm{O})-\mathrm{N}(-\mathrm{C})-\mathrm{C}$ |
| N*UREA*H | $\mathrm{C}^{*}-\mathrm{N}-\mathrm{C}(=\mathrm{O})-\mathrm{NH}$ or $\mathrm{C}^{*}-\mathrm{NH}-\mathrm{C}(=\mathrm{O})-\mathrm{N}$ |
| NITRILE | Synonym for CYANO |
| NITRITE | C*-O-N=O |
| NITRO | $\mathrm{C}^{*}-\mathrm{N}^{+}(=\mathrm{O})-\mathrm{O}^{-}$ |
| NITROSO | $\mathrm{C}^{*}-\mathrm{N}=\mathrm{O}$ |
| O*CARBAMATE | $\mathrm{C}^{*}-\mathrm{O}-\mathrm{C}(=\mathrm{O})-\mathrm{N}$ |
| O*CARBONATE | $\mathrm{C}^{*}-\mathrm{O}-\mathrm{C}(=\mathrm{O})-\mathrm{O}-\mathrm{C}$ |
| O*SULFONATE | C*-O-SO ${ }_{2}-\mathrm{R}[\mathrm{R}=\mathrm{C}$ or H$]$ |
| O*SULPHONATE | Synonym for O*SULFONATE |
| OLEFIN | C*=C |
| OXIME | $\mathrm{C}^{*}=\mathrm{N}-\mathrm{OH}$ |
| PEROXIDE | $\mathrm{C}^{*}$-O-OH or $\mathrm{C}^{*}$-O-O-C |
| PHOSPHINE | C*-P(-C)-C |
| PHOSPHONATE | C*-O-P |
| SELENIDE | C*-Se-C |
| SILYLENOLETHER | $\mathrm{C}=\mathrm{C}^{*}-\mathrm{O}-\mathrm{Si}(-\mathrm{C})(-\mathrm{C})-\mathrm{C}$ |
| SULFIDE | C*-S-C |


| SULFONE | $\mathrm{C}^{*}-\mathrm{SO}_{2}-\mathrm{C}$ |
| :--- | :--- |
| SULFOXIDE | $\mathrm{C}^{*}-\mathrm{S}(=\mathrm{O})-\mathrm{C}$ |
| SULPHIDE | Synonym for SULFIDE |
| SULPHONE | Synonym for SULFONE |
| SULPHOXIDE | Synonym for SULFOXIDE |
| THIOCYANATE | $\mathrm{C}^{*}-\mathrm{S}-\mathrm{C} \equiv \mathrm{N}$ |
| THIOESTER | $\mathrm{C}^{*}(=\mathrm{O})-\mathrm{S}-\mathrm{C}$ |
| THIOL | $\mathrm{C}^{*}-\mathrm{SH}$ |
| TRIALKYLSILOXY | $\mathrm{C}^{*}-\mathrm{O}-\mathrm{Si}(-\mathrm{C})(-\mathrm{C})-\mathrm{C}$ |
| TRIALKYLSILYL | $\mathrm{C}^{*}-\mathrm{Si}(-\mathrm{C})(-\mathrm{C})-\mathrm{C}$ |
| TRIHALIDE | $\mathrm{C}^{*} \mathrm{XYZ}[\mathrm{X}, \mathrm{Y}, \mathrm{Z}=\mathrm{F}, \mathrm{Cl}, \mathrm{Br}$, or I] |
| VIC*DIHALIDE | $\mathrm{X}-\mathrm{C} *-\mathrm{C}-\mathrm{Y}[\mathrm{X}, \mathrm{Y}=\mathrm{F}, \mathrm{Cl}, \mathrm{Br}$, or I] |
| VINYLSILANE | $\mathrm{C}=\mathrm{C} *-\mathrm{Si}(-\mathrm{C})(-\mathrm{C})-\mathrm{C}$ |

a) Note that some group names are not unique - e.g. ALCOHOL and HYDROXYL are synonyms
b) The atom on which the group is centred is indicated with an asterisk. It is always a carbon atom.
c) Note that the definition for ACETAL in the original CHMTRN, and currently retained, includes ketals.

Table S8. Other substructures

| Reserved word | Notes |
| :---: | :---: |
| ALKYL*ATOM(S) | An sp ${ }^{3}$ hybridised carbon atom with no bonds to heteroatoms. |
| APPENDAGE(S) | Allows a pair of substructures within a structure to be compared. See the section entitled 'Appendages'. |
| ISOPROPYL | $\mathrm{CH}_{3}-\mathrm{CH}^{*}-\mathrm{CH}_{3}$ |
| METHYL*ATOM(S) | A carbon atom with three hydrogens (explicit and/or implicit) attached to it. |
| PHENYL | $\mathrm{CH}=\mathrm{CH}$ |
|  | 11 |
|  | C* CH |
|  | 11 /1 |
|  | $\mathrm{CH}-\mathrm{CH}$ |
| T*BUTYL | $\mathrm{CH}_{3}-\mathrm{CH} *\left(-\mathrm{CH}_{3}\right)-\mathrm{CH}_{3}$ |

Table S9. Group Properties

| Reserved word | Notes ${ }^{\text {a }}$ |
| :---: | :---: |
| DONATING | Any of ACETAL, ALCOHOL, AMIDZ, AMINE, ESTERX, ETHER, HALOAMINE, HEMIACETAL, HYDROXYLAMINE, N*CARBAMATE, $\mathrm{N}^{*}$ UREA*C, $\mathrm{N}^{*}$ UREA*H, O*CARBAMATE, O*CARBONATE, PHOSPHINE, SELENIDE, SULFIDE, THIOL, TRIALKYLSILOXY, or VINYLD |
| EXPANDABLE*WITHDRAWING ${ }^{\text {b }}$ | Any of ACID*HALIDE, ALDEHYDE, AMIDE, ANHYDRIDE, CARBONIUM, ESTER, KETONE, IMINE, NITRILE, OXIME, THIOESTER, TRIHALIDE, or |


|  | VINYLW |
| :---: | :---: |
| GOOD*LEAVING | Any of AZIRIDINE, BROMIDE, CHLORIDE, <br> EPISULFIDE, EPOXIDE, IODIDE, O*SULFONATE, PHOSPHONATE, GEM*DIHALIDE, or TRIHALIDE |
| GOOD*LEAVING*GROUP | Synonym for GOOD*LEAVING |
| INTERFERING | The transform writer can designate a group as one that will interfere with the reaction described by a transform and is likely to be unprotectable, so that the end-user's attention can be drawn to the problem. An automatic decrement may also be applied to the transform by the program. See also PROTECTED. |
| LEAVING | Any of ACETAL, ALCOHOL, AMIDZ, AZIRIDINE, CARBAMATE ( N - and O -), DITHIOACETAL, <br> DITHIOKETAL, <br> EPISULFIDE, EPOXIDE, ESTERX, ETHER, GEM*DIHALIDE, HALIDE, N*UREA*C, N*UREA*H, O*CARBONATE, PHOSPHONATE, SELENIDE, SULFIDE, SULFONE, SULFONATE (C- and O-), THIOCYANATE, TRIALKYLSILYOXY, or TRIHALIDE |
| LEAVING*GROUP | Synonym for LEAVING |
| NON*EXPANDABLE*WITHDRAWING ${ }^{\text {b }}$ | NITRO, C*SULFONATE, SULFONE, or SULFOXIDE |
| PROTECTED | The transform writer can designate a group as one that the program should treat as being protected while the reaction is carried out, so that it can be appropriately labelled in displays and does not trigger automatic decrements to the transform rating. See also INTERFERING. |
| PARTICIPATING | The transform writer can indicate that what might appear to be a functional group that would compete with the reaction described by a transform is not to be taken into account when assessing the need for protection. |
| VINYL*D | $\mathrm{C}^{*}=\mathrm{C}-\mathrm{D}[\mathrm{D}=$ donating group, q.v.] |

VINYL*W
WITHDRAWING
$\mathrm{C}^{*}=\mathrm{C}-\mathrm{W}[\mathrm{W}=$ withdrawing group, q.v.]
Any EXPANDABLE*WITHDRAWING or NONEXPANDABLE*WITHDRAWING group
a) The values of properties are Boolean.
b) An expandable withdrawing group is one that activates the atom(s) alpha to the origin of the group, (e.g., ketone, ester); a non-expandable withdrawing group actives the origin of the group itself (e.g. nitro). So WITHDRAWING BONDs associated with a non-expanding withdrawing group are the bonds on the origin itself (excluding those that are defined as part of the group); WITHDRAWING BONDs associated with an expandable withdrawing group are the bonds on the atom(s) alpha to the origin (excluding the atoms that are defined as part of the group).

Table S10. Atom and Bond Sets

| Reserved word ${ }^{\text {a }}$ | Notes |
| :--- | :--- |
| ALPHA | The atoms alpha to a specified atom and the bonds on them, <br> excluding the bonds back to the specified atom (e.g. <br> "ALPHA TO ATOM*1"). |
| ATOMS*TO*ATOMS | The atoms alpha to a specified atom or set of atoms. <br> The bonds on a specified atom or set of atoms. |
| ATOMS*TO*BONDS | The atoms and bonds beta to a specified atom or set of atoms <br> (e.g. "BETA TO ATOM*1"). |
| BETA | The set of bonds in a structure/substructure. |
| BOND(S) | The atoms attached to a specified bond or set of bonds. |
| BONDS*TO*ATOMS |  |
| BONDS*TO*BONDS |  |
| CURRENT*RING |  |
| bonds (excluding the specified bond(s)). |  |


| GAMMA | lowest numbered atom. <br> The atoms and bonds gamma to a specified atom or set of <br> atoms (e.g. "GAMMA TO ATOM*1"). |
| :--- | :--- |
| IT | The atom, bond, or set of either, identified in the preceding <br> CHMTRN statement. <br> Synonym of IT. Intended only for use together with the <br> buzz word PREVIOUS for clarity (e.g. "IF THERE IS A <br> NITROGEN ATOM ALPHA TO THE PREVIOUS <br> LOCANT"). |
| LOCANT(S) |  |
| RING(S) | The set of all rings in a structure/substructure. Supports <br> statements such as "IF ATOM*1 IS IN A RING..." |
| SAVED*ATOM(S) n | The set of atoms saved using "SAVE AS n...". |
| SAVED*BOND(S) n | The set of bonds saved using "SAVE AS n...". |
| SPECIFIED*ATOM(S) n | A set of atoms passed to a subroutine or an atom passed to a <br> DO loop. |
| SPECIFIED*BOND(S) n | A set of bonds passed to a subroutine or a bond passed to a <br> DO loop. |
| SPANNING | The bond between a specified pair of atoms and/or bonds. <br> Both atoms must be specified, separated by the word <br> SPANNING (e.g. "IF THE BOND SPANNING ATOM*1 |
| AND ATOM* IS A DOUBLE*BOND THEN ..."). |  |

a) $n$ represents an integer.

Table S11. Properties of Sets

| Reserved word ${ }^{\text {a }}$ | Notes ${ }^{\text {b }}$ |
| :---: | :---: |
| BRIDGE(D) ${ }^{\text {c }}$ | See the section entitled 'Properties Relating to Rings'. |
| CARBOCYCLIC | Ring property - the ring contains only carbon atoms. |
| HETEROCYCLIC | Ring property - the ring contains at least one noncarbon atom. |
| FEWER THAN aaa | Tests the number of objects in a set (e.g. "IF THERE ARE FEWER THAN TWO HYDROGENS ON ATOM*1 ...". |
| FEWER*THAN aaa | Synonym for FEWER THAN. |
| LARGER | See OR LARGER. |
| NUMBER | The number of members of a set of atoms or bonds. Returns an integer (e.g. "MYVAR = THE NUMBER OF CHLORINE ATOMS ALPHA TO ATOM*1"). |
| OR LARGER | Used only in connection with ring size (e.g. "IF ATOM* 1 IS IN A RING SIZE 7 OR LARGER ..." |
| MORE THAN aaa | Tests the number of objects in a set (e.g. "IF THERE IS MORE THAN ONE CHLORINE ALPHA TO ATOM*1 ...". |
| MORE*THAN aaa | Synonym for MORE THAN. |
| SIZE(S) n SIZE(S) n THROUGH n | Used only together with RING, to specify the ring size (e.g. "IF ATOM*1 IS IN A RING SIZE 5"; "IF ATOM* 1 IS IN A RING OF SIZE 5 THROUGH 7"). |

a) n represents a positive integer; aaa represents an English language word for a number between one and twenty (i.e. included in Table S12).
b) The values of properties are Boolean unless otherwise indicated.
c) BRIDGE and BRIDGED are synonymous. "BRIDGE(D)" (i.e. with "D" in parenthesis) is not part of the language.

Table S12. Named Numerical Values

ONE
TWO
THREE
FOUR
FIVE
SIX
SEVEN
EIGHT
NINE
TEN
ELEVEN
TWELVE
THIRTEEN
FOURTEEN
FIFTEEN
SIXTEEN
SEVENTEEN
EIGHTEEN
NINETEEN
TWENTY

Table S13. Logical Variables

| Reserved word | Notes |
| :--- | :--- |
| FAIL | Boolean False (return value from a subroutine). |
| SUCCESS | Boolean True (return value from a subroutine). |
| SUCCESSFUL | Tests whether SUCCESS was returned by a preceding call to a <br> subroutine. |
| UNSUCCESSFUL | Tests whether FAIL was returned by a preceding call to a subroutine. |

Table S14. Operators ${ }^{\text {a }}$

| Reserved word ${ }^{\text {b }}$ | Notes |
| :--- | :--- |
| - | Arithmetic subtraction. |
| $\%$ | Integer arithmetic modulus. |
| $/$ | Arithmetic division. |
| + | Arithmetic addition. |
| $=$ | Arithmetic equals symbol. |
| .EQ. | Relational equality operator. |
| .GE. | Relational greater than or equal operator. |
| .GT. | Relational less than or equal operator. |
| .LE. | Relational less than operator. |
| .LT. | Relational not equal operator. |
| .NE. | Boolean operator for use in arithmetic conditional statements <br> - e.g. "IF (VAR1 .EQ. 2 .OR. VAR1 = 3) THEN ...". |
| .OR. |  |


| ADD n | Increases the numerical rating for the transform by the specified amount. See the section entitled 'Rating'. |
| :---: | :---: |
| AND: | See the section entitled 'Conditional Statements' in the main paper to which this is a supplement. |
| CLEAR | Empties an atom or bond set. Removes a previous designation (see DESIGNATE). E.g. "CLEAR SAVED*ATOM 1"; "CLEAR THE CURRENT*RING") |
| CONTAINED | Tests whether a query is a member (or subset) of a set (e.g. "IF ATOM*3 IS CONTAINED IN THE CURRENT*RING ..."). See also SAME. |
| DECREMENT | Synonym for SUBTRACT (q.v.). |
| DESIGNATE | Assigns the contents of a set, or attaches a property to a set (e.g. "DESIGNATE AS THE CURRENT*RING THE HETEROCYCLIC RING CONTAINING ATOM*3"; "DESIGNATE THE GROUP ON SAVED*ATOM 2 AS INTERFERING"). |
| GETS | Initialises the contents of a set that has been declared by the writer (e.g. "MYSET GETS BENZYLIC ATOMS"). |
| INCREMENT | Synonym for ADD. |
| LOWER*RATING ${ }^{\text {c }}$ | See the section entitled 'Rating' in the main paper to which this is a supplement. |
| NO | Synonym of NOT (q.v.) for use in statements such as "IF THERE IS NO HYDROGEN ON ATOM* $1 . . . "$. |
| NON | Tests whether the query object is a member of the set of objects not having a specified property (e.g. "IF ATOM*1 IS NON AROMATIC ..." means "if atom 1 is a member of the set of atoms that are not aromatic"). See also NOT. |
| NOT | Tests whether the query object is absent from a specified set of objects having a specified property (e.g. "IF ATOM*1 IS NOT AROMATIC ..." means "if atom 1 is not a member of the set of atoms that are aromatic"). See also NON. |
| OR: | See the section entitled 'Conditional Statements' in the main paper to which this is a supplement. |

PERMUTE

PUT:

RAISE*RATING ${ }^{\text {c }}$

REMOVE:

SAME

SAVE AS n

SAVE*SEQUENTIALLY STARTING FROM n

SUBTRACT

THROUGH
TOGGLE BIT n

TOGGLE FLAG aaa

Changes the formal bond order of a specified bond in an aromatic ring from single to double or vice versa, adjusting the orders of the other bonds in the aromatic system accordingly (e.g "PERMUTE AT BOND*4").

Adds an object to a set (e.g. "PUT: THE ATOMS ALPHA TO ATOM* 1 INTO MYATMS").

See the section entitled 'Rating' in the main paper to which this is a supplement.

Removes a member from a set (e.g. "REMOVE: ATOM*1 FROM MYATMS"). Note the inclusion of the colon - see also REMOVE in Table S15.

In terms of program behaviour, SAME is a synonym of CONTAINED (i.e., it tests whether the query is a member of a specified set). However, to avoid confusing wording, SAME should be used only when the query is known to be a single object and the second reference is to a set that is known to have only one member (e.g. "IF ATOM*1 IS THE SAME AS SAVED*ATOM 4", where it is known that there is only one atom in SAVED*ATOM 4; if that is not known, it is preferable to write "IF ATOM*1 IS CONTAINED IN SAVED*ATOM 4").

The keying word is "SAVE" alone but it is normally followed by "AS" for clarity. 32 atom sets and 32 bond sets are available for use by the CHMTRN writer (e.g. "SAVE AS 5 THE ATOMS WITHIN GAMMA TO ATOM*2" would put those atoms into saved atom set 5).

Creates saved atom sets each containing a single member from a specified input set. The saved sets are numbered sequentially starting from n (e.g. "SAVE*SEQUENTIALLY STARTING FROM 3 THE ATOMS WITHIN GAMMA TO ATOM*2").

Decreases the numerical rating for the transform by the specified amount. See the section entitled 'Rating'.

Used only with SIZE(S), q.v.
Changes BIT n in the SPECIAL*SET from 1 to 0 or vice versa. The value of n must be between 1 and 64

Changes the value of FLAG aaa from true to false or vice versa.

TURN OFF BIT n
TURN OFF FLAG aaa
TURN ON BIT n
TURN ON FLAG aaa
ZERO

Sets BIT n in the SPECIAL*SET to 0 .
Sets the value of FLAG aaa to false.
Sets BIT n in the SPECIAL*SET to 1 .
Sets the value of FLAG aaa to true.
The empty set (e.g. "MYSET GETS ZERO" makes MYSET an empty set).
a) Operators used in mechanism statements are listed in Table S15.
b) aaa represents a previously-declared name for a Boolean variable; n represents an integer.
c) Terms that are not part of the original CHTMRN language. They have been added in the current project.

Table S15. Terms used in mechanism statements

| Reserved word | Notes |
| :--- | :--- |
| ANIONIZE | Assigns a negative charge to a specified atom. <br> Attaches a functional group to a carbon atom, which <br> becomes the origin of the group (e.g. ATTACH AN <br> ALDEHYDE TO ATOM*3 makes atom 3 the carbon atom <br> of a new aldehyde). See also INTRODUCE. <br> Removes a bond of any order (e.g. BREAK BOND*1). |
| BREAK | Assigns a positive charge to a specified atom. <br> CATIONIZE <br> DELETE |
| Removes an atom and the bonds connected to it (e.g. <br> DELETE ATOM*1). <br> DEFINED*ANTI <br> Defines the stereo-relationship between substituents on a <br> newly-formed double bond or the diastereomeric <br> relationship between a pair of atoms and/or bonds one of <br> which is on a new asymmetric centre. See the section <br> entitled 'Stereochemistry'. <br> DEFINED*CIS | Defines the stereo-relationship between a newly-created <br> stereocentre on a ring and an existing one (e.g. <br> INTRODUCE A NITRILE ON ATOM*2 DEFINED*CIS |



SEPARATE

SINGLE
TRIPLE
precursor the same as in the target (e.g. RETAIN AT ATOM*3).

Removes completely any bonding between a pair of atoms (e.g. SEPARATE ATOM*1 AND ATOM*2).

Changes the order of a bond to single.
Changes the order of a bond to triple.
a) If a group containing an alkyl group is attached or introduced, the alkyl group defaults to a carbon atom carrying no other carbon or heteroatoms. E.g. ATTACH AN ESTER TO $\mathrm{ATOM}^{*} 1$, causes ATOM* 1 to be extended to $\mathrm{C}(=\mathrm{O})-\mathrm{O}-\mathrm{C}[\mathrm{HS}=3]$.

Table S16. Control and conditional terms

| Reserved word ${ }^{\text {a }}$ | Notes |
| :---: | :---: |
| ... | The contents of a line of code that follow are comments. |
| $\ldots$ | Marks the start and end of the mechanism statements (see the section entitled 'Mechanism Statements'). |
| () | Used in the usual way in programming code that handles arithmetic and associated relational operations [e.g. "VAL = (V1 + V2 + V3)/ V4"; "IF (V1 .GT. V2 OR V3 .EQ. V4) THEN ..." |
| AND:IF | See the section entitled 'Conditional Statements'in the paper to which this is a supplement. |
| AND*THEN | See the section entitled 'Conditional Statements' Statements'in the paper to which this is a supplement. |
| BEGIN aaa | See the section entitled 'Conditional Statements' Statements' in the paper to which this is a supplement. |
| BLKEND aaa | See the section entitled 'Conditional Statements' Statements' in the paper to which this is a supplement. |
| BRANCH | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |


| CALL aaa | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| :---: | :---: |
| DO aaa | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| END*TRANSFORM | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| ENDDO | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| EXIT aaa | See the section entitled 'Conditional Statements' Statements' in the paper to which this is a supplement. |
| FOR EACH | Causes an operation to be applied iteratively for members of a set of atoms or bonds. See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| FOR*EACH | Synonym for FOR EACH. |
| FOREACH | Synonym for FOR EACH. |
| KILL | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| GO (TO) aaa | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| OR:IF | See the section entitled 'Conditional Statements' Statements' in the paper to which this is a supplement. |
| THEN | See the section entitled 'Conditional Statements' Statements' in the paper to which this is a supplement. |
| THEN*TO aaa | See the section entitled 'Control Statements' Statements'in the paper to which this is a supplement. |
| TRANSFORM n | A unique integer identifier for the transform. |
| RETURN TRUE | Causes a subroutine to return control to the calling transform, returning SUCCESS. |
| RETURN FALSE | Causes a subroutine to return control to the calling transform, returning FAIL. |
| SUBROUTINE | The header for a block of code that is not part of a transform but can be called from any transform. |

A message to be displayed to the end user. The integer, $n$, is a globally unique identifier for the message; aaa is the text of the message.
a) aaa represents a string of characters; n represents a positive integer.

Table S17. Other reserved words

| Reserved word | Notes |
| :--- | :--- |
| AND | Word separator required in statements such as "THE BOND <br> SPANNING ATOM*2 AND ATOM*5". <br> BIT n <br> One of the bits in the SPECIAL*SET (q.v.). n must be an integer <br> between 1 and 64. <br> EMPTY |
| Fsed only in connection with the SPECIAL*SET (q.v.). "IF |  |
| THE SPECIAL*SET IS EMPTY ..." returns true if all bits in the |  |
| SPECIAL*SET are off (i.e. all bits are binary zero) and otherwise |  |
| returns false. |  |

Table S18. Words Used in Declarative Statements

| Reserved word | Notes $^{\text {a }}$ |
| :--- | :--- |
| SET(S) aaa, aaa ... | Declares a list of names of sets. $^{\text {SET*ARRAY(S) aaa, aaa ... }}$ |
| VARIABLE(S) aaa, aaa ... | Declares a list of names of arrays of sets. |
| VARIABLE*ARRAY(S) aaa, aaa ... | Declares a list of names of variables. |
| a) list of names of arrays of variables. |  |
| a) Declaration of sets, arrays, and variables is optional in the current implementation. |  |

Table S19. Buzz Words

A
ALSO
AN
AND:FOR
ANOTHER
ANY
ANYWHERE
ARE
ARRAY
ARRAYS
AS
AT
ATOM
ATOMS
ATTACHED
AWAY
BE

BEARING
BETTER
BONDED
BY
CAN
CENTER
CENTRE
CHANGE
CHARACTER
CHARGE
CODE
CONTAIN
CONTAINING
CONTAIN
CONTAINS
COPY
DIMENSIONED
DOES
DUPLICATE
ELEMENT
ELEMENTS
EMPTY
EXACTLY
EXO
FACE
FOR
FROM

## GROUP

## GROUPS

HAS
HAVE
IDENTICAL
IF
IN
INCLUDING
INTO
ION
IS
LENGTH
LIST
MECHANISM
MEMBER
MEMBERS
MOLECULE
MOLECULES
MUST
NEXT
OF
OFFSPRING
OLD
OLEFIN*TYPE
ON
ONLY
ORDER

## PARENT

PATH
PATHS

## PRECURSOR

PRESENT
PREVIOUS
REQUIRED
SHORTEST
SIMULTANEOUSLY
STRUCTURAL
TARGET
TEST
THAT
THE
THERE
TO
UNIT
UNITS
UP
VALUE
VALUES
WHICH
WITH
WORD
WORDS
WRITE

