## LITERATURE CITED

(1) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968, p 14.
(2) Palmer, G., "Wiswesser Line-Formula Notation," *Chem. Brit.*, **6**, 422–426 (1970).
(3) Campey, L. H., Hyde, E., and Jackson, A., "Interconversion of Chemical Structure Systems," *Chem. Brit.*, **6**, 427–430 (1970).
(4) Granito, C. E., and Garfield, E., "Substructure Search and Correlation in the Management of Chemical Information," *Naturwissenschaften*, **60**, 189–197 (1973).
(5) Feldman, R. J., and Koniver, D. A., "Interactive Searching of Chemical Files and Structural Diagram Generation from Wiswesser Line Notation," *J. Chem. Doc.*, **11**, 154–159 (1971).
(6) Miller, G. A., "Encoding and Decoding WLN," *J. Chem. Doc.*, **12**, 60–67 (1972).
(7) Farrell, C. D., Chauvenet, A. R., and Koniver, D. A., "Computer Generation of Wiswesser Line Notation," *J. Chem. Doc.*, **11**, 52–59 (1971).
(8) Bowman, C. M., Landee, F. A., Lee, N. W., and Reslock, M. H., "A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures," *J. Chem. Doc.*, **8**, 133–138 (1968).
(9) Reference 1, p 237.
(10) Reference 1, p 16.


# Computerized Management of Structure–Activity Data. III. Computerized Decoding and Manipulation of Ring Structures Coded in WLN


DAVID ELKINS,† A. LEO,* and CORWIN HANSCH

Department of Chemistry, Pomona College, Claremont, California 91711

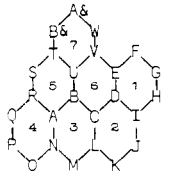

Received December 21, 1973

**Construction of the locant path, a table of interatom connections, and the molformula from the WLN for complex ring systems (including polycyclic, spiro, perifused, bridged, and pseudo-bridged rings) is described. The procedure has been programmed for a digital computer and has been found useful in manual decoding also.**

WLN Ring Code analysis can be broken down into three principal tasks: (1) dissection of the WLN into its constituent parts (ring locants, ring numerals, nonconsecutive locant pairs, bridges, etc.), (2) construction of the locant path and a table of interatom connections, and (3) molformula calculation. The first step presents no real problem to the experienced human decoder and it turns out to be a straightforward task in programming. For this reason it will not be discussed in this paper.

## LOCANT PATH

The guiding principle which underlies the procedure for developing the locant path from the WLN is simple: extend the locant path by using successively later alphabetic symbols, *only after* making sure that the information already processed has not established part of the particular ring pathway being sought. In other words, in the later rings of a fused ring system, it is often possible to get a "running start" on a new ring by using the locants which are common to some of the earlier rings.



1

T D6 C6666 B6 T6 5ABCDU B&J

In the fused ring system 1, the pathway begins at "A" but does not complete its first ring until it reaches "I." The second ring begins with the earliest locant, "C", and continues to later (not always *successively* later) locants, as will be explained below, until the ring contains the number of atoms specified. This also holds true for rings 3 and 4. Ring 5 begins with "A" and the *forward* path is along the route of *latest* previous connection which is to "R." The six-membered ring is not completed with "V," however. The information already in hand that the bonds to "A" were fulfilled with ring 4 should have been used to take a "running start" at ring 5 and begin it with "B." Similarly, ring 6 had "B" as its earliest locant, but the information already processed leads us to begin its running start as E–D–C–B and to complete the ring with "U" and "V."

This procedure is useful for both automatic and manual decoding. It has the advantage over graphical methods in that the structural diagram need not be attempted until the analysis is complete. Granito, *et al.*,[1] mention a similar approach in converting WLN to Ring Codes. Although it is a simple principle, its application is somewhat complex if it is to be applied by computer to the most complex perifused ring systems, as will be seen in the following section.

The first step in the process is the creation of a seven-column "Ringdecod" table containing a row for each locant in the notation. See example number 1. (It will be readily apparent that for manual use this rather formidable table can be greatly simplified or dispensed with altogether. It is a convenient format for computer use and for explanatory purposes only.) The row order is "A" through "W," "A&" through "W&" followed by the cited branch locants in the same order. If the "last locant" is not cited in the WLN, an estimate can be made and space left for possible additions, or it can be calculated from the formula[2] $n = s - 2(r - 1) - b$, where $n$ = the numerical equivalent in the alphabet of the "last locant," $s$ = the sum of the ring numerals in the notation, $r$ = the number of ring numerals, and $b$ is the number of branch locants. The contents of the columns in the "Ringdecod" table are as follows:

† Present address: G. D. Searle & Co., Chicago, Ill. 60680.
* To whom correspondence should be addressed.

(1) The locant.
(2) The corresponding branch locant, if cited.
(3) The lower member of a cited nonconsecutive locant (NCL) pair.
(4) Other locants connected to locant in (1).
(5) Maximum number of connections allowed for locant in (1).
(6) Indices of the rings in which locant in (1) appears.
(7) Maximum number of rings in which locant in (1) may appear.

At the outset, the following *initial* conditions are assumed.

(a) Each locant (except a branch locant) is connected *only* to the locants adjacent in the locant alphabet. For manual use these are given the subheading of "given."

(b) As the ring structure is developed, each locant may be connected to no more than three locants except that bridges must have only two connections and X symbols must have four. These additional connections may be subheaded "developed."

(c) Each locant may appear in no more than two rings except that a bridge locant must appear in one more ring than the number of times it is cited as a bridge, and multicyclic points must appear in two more rings than the respective number of citations for each in the MCP list.

## RING INITIATION

Each ring is initiated by entering the cited (or "A" understood) locant as its tail and the *highest available known connection* as the head. This serves to establish one edge of the ring as well as giving direction to a pathway. Ring initiation proceeds in the order cited in the WLN.

At this point, a few definitions are required.

(1) The *tail* and *head* of the ring, at any given stage of processing, are the most recent entries to the backward and forward paths, respectively.

(2) A locant is *available* for entry into a ring if it has not already appeared in that ring, and if it has not yet appeared in the maximum number of rings permitted it.

(3) A *branch jump* occurs in the backward path when the tail has an available branch locant in which case the branch locant becomes the new tail (*e.g.*, the current tail is "E" and "E-" is cited in the WLN and is available).‡

(4) An *NCL jump* occurs in either path when the head or tail is the later of a cited NCL pair and the earlier of the pair is available. An NCL jump may occur only once per ring system per cited NCL pair. The earlier locant becomes the new head or tail.

## PATH INITIATION

After initiation, each ring is expanded in both directions but the *backward* path, if allowed, must be *exhausted* before the forward path is begun ("getting a running start"). The backward path is initiated if (1) there is a branch jump specified for the tail locant and the tail locant is not an "X" (whence the branch locant becomes the new tail; see example 2, ring 1), or if (2) the tail locant is cited as a bridge and both of its connections are known (whence the earlier of the known locants becomes the new tail; see example 3, ring 4), or if (3) the tail locant is a multicyclic point (MCP) with all its connections known *and* if the next later locant is an available MCP or bridge (whence it becomes the new tail; see example 3, ring 5).

Once initiated, the backward path is *continued* if any of the following apply to the current tail locant: (1) a branch jump is specified or (2) an NCL jump is specified (see example 3, ring 6, locant "S") or (3) it is an MCP with all connections known and the next later locant is an available MCP or bridge (whence the next later locant is the new tail) or (4) all connections are known (whence the latest available of these becomes the new tail; see example 3, ring 6, locants "R" and "S").

Only if all four criteria fail to be met can the forward path be initiated. It should be noted that the criteria are hierarchical; *i.e.*, (1) must fail before (2) can be tried, etc.

The forward path proceeds as follows: an NCL jump is taken whenever possible; otherwise add to the head the highest available locant already known.

## PATH COMPLETION

It can be readily seen that in filling the final position in a ring the head must be connected to, or connectable to, the tail. Occasionally the head will give no information regarding the final entry to the ring in which case the roles of head and tail may be reversed for the preceding step. This is illustrated by example 2 where reversal is required in closing the final ring. Some structures will be encountered where this reversal procedure is required to close an earlier ring, for example, in L6 G656 Q6 O5 N6 2AO C&J.[3]

As each position in the ring is filled, the "Ringdecod" table is updated. If at any stage there is no entry available or if the total number of connections or rings for MCPs, bridges, and X symbols is not correct at the end of the process, then the encoding of the WLN should be suspect.

## MOLFORMULA

Molformula calculation for the ring kernel is performed separately from that of the substituents. The latter are derived using the routine designed to handle acyclic branching chains and multiplied groups (see article II in this series). The molformula of the ring kernel is calculated using the Symbol Equivalence Table (Table I of article II in this series) together with the following two formulas:

$$C = n + b - m \quad (1)$$

where

$C$ = number of carbon atoms, not V, X, or Y
$n$ = last locant converted from alphabetic to numeric equivalent
$b$ = number of branch locants
$m$ = number of ring segment symbols (excluding "U")

$$H = \& + (T_2) + 2(T_1) + 2(BT) + (B\&) + h - u - 2(uu) \quad (2)$$

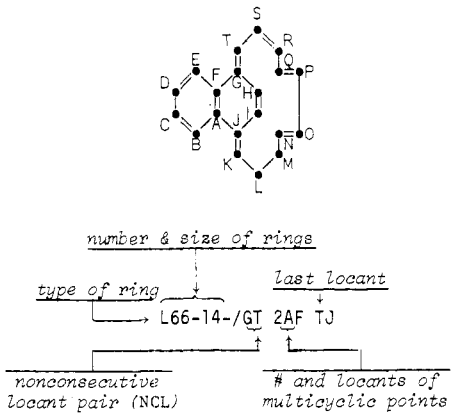
where

$H$ = number of hydrogen atoms on carbon atoms
& = number of *locants* in only one "&" ring
$T_2$ = number of locants *exclusively* in T rings and appearing more than once, but *not* the locant of a V, X, Y, bridge, or heteroatom
$T_1$ = number of locants in only one T ring and not V, X, Y, or hetero
$BT$ = number of bridge locants, *not* V, X, Y, or hetero and in *exclusively* T rings
$B\&$ = same as $BT$, but in at least one "&" ring
$h$ = hydrogen atoms cited in ring kernel
$u$ = number of *locants* connected to U symbols (not merely those cited), but *not* in "&" rings and not of heteroatoms
$uu$ = same as $u$ but connected to UU symbol

For the cited ring segment symbols, the atom count for the heteroatoms (N, O, S, SW, etc.) is the same as for the acyclic portion, but note that the symbol equivalence for K, X, and Y differs if they occur inside a ring. These are added to (1) and (2) for final molformula.

‡ Note that a branch jump is not taken on the forward path. The rules governing their use may prevent a branch locant from being encountered on the forward path. It can occur only once per locant per ring system. Also note that for the "Ringdecod" table, a branch locant is considered as coming later than the last locant (*e.g.*, "E" is followed by "F" and not "E-").

**Example 1**




Ring Index #8691

**Table I.** "Ringdecod" Table

| [1] Locant | [2] Branch jumps | [3] NCL jumps | [4] Connections Given | Derived | [5] Max no.[a] of connections | [6] Ring index | [7] Maximum no.[b] of rings |
|---|---|---|---|---|---|---|---|
| A | | | B | F,J | | 1,2,3 | 3 |
| B | | | A,C | | | 1 | |
| C | | | B,D | | | 1 | |
| D | | | C,E | | | 1 | |
| E | | | D,F | | | 1 | |
| F | | | E,G | A | | 1,2,3 | 3 |
| G | | | F,H | T | | 2,3 | |
| H | | | G,I | | | 2 | |
| I | | | H,J | | | 2 | |
| J | | | I,K | A | | 2,3 | |
| K | | | J,L | | | 3 | |
| L | | | K,M | | | 3 | |
| M | | | L,N | | | 3 | |
| N | | | M,O | | | 3 | |
| O | | | N,P | | | 3 | |
| P | | | O,Q | | | 3 | |
| Q | | | P,R | | | 3 | |
| R | | | Q,S | | | 3 | |
| S | | | R,T | | | 3 | |
| T | | G | S | G | | 3 | |

[a] Value is 3 if not specified. [b] Value is 2 if not specified.

Entries in Table I in the locant column are made up to "T" since it is specified as the last locant. There are no branch jumps to be entered (all branch locants must be citable by Rule 30c) and only one NCL pair which requires the entry of "G" in column [3], row "T." Under the subheading "given," the adjacent two connections for each locant can then be entered except for the first and last which initially have only one adjacent locant known. No bridges or X symbols appear in the WLN and so no entries are needed in column [5] with all assumed to be 3. Column [7] can be completed with the entry of "3" in rows "A" and "F" as indicated by the two cited multicyclic points.

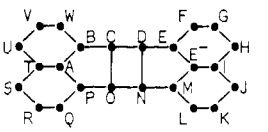
```
Ring  tail
        |  head
        |  |
(1)  A B C D E F
(2)  A F G H I J
(3)  A J K L MN O P Q R S T G F
```

Computed molformula: $C_{20}H_{16}$

Locant "A" is the initial tail for all three of the rings, and it is seen that none of the three criteria for initiating a backward path are met in any of the rings. Entries are made in the Ring Index column [6] as the first two rings are completed in a straightforward manner, *i.e.*, adding to the head the highest available locant already known, as shown in column [4]. There is no problem in closing either of the first two rings.

In ring (3), the NCL jump requires "G" to follow "T" in the head position. The final (14th) locant for this ring must now be found, one which is already connected to "A." Both "H" and "T" are connected to "G," but "T" has just been used in this ring and "H" is *not* connected to "A"; therefore an earlier locant, "F," must be chosen to make the final ring closure. Note that this information was obtained from the "Ringdecod" table without need to refer to the structural diagram.

**Example 2**



L E6 E-6 D5 C4566 2AE- WTJ[H]

```
Ring  tail  head
        |     |
(1)E- E F G H I
(2)   E-I J K L M
(3)   D E E-M N
(4)   C D N O
(5)   A B C O P
(6)   A P Q R S T
(7)   A T U V W B
```

Computed Molformula: $C_{24}H_{36}$

Criterion (1) initiating a backward path in ring (1) is met, and locant "E-" becomes the new tail. All four criteria for continuation of the backward path fail ("E-" is a multicyclic point but its connections are not all known) so the forward path is begun at the head, "F."

In ring (7) the locant "W" gives no information for the final position so the three connections to "A" are examined: "T" is already in this ring, "P" has already appeared in its maximum of two rings and is not available, leaving "B" which is connectable to "W."

For example 3, the program can print out the diagnostic information as shown.

Note that the backward paths are *initiated* by bridge "A" in ring (4) (criterion 2) and by MCP "B" in ring (5) (criterion 3). Both backward paths terminate because the new entry is not fully connected; *i.e.*, they do not meet any of the four criteria for backward path *continuation*. However, in ring (6) there is an example of the backward path completing the ring with no forward path needed after the initial head locant "O" is assigned. Note that "R," "S," and "N" are neither bridges nor MCP's. Also note the NCL jump from "S" to "N."

The WLN ring decoding program described above presently has the following limitations.

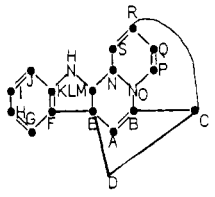
(1) The decoding of a "ring of rings" system is not implemented.

(2) Only "L," "L&," "L-," and "L&-" locant types are recognized.

(3) The maximum bonds per ring atom is 4.

(4) An MCP can be shared by no more than 4 rings.

**Example 3**



T5 F6 E56 B6 B6/CR/NS A 5BBCEO S EX LM ON AU CH MH NHT&&T&&J[5]

Computed Formula (including core symbols):

$C_{17}H_{14}N_2$

Stored Formula: same

| Ring | backward path | tail | head | forward path |
|---|---|---|---|---|
| (1) | ← | A | B C D E | → |
| (2) | | F | G H I J K | |
| (3) | | E | F K L M | |
| (4) | | B A | E M N O | |
| (5) | | C B | O P Q R | |
| (6) | N S R C B | O | | |

| | Ring types | | MCP's | Core symbols |
|---|---|---|---|---|
| 19 normal locants | A | 5T | B | E = X |
| 0 branch locants | F | 6& | B | L = M |
| 3 core symbols | E | 5& | C | O = N |
| 16 = normal + branch − core | A | 6T | E | |
| 7 & type locants | B | 6& | O | |
| 0 $T_2$ type locants | B | 6& | | |

| | | Unsaturations |
|---|---|---|
| 1 $T_1$ type locants | | A = U |
| | NCL's | |
| 1 $BT$ type locants | C R | B = U |
| 0 $B$& type locants | N S | |
| | | Saturations |
| 2 $h$ cited | Bridges = A | C, M, N |
| 13 = & + $T_2$ + 2$T_1$ + 2$BT$ + $B$& + $h$ | Last locant = S | |

(5) A bridge can be shared by no more than 2 rings.

(6) A given locant can be the later of only one cited NCL pair.

(7) An "H" or "W" substituent on a reapeated spiro atom is not removed.

In our files only the first of these is encountered with sufficient frequency to suggest a real need for it in the program.

## SYNTAX CHECKS

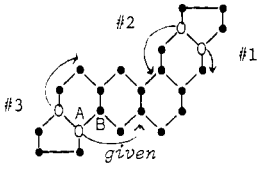
Many of the syntactical encoding rules are designed to produce a *unique* notation for each unique structure and thus a WLN with a syntactical error may yield the correct molformula and connection table and could, therefore, properly respond to search questions. However, since at present most WLN files are *not* checked for CT uniqueness as a requisite for entry and are not searched *via* CT, it is very important to reduce this type of error to a minimum. Of course there is a trade-off between encoding effort and computing effort. While it is conceivable that a foolproof syntax checking program could be designed, it would be forbiddingly expensive to operate.

Rule 30, which defines the path selection for polycyclic fused rings, presents problems of this magnitude if all ($n$) subsections are to be implemented. A program has been described[6] which accepts as input any trial WLN which completes the ring pathway. It then examines all pathways and, considering the first seven subsections of Rule 30, saves the one most proper canonically. However, unless some starting locants are forbidden to it (not very "risky" for an experienced encoder), it can still take a large computer (*e.g.*, IBM 370/155) hours to explore the millions of paths possible in a very complex ring system.

Some of the most common errors in interpretation of Rule 30 involve subsections d, e, g, h, i, j, k, and l. Many of these have been averted by the inclusion of a relatively simple subroutine in the ring decoding procedure described above. For rings containing no bridges or MCP's, it tests three additional paths.

In example 4, let $L_1$ = the latest cited ring fusion locant in the given notation ("I"), $L_2$ = the latest locant connected to $L_1$ ("M"), and $L_3$ = the next-to-latest locant connected to "A" ("U"). Then the three test paths are: (#1) starting at $L_1$ in the reverse direction of the given path, (#2) starting at $L_2$ in the direction of the given path, and (#3) starting at $L_3$ in the reverse direction of the given path.

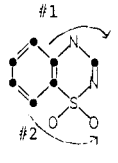
**Example 4**




Given = T I5 G6 D6 B665J[7]

From the "Ringdecod" table the program computes path #1 and finds as fusion locants: "I," "H," .... It discards path #1 because "H" is later than "G" and proceeds to path #2. This develops as fusion locants "I," "F," "D," "C," and "A" and prints a warning that it has found a superior path. Path #3 is not superior.

**Example 5**




WLN for Path #1: T66 BM DNSWJ

WLN for Path #2: T66 BSWN EMJ

An inexperienced (or careless) encoder might enter WLN #1 because it produced an earlier set of segment symbol locants *in the notation* ("B," "D" *vs.* "B," "E"). The subroutine, however, makes sure that the earliest set *in the ring* is chosen and prints out a warning message that path #2 is preferred ("B," "C," "E" *vs.* "B," "D," "E"). Paths #3 and #4 are obviously not in contention.

The procedure described has been implemented in PL/I on an IBM 360/40 computer at Pomona College. Using the same test files as reported in the previous paper, the program decoded and checked molformulas of carbocyclic and heterocyclic rings at a rate of about 75/min.

## LITERATURE CITED

(1) Granito, C. E., Roberts, S., and Gibson, G. W., "The Conversion of Wiswesser Line Notations to Ring Codes. I. The Conversion of Ring Systems," *J. Chem. Doc.*, **12**, 190–196 (1972).
(2) Smith, E. G., "The Wiswesser Line-Formula Chemical Notation," McGraw-Hill, New York, N. Y., 1968, p 187.
(3) Reference 2, p 118; note that the WLN for No. 6 in this reference is incorrect.
(4) Reference 2, p 122, No. 15.
(5) Reference 2, p 211, No. 26, with double bonds added.
(6) Bowman, C. M., Landee, F. A., Lee, N. W., and Reslock, M. H., "A Chemically Oriented Information Storage and Retrieval System. II. Computer Generation of the Wiswesser Notations of Complex Polycyclic Structures," *J. Chem. Doc.*, **8**, 133–138 (1968).
(7) Reference 2, p 105, No. 9.


# Production of Printed Indexes of Chemical Reactions. II. Analysis of Reactions Involving Ring Formation, Cleavage and Interconversion


ROBERT CLINGING and MICHAEL F. LYNCH*

Postgraduate School of Librarianship and Information Science, University of Sheffield, Western Bank, Sheffield, S10 2TN, England


Received February 25, 1974


**Simple algorithms designed to identify ring changes in records of chemical reactions are described. They operate on the WLN notations of reactant and product molecules sampled from *Current Abstracts of Chemistry* and *Index Chemicus*. They enable summaries of ring changes, including formation, cleavage, and interconversion, to be produced, and account for approximately 22% of reactions in the sample.**

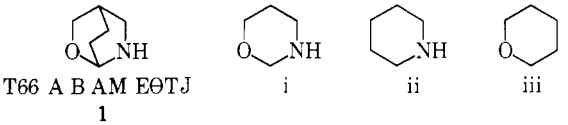

In a previous paper of this series[1] a method was described by which functional group interconversion reactions could be identified by analysis of the Wiswesser Line Notations (WLN's) of the reactant and product molecules of organic reactions. Using a specified list of reactions it was shown that about 20% of a file based on *Current Abstracts of Chemistry* and *Index Chemicus* could be successfully analyzed. Further work has since shown that this can be considerably enhanced if a more comprehensive list of reactions is used. The proportion of reactions analyzed is highly dependent on the source of the data base, although it is largely consistent over three monthly files from *Current Abstracts of Chemistry* and *Index Chemicus*.

In order to extend the usefulness of this system, an approach has now been made to the analysis of reactions involving changes in rings (formation, cleavage, and interconversion).

At this stage, the description of the rings has been limited to the information explicit in the parts of the notations describing the ring systems.

## IDENTIFICATION OF INDIVIDUAL RINGS

WLN's describe complete ring systems, so that it was necessary to develop a routine to analyze these to provide descriptions of the individual rings present. The routine developed here uses the method of Granito, Roberts, and Gibson[2] to identify the atoms within each ring; each is described by its atomic symbol and its degree of saturation or substitution, if this information is available within the ring notation. More specific information identifying fusion points, bridge atoms, etc., is also determined but is not used in the work described here. Only those rings described explicitly by the notation are dealt with; for instance, in compound **1**, only rings i and ii are identified; ring iii is also present but is only described implicitly.




T66 A B AM EΘTJ — i — ii — iii

**1**

The routine detects certain illegal character sequences, and these cause it to set error codes and exit. In addition, there are certain restrictions on the size and complexity of the rings and ring systems involved (*e.g.*, systems containing nonconsecutive locant paths or rings of greater than 20 atoms are excluded), and these are also indicated by exception codes. Each atom in a ring system is described by three characters. The first two characters contain the atomic symbol (for two character symbols the Wiswesser symbol is used if it differs from the standard atomic symbol), and the third defines its degree of saturation. For carbon atoms, this is H for a saturated atom and U if the atom is unsaturated regardless of whether the double bond is endo- or exocyclic. For nitrogen atoms it is not always possible to define the degree of unsaturation unambiguously from the cyclic part of the notation, and, therefore, the Wiswesser symbols M, N, and K are used unaltered. For other atoms the third character contains a zero.

The atoms within the ring are listed starting with the atom whose atomic symbol has the latest alphabetic position and proceeding around the ring so as to give the latest possible position for the whole ring description. If this fails to give an unambiguous result, the characters describing saturation are also taken into account. The result of this is that a particular ring will always lead to the same description, whatever its environment. For example, the description of the ring i is

| Θ O | C H | N M | C H | C H | C H |
|---|---|---|---|---|---|
| atom 1 | atom 2 | atom 3 | atom 4 | atom 5 | atom 6 |

* Author to whom correspondence should be addressed.