Augmenting Connectivity Information by Compound Name Parsing: Automatic Assignment of Stereochemistry and Isotope Labeling

Wolf-Dietrich Ihlenfeldt and Johann Gasteiger*

Computer-Chemie-Centrum, Institut für Organische Chemie, Universität Erlangen-Nürnberg, Nägelsbachstrasse 25, D-91052 Erlangen, Germany

Received January 2, 1995[⊗]

Compound names found in catalogs of fine chemical manufacturers typically contain auxiliary information besides the basic verbose connectivity description. The distribution format of computer-readable catalogs is usually only a plain connection table, but the name often contains additional information. By parsing the name string and analyzing indicator token contexts it is possible to add valuable information to this connection table. Indicator tokens (characteristic name fragments) present in compound names include those describing or denying stereochemistry, isotopic labeling, and the purity of the compound. The algorithms and heuristics developed to automatically assign the location and type of stereocenters and isotope labeling positions in structures of the complexity displayed in typical commercially available compounds are described. The operation principles are based on the intelligent coordination of topological information extracted from the bare unannotated connection table and token statistics generated from the compound name.

1. CHEMICAL NAMES AND CHEMICAL STRUCTURES

The naming of chemical structures has gone through a long process of gradual evolution and is now stuck in an awkward position. Today's IUPAC systematic nomenclature system[1] is far from easy and unambiguous. Furthermore, many trivial names for compounds and structural fragments both blessed by IUPAC and officially deprecated but still in widespread use exist and generate their share of confusion. Determining the structure of a compound from a name or vice versa is often not an easy task under these circumstances, neither for humans nor for computers. Nevertheless, a number of programs, even commercial ones, have appeared in recent years which generate names from connection tables, connection tables from names, and connection tables from scanned graphics. Some well-known examples are AUTONOM[2] (structure to name) and Kekulé[3] (graphic to structure), and some earlier attempts[4−12] include examples for name parsers which output structures as well as predecessors of the above mentioned commercial systems. However, these approaches assume that there exists complete information on one side (compound name or connection table or graphic) and an informational void on the other side(s). The algorithms attempt to transfer as much information as possible from one side to the other, a procedure which is difficult to make lossless. On the other hand, there are certain situations where partial information is available on both sides. By extracting and combining information tidbits from multiple sources (name and connectivity table) and subsequent isolation and merging of nonoverlapping information the overall knowledge about a structure may be enhanced on one or both sides. This situation does not seem to have been dealt with so far.

Such a situation has arisen in our group when catalog data from fine chemical manufacturers was processed for the generation of the starting material libraries of the WODCA synthesis planning system.[13,14] One of the largest and most important catalogs of the system is the complete Janssen Chimica catalog. At the time it was obtained (1988, things have changed now, and Janssen Chimica is now Acros Organica Division), the format was completely nonstandard: A dump of a relational database file which contained names, quantities, prices, and catalog numbers and another such file which contained once more the catalog number as access key and a structure drawing consisting of line coordinates and character symbol locations. The file did not contain a connection table. Converting the structure plot into a connectivity table was not a major problem, although some unexpected minor obstacles had to be overcome.[15] A number of names were only present in the order information file, and some plots had no order information or full name counterpart. Structures without a plot were discarded, while compounds without a name or pricing information were retained. The names of the compounds without an entry in the catalog data file were set to the original cross reference key. Although this string does not contain a compound name, it still may contain some stereochemistry indicator such as "1531384 3257-18-9 L" (the "L" is the indicator token) which can be exploited in the framework of this study.

The total number of accepted compounds was 8464. However, due to the nature of the connectivity table generation, the structural information was very limited. While there were obviously a large number of structures in this data set which were manufactured with a defined stereochemistry, no such information, not even wedged or dashed bonds, was contained in the original plot or the basic connectivity matrix. Likewise the names hinted that there were quite a number of structures with deuterium isotopic labeling, but the initial structures contained only unmarked, mostly automatically added hydrogen. The structure plots either did not show these hydrogen atoms at all (implicit hydrogen atoms) or contained them as plain "H" characters.

While the original work was done within the framework of the WODCA system mentioned above, this paper actually describes the reimplementation of the algorithms with some refinements in the CACTVS system.[16] The used input data are identical, but due to improvements in the algorithms the

[⊗] Abstract published in *Advance ACS Abstracts,* June 1, 1995.

0095-2338/95/1635-0663$09.00/0

Table 1. Available Information: A Catalog Is More Than a Connectivity Description

p-xylene-alpha,alpha,alpha,alpha',alpha',alpha'-D6, 99+ atom% D
alpha,alpha,alpha-tris-(hydroxymethyl)-methylamine P.A. for biological applications
phosphoricacidtrimethylester 97% *import license required*
2,4,7-trinitro-9-fluorenone, 80% moist product, contains 20% water
N-(2,2,2-trifluoroethyl)-hydrazine 70WT% solution in water
2276264 26910-17-8 L L L
o-tolylaldehyde 98% (stabilized with 0.1% H.Q.)
2,2,6,6-tetramethylpiperidin-1-oxyl 98% free radical
2-mercaptoquinoline 97% *** till depletion of stock ***
4-piperidone monohydrate hydrochloride 98% titr. on dry product
phenylacetaldehyde 85%, rest phenetylalcohol
phenol 99+%, loose crystals, biochemical grade
perfluorodecaline 95% cis-trans isomer mixture
1-methoxy-1-butene-3-in, 50 WT% solution in methanol/water (4:1)
magnesiumsulfate, dried, clean, contains 3 to 4 mol water
DL-lactic acid, Eur.Pharm., B.P.
D-(-)-tartaric acid 99% produced from artificial tartaric acid

success rate reported in this paper is somewhat higher than in the original implementation. The reimplementation uses the preconverted data files not the raw database dump.

The original names in the catalog were German. However, in order to honor the international character of this journal, they have been translated into their English counterparts. The results were not influenced by this editing.

2. ENHANCING BASIC CONNECTIVITY INFORMATION

The names in the data base file contained valuable information which was not mirrored in the elementary connection table or ordering information fields. Some representative naming examples are shown in Table 1. However, 8464 compounds is a number where manual structure browsing, editing, and input becomes unfeasible. Therefore, we set out to process this data set with a series of programs in order to automatically extract a comprehensive set of auxiliary information from the compound name and to append the gained information in a readily computer-readable format for further use.

Two such information classes present in the name (these are isotope labeling and stereochemistry) will be dealt with in detail, but abundant other auxiliary data was also present: Purity grade, stabilizers, impurities, moisture content, solvent, stock limitations, application targets, pharmaceutical registry numbers, import and export restrictions, etc. are but a small selection of other data found. The only information from this set outside of the isotope labeling and stereochemistry set which was of interest to the synthesis planning program was purity, which was relatively straightforward to extract (see Appendix).

The presence or absence as well as the locations of isotope labeling and stereochemistry are of primary interest in this paper. The first step to work on these problems is a general cleanup and normalization of the compound names. Details of this somewhat messy step are explained in the Appendix.

2.1. Isotope Labeling. Isotope labeling in the catalog file data set was limited to deuterated species. Basically two types of markers can be distinguished: Those which give a count (such as "benzene-D6", "methanol-D1") and general labeling hints without numerics (such as "deuterium bromide"). These marker tokens are extracted and counted together with their attached location counts by regular expression string match codes (*regexps*). A special problem is posed by compounds with "D" markers without counts (D*n*). If a human reads "ethanol-D" it is immediately clear to him or her that this is an isotope labeling because ethanol contains no stereocenters. On the other hand, "*D*-camphor" will be understood as a stereochemistry specification. In cases such as "2290614 7772-79-4 D" (structure without catalog ordering information) it is necessary even for humans to refer to the structure plot. Trivial names such as "cytochalasin D" or "vitamin D3" and compounds with reduced deuterium content ("water, dedeuterated") complicate the situation additionally. As a rule of thumb isotope labeling positions follow the name, while stereochemistry specifications prefix the name or are embedded, but this is not generally true for the Janssen data set. Recognizing a name fragment as an isotope labeling indicator cannot be completely separated from the stereochemistry assignment problem. Since the complexity of commercial isotope labeled compounds is limited, the rule holds that uncounted "D" markers are stereochemistry markers if the compound can exhibit atomic stereochemistry at all. Refer to paragraph 2.2 for the algorithmic details of the detection of potential stereochemistry. Furthermore, a simple upper limit of the size of molecules considered eligible for isotope marking helps to eliminate the notorious "vitamin D*n*" type of problems without stereochemistry analysis (which would obviously suffice in this case, too). While the method of stereospecific isotope labeling has a rich history in the examination of biochemical pathways and reaction mechanisms, this kind of compound is rarely found in general fine chemical catalogs. Very few labeled compounds with potential stereochemistry (such as decahydronaphthaline-D18) are listed, and these are not sold stereodifferentiated and have attached substitution counts at the marker.

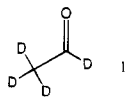
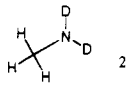
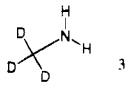
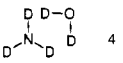
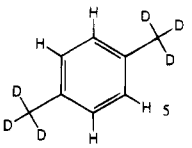
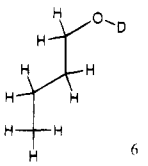
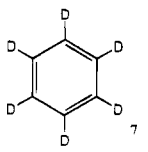
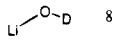
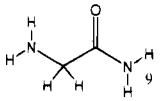
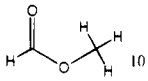
Basically, the name scan for the presence or absence of isotope labeling locates "D" markers, extracts a count if present, and recognizes "deuter" as another marker. The various complications to this simple scheme listed above are also taken in account.

Having established the presence of isotope labeling in a compound, the next step is the assignment of the actual labeling positions. The following rule set reliably works with the restricted compound complexities of starting materials:

In case of present label position counts the following should be done:

- Check if the count matches the total hydrogen count, the heteroatom-bonded hydrogen count (but not the carbon-bonded hydrogen count), or the carbon-bonded hydrogen count (but not the heterobonded hydrogen count, in this order) in the target molecule. In case of fit, assign the labeling positions to this hydrogen subset. These are examples 1−3 in Table 2.
- If besides the count also the "deuter" token has been found, assume that the compound is dissolved in D_2O or similar solvents and substitute all hydrogen atoms, regardless of matching counts. See example 4 in Table 2.
- Group the hydrogen atoms into topological equivalence classes. This study uses for this purpose a fast and reliable hashcoding algorithm.[17] If the count matches the population count of a single hydrogen atom group, assign the labeling positions to this group. This is example 5 in Table 2.

Table 2. Examples of Isotope Labeling Assignments

name	method	plot
acetaldehyde-D4	counted all H	 1
methylamine-D2	counted X-H	 2
methyl-D3-amine	counted C-H	 3
ammonia-D4-deuteriumoxide	counted deuterium override	 4
p-xylole-a,a,a,a',a',a'-D6	counted equivalence group	 5
1-butan(ol-D)	uncounted X-H	 6
benzene-D	uncounted equivalence class	 7
lithiumdeuteriumoxide	uncounted deuterium override	 8
glycine-D5	discarded as coding error	 9
formic acid methylester-D	unassigned	 10

• Otherwise discard as a probable coding error. Refer to example 9 in Table 2.

Without label position counts:

• First check whether there is a single heterobonded hydrogen atom (possibly besides carbon-bonded hydrogen atoms) as an obvious candidate. An example is molecule 6 in Table 2.

• If not found, but all hydrogen atoms belong to the same equivalence class, make all hydrogen atoms labeled. Compare example 7 in Table 2.

• If a "deuter" token is present, assume again full substitution. This is example 8 in Table 2.

• Otherwise leave unassigned.

Both rule sets apply their rules in the order given. Representative examples for all these assignment classes are found in Table 2, and the statistics of the catalog assignment

Table 3. Isotope Labeling Assignment Results

total number of compds	8464
with isotope labeling indicator	90 (1.1%)
counted marker	63
uncounted marker	27
deuterium marker	16
counted assignment of *all* hydrogens	54
counted assignment of all *hetero* hydrogens	2
counted assignment of all *carbon* hydrogens	4
counted assignment by equivalence groups	1
uncounted assignment of single *hetero* hydrogens	8
uncounted assignment of equivalent hydrogens	5
uncounted assignment for *deuterium* markers	14
structure coding errors	1
unassigned	1
successful assignments	88 (98%)

are listed in Table 3. Using this simple set of rules, nearly all the substitution positions of the structures without obvious coding errors could be assigned automatically. The single exception is "formic acid methyl ester-D" (structure 10 in Table 2). This is an example for an uncounted marker with topologically different hydrogen atoms. A human chemist will not be sure about this compound, either. The introduction of another rule to handle this kind of ambiguity is not justified by a single example without a known correct solution.

2.2. Stereochemistry. The final and most complicated problem is the automatic assignment of stereochemistry. The basic approach chosen to solve this problem is similar to the methods applied successfully in the case of isotope labeling detection and location assignment. A primary name scan isolates tokens which hint toward presence or absence of stereochemistry. Afterwards, the molecular structure is analyzed for potential stereocenters, and this information is cross-checked against the information obtained from the name scan.

The number of tokens indicating stereochemistry is significantly larger than the isotope labeling vocabulary. Two completely independent naming schemes (*R/S* vs *D/L* and *E/Z* vs *cis/trans*) plus a variety of implicit stereochemistry in trivial names (*D*-glucose), ring-substituent oriented cis/trans nomenclature (*trans*-1,2 dichlorocyclohexane), 3D-related keywords (*endo/exo*), compound class specific naming conventions (α and β sugars), and various other difficulties obfuscate the picture. Additionally, name constructions such as "*cis/trans* mixture" must not be mistaken as indicators for stereochemistry but rather the opposite. Keywords like "mostly *trans*" lead to the question how much enantiomeric or diastereomeric excess is needed in order to consider a structure as stereochemically defined. Table 4 gives an overview of the encountered naming schemes for stereo compounds from the Janssen Chimica catalog.

The data structure in the WODCA system presently does not allow for stereochemical purity specifications. In the presence of an indicator showing that a compound contains some excess of an enantiomer or diastereomer it is coded as stereodefined the same way as truly pure compounds (if such compounds exist at all). The explicit denial of potential stereochemistry by mixture indicators is not specifically stored. In this environment it is currently not possible to express knowledge about absence of stereochemistry as opposed to lack of knowledge. These compounds are therefore passed as unassigned. The routines which perform the primary assignment of stereochemistry just attach the found and checked descriptor in its original form to the bond

Table 4. Many Ways To Express Presence or Absence of Stereochemistry

L-abrine 99%
S-trityl-L-cysteine
2277072 926-79-4 L L L L
O-acetyl-DL-mandelic acid chloride
N-acetyl-L-phenylalanyl-3,5-diiodo-L-tyrosine
cis-2-butene-1,4-diol
1,4-dibromo-2-butene 99% mostly trans
1814910 3306-06-7 1S 2S +
2-benzoyl-5-norbornene techn. 90% endo and exo mixture
endo-3-bromo-D-camphor
2285358 S-L
cis,cis,cis,cis-cyclopentane-1,2,3,4-tetracarbonic acid
L-threo-1,4-dimercapto-2,3-butandiol
(-)-isopulegol
(S)-(-)-1-phenylethylamine
quinine
quinidine
5-thio-D-glucose mostly alpha-anomer

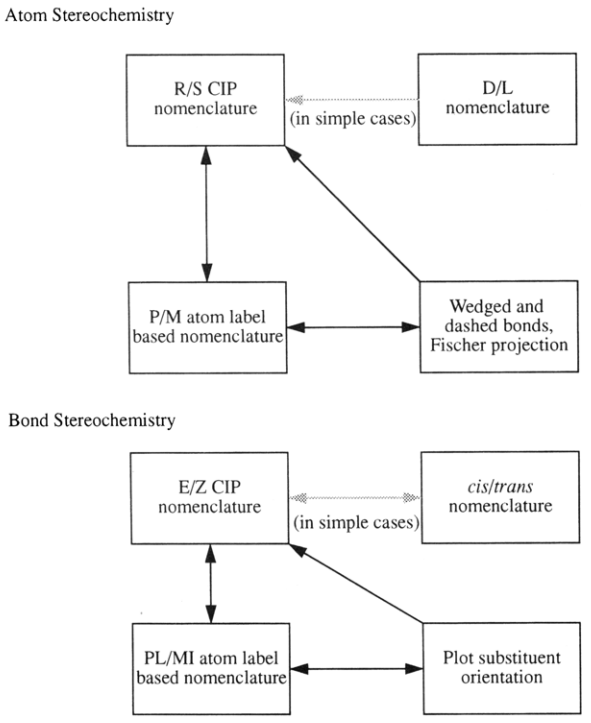


Figure 1. Stereochemical descriptor conversion routes in the WODCA system.

or atom. WODCA contains subroutines which convert some form of stereochemical descriptors into other forms on demand. The supported conversion routes are listed in Figure 1.

2.2.1. Determination of Potential Stereochemistry. A few computer implementations of algorithms to detect and handle CIP stereochemistry have appeared, most notably those included in CHIRON[18] and LHASA[19] as well as a stand-alone program.[20] Abundant literature on the internal representation of stereochemistry on computers and operations on those data structures is available.[21-25]

The first step toward assignment of stereochemical descriptors is the determination of potential stereocenters and stereobonds in the compounds under examination. This has to be achieved solely with the information in the connection table. The structure plots in the data file obtained from the supplier did not contain any wedge bonds or similar graphical hints. Some structures were drawn in a perspective projection, but errors were not uncommon. For example, *D* and *L* camphor fragments were misused in several places. Even *trans* double bonds were plotted in at least one instance as *cis* in order to shape the molecule to fit into the roughly square plotting area. Because of these observations, no attempt was made to include information from the plot coordinates in the detection process.

The stereochemistry detection subroutine is actually a part of a CIP descriptor computation package. This module follows basically the definitions and algorithms outlined in refs 26 and 27. Although completely independent in implementation, the basic operation principles are very similar to those reported in refs 18 and 19. This is hardly avoidable because the CIP definition is itself a very algorithmic paper.

The input information needed for the determination of potential stereochemistry include charge, free electron pairs, rings, aromaticity, and isotope labeling in addition to the elementary connectivity information. Routines to compute this information were available from the EROS system libraries[28] or, in the CACTVS reimplementation, are part of the computational core capabilities and are not described here.

Stereochemistry treated in this paper is limited to tetrahedral atoms (including those with spatially localized free electron pairs), double bonds, allenes, and *cis/trans* diastereoisomerism based on ring systems. Stereochemistry originating from higher coordination numbers is not handled and was not present in the data set. Examples of allenic stereochemistry, although principally detected, were not present either. Pseudoasymmetric stereocenters and bonds are treated, and such atoms, but no such bonds, do occur in the data set.

The first step is a basic check whether an atom or a bond can exhibit stereochemistry at all, judged from its element type and first-sphere neighbors. Text book rules are applied to determine whether the atom type (C, N, Si, P, S, Ge, As, Se, Sb), charge, valencies, and number of neighbors are acceptable in principle. No further examination of the neighbor sphere is carried through in this first filter step, with a few exceptions: The very common case of two plain, i.e., no deuterium or tritium labeling, hydrogen atoms bonded to the same atom disqualifies the candidate atom. Furthermore, for certain stereochemistry types stemming from localized free electron pairs, a minimum number of neighbors must be strongly electronegative, i.e., should be oxygen or nitrogen. Atoms in aromatic rings are always excluded. A similar set of rules is applied to filter the bond set. Only double bonds are allowed which are neither aromatic nor in small or medium sized rings (where *cis* is implied) and have suitable first sphere neighbors in certain cases. Oximes but not imines are accepted as potential stereobonds where a localized electron pair acts as a substituent. Bonds are immediately rejected if two unlabeled hydrogen atoms are ligands at the same bond atom. A complete set of rules can be found in any stereochemistry textbook and is easily implemented. Generally, the number of atoms and bonds within a molecule surviving this first filtering step is less than 10%. Surviving atoms and bonds are promoted to a "maybe" state, while the rest remain marked with a plain "no" for the potential stereochemistry characteristic.

The following detailed analysis is carried through only on this subset. Repeated attempts are made to evaluate the possibility of stereochemistry for all "maybe" atoms, all "maybe" bonds, and all those atoms again and again as long as in a bond/atom cycle at least one successful status determination (either "checked-no" or "checked-yes") has taken place. Assignments cannot generally be completed in a single cycle because their status may depend on the yet undetermined status of atoms or bonds encountered in a neighbor sphere. In certain complicated cases atoms with unclear stereochemical status are still present when the loop terminates. These atoms are handled in a final cleanup phase with heuristic rules.

In order to evaluate the possibility of stereochemistry, neighbor spheres are expanded around the central atom (in the case of tetrahedral stereochemistry) or the terminal atoms (for double bond and allene type stereochemistry). Each atom entered into a neighbor sphere is assigned a 32 bit unsigned signature ID, which encodes, in the order of CIP significance, PSE number, isotope, character of the bond this atom was reached over (aromatic or not), phantom status, and stereochemistry. Phantom atoms are defined in the CIP rules: the partner atom of an atom in a multiple bond is entered a number of times corresponding to the bond order, but only the first, i.e., nonphantom entry is expanded in further spheres. Free electron pairs are not explicitly coded. Spheres are expanded stepwise by one bond.

After each extension a check is made whether a status assignment is possible or not. If not, but new atoms have been entered in the last expansion cycle and no atoms which are themselves yet undecided were encountered, another sphere is added until a conclusion is reached; currently undecided atoms were encountered or the reachable atoms are exhausted. The expanded atom sphere tree is always kept in sorted order. After adding the atoms around a former leaf atom node, this set is sorted according to the signature values. On the way back from the terminal sphere through the recursive call chain, the inner positions are also sorted by systematic comparison of the subtrees leading to the leaf nodes just returned from. Subtrees are compared by a function which returns one of four possible return values: identity, CIP preference for one of the two subtrees, or indecision. Indecision is returned if no difference in the signature ID's with the flags for potential stereochemistry masked out can be found, but in at least one of the signatures of the involved atoms an uncleared possibility flat is present. The comparison function is itself recursive. If no difference and no reason for early abortion is found, the function calls itself recursively with the children of the current subtree node, working itself down on the tree to the leaves. Since the subtrees are always kept sorted, this scheme is easy to implement. Free electron pairs are not coded explicitly but are always of the lowest priority, so a precedence return value is generated when the children of higher priority yield no decision but the number of children is different. Atoms may temporarily be entered multiply in a subtree to form a single-atom overlap of different paths at maximum, but a flag is bubbled upwards indicating whether any nonduplicate atoms were entered during an expansion cycle. Bonds, however, are never traversed back and forth within two directly following levels. In any case, the tree growth is terminated when no further fresh atoms can be added in any subtree. Another early termination criterion is if any two of the three (in case of a free electron pair) or four subtrees around a given center (atom or bond/allene terminals), which could not be extended with fresh (nonoverlapping) atoms during a cycle are found decidably equal. This outline of the tree growth and comparison processes may seem complicated, but the method is actually rather easy to implement in a language which allows recursion and dynamic memory allocation. The double recursive scheme of tree growth and tree sort with the comparison function which is itself recursive for the sort of the expanded tree on the way back to the upper call levels has a certain elegance when expressed concisely in about 50 lines of code.

In some cases atoms remain with a status that cannot be resolved with this scheme. Two main cases exhibit this behavior: Ring atoms which express a ring-oriented *cis/trans* relationship to similar atoms in a position on the opposite side in the ring (i.e., 1,3-relationships in four-membered rings, 1,4-relationships in six-membered rings, or more complicated structures spanning multiple connected rings) and bridgehead atoms in cage-ring systems. This type of stereogenic unit can only be handled with the 1982 CIP revision,[27] not with the older definition.[26] Before the CIP revision, *cis/trans* descriptors were abused to express these facts, and this scheme is also still standard in catalog names. It is therefore useful to separate this case for this application and to shortcut the normal strict CIP rule processing (rule 4 in the 1982 revision). In order to resolve these cases, the following heuristic rules are applied in the order listed. Later rules are only applied if no fitting rule earlier in the list was found:

- Unresolved atoms not in a ring are potential stereocenters.
- Unresolved bridgehead atoms (member of more than two rings in the ESSSR[29]) are no stereocenters.
- Unresolved atoms in a ring are stereocenters with a special "ring *cis/trans* diastereoisomerism" attribute, if there are other unresolved atoms in the ring, otherwise they are not.

This set of rules works perfectly with the catalog data set but has some implicit assumptions and is therefore not applicable in the general case. For example, the second rule assumes that for molecules of the size and complexity found in the catalog no inside/outside isomers exist. This is certainly true for this data set but cannot be taken for granted in arbitrary data collections.

Some examples may be helpful to make the methodology clearer. The example molecules are shown in Figure 2.

The classical example of glycerole aldehyde (structure **11** in Figure 2) is one of the simplest cases. The early filtering step leaves only the central carbon atom as a candidate. The first sphere expansion around this atom does not yet lead to a decision, but no other unassigned stereocenters, equivalent unexpanded subtrees, or exhaustion of fresh atoms stops the analysis. The second step brings the decision: Two oxygen atoms from the aldehyde group (one real atom and one phantom atom, generated because of the bond order of the carbon-oxygen double bond) versus a single such atom from the opposite hydroxy group are a difference, and so the central carbon atom is recognized as a potential stereocenter.

In the case of citral (structure **12** in Figure 2), the initial screening leaves two double bonds but no atomic or allenic centers as candidates. The impossibility of stereochemistry on the dimethyl-substituted double bond is detected in the third cycle. In this cycle the expansion around the hydrogen atoms of the methyl groups fails. At the subtree comparison

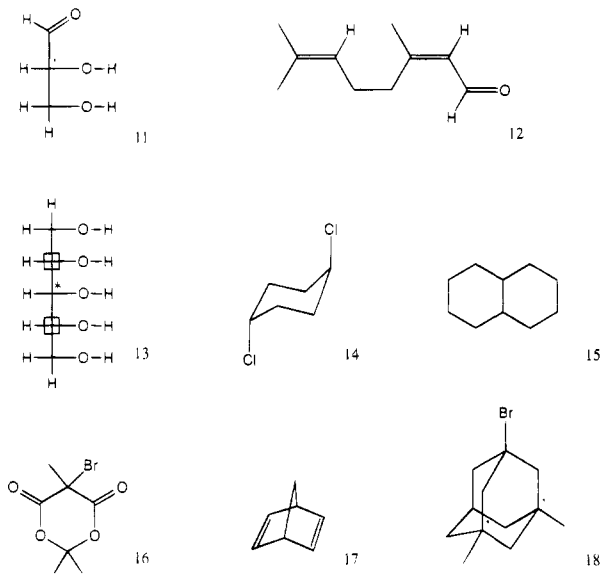


Figure 2. Example molecules explaining the algorithms used to detect potential stereocenters or stereobonds (see text).

step the identity of the two subtrees leading to these groups is detected, and since neither of them was expanded in this cycle, their final equality and therefore the impossibility of stereochemistry at this bond is detected.

Assuming an unfortunate internal numbering of the atoms (the central atom, marked with an asterisk, is assumed to be the atom numbered 1), compounds such as arabit (compound **13** in Figure 2) need more than one cycle for the assignment. All three inner carbon atoms survive the filtering step. However, the analysis of the central atom is aborted as soon as the expansion reaches the yet undecided boxed neighbor atoms. Only after these have been found to be able to carry stereochemistry, is the middle atom decided upon in a second atom/bond assignment cycle.

No decision by path expansion is possible for compounds with structures similar to 1,4-dichlorocyclohexane (**14** in Figure 2). Starting from either carbon atom with the chlorine substituent, the expansion finally hits the undecided stereocenter on the opposite side of the ring. According to the list of rules to be applied in such cases, both atoms are finally assumed to serve as anchors of ring *cis/trans* diastereoisomerism but are not enantiomers. The detection of *cis/trans* bond diastereoisomerism in compounds such as decalin (**15** in Figure 2) follows the same paths. Keep in mind that this is not processing strictly according to the CIP rules.

In the case of 2-bromo-2-methylmalonic acid acetonide (**16** in Figure 2) a postponed decision can be resolved in the second cycle. If the bromo-substituted atom comes first, the sphere expansion will hit the undecided dimethyl-substituted carbon atom at the other side of the ring. After delaying the decision on this subject, the nonstereospecific character of the acetone fragment center atom is established. In a second try on the bromo-substituted atom, all atoms are exhausted in the expansion process. At the moment of exhaustion, both ring branches still compare equal, so there is no stereochemistry in this case.

A further example with a bridgehead atom is 2,5-norbornadiene (**17** in Figure 2). Like in the case of 1,4-dichlorocyclohexane, no decision is possible, but due to the bridgehead nature of the atoms in question no stereochemistry is assumed to be present. Of course there are bridgehead atoms in other compounds which are stereocenters such as in some adamantane derivatives. These are correctly detected in the sphere expansion process. An example of such a compound is 1-bromo-3,5-dimethyladamantane (**18** in Figure 2).

Table 5. Potential Bond Stereochemistry, Full Data Set

no. of potential bonds	structure count	no. of potential bonds	structure count
1	268	4	4
2	31	more	none
3	3		

Table 6. Potential Bond Stereochemistry, Indicator Token Data Set

no. of potential bonds	structure cound	no. of potential bonds	structure count
1	71	3	1
2	10	more	none

Table 7. Potential Atom Stereocenters, Full Data Set

no. of potential centers	structure count	no. of potential centers	structure count
1	805	9	9
2	272	10	8
3	92	11	3
4	128	12	0
5	66	13	0
6	28	14	2
7	19	15	1
8	22	more	none

Table 8. Potential Atom Stereocenters, Indicator Token Data Set

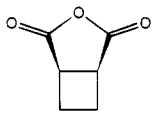
no. of potential centers	structure count	no. of potential centers	structure count
1	213	9	2
2	126	10	3
3	35	11	1
4	38	12	0
5	42	13	0
6	3	14	1
7	0	more	none
8	0		

The analysis of the frequency distribution of potential stereocenters or stereobonds in the catalog data set is instructive. Tables 5−8 show the distributions of potential bond and atom stereochemistry for the full catalog and those entries where the name contains an identified stereochemistry indicator token. The maximum number of potential stereobonds found in the data set is four (Tables 5 and 6), but the majority of the entries has only one potential stereobond. This weighted distribution makes attempts for the automatic assignment of bond stereochemistry more promising. The situation is similar for atoms (Tables 7 and 8). The astonishing maximum number of potential atom stereocenters found is 15 for "maltotriose 93%" without an indicator and 14 with "D-(+)-raffinose pentahydrate". However, in this case also there is a pronounced and reassuring bulk of structures with lower counts. Not surprisingly, compounds with only one or two potential atomic stereocenters are the majority. Simple sugar derivatives introduce a flat spectrum of three to five potential stereocenters because C5 open chain sugar alcohols possess three centers and C6 pyranoses five centers. Many of the compounds with two stereocenters are isoleucine derivatives of dipeptides.

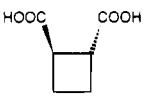
2.2.2. Stereochemistry Token Scan. The cleaned names are submitted to a token scan procedure similar to those performed to isolate isotope labeling indicator tokens (para-

Table 9. Many Ways To Express Mixtures of Different Kinds

2-acetyl-5-norbornene, endo-exo isomer mixture
N-acetyl-DL-homocysteine-thiolacton
0850950 68697-61-0 DL HCl
(-)-limonene oxide, cis-trans isomer mixture
lanthionone 98%, DL-meso mixture
P-dimethylamino-cinnamonicacidnitrile, cis-- trans
DL-alanyl-DL-norvaline
2-amino-5-chlorobenzophenonoxime syn/anti mixture
trans-cyclohexane-1,2-dicarboxylic acid (see text, this is an *R/S* mixture)



Label: *cis*
Actually: *meso*-type *R,S* diastereomer, random choice of *R* and *S* attachment points for the stereocenters is possible



Label: *trans*
Actually: *R,R* and *S,S* enantiomer mixture, analogous to *DL* tartaric acid derivatives

Figure 3. Hidden *meso* or *DL* mixture character in *cis/trans* ring compounds.

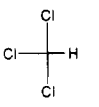
graph 2.1). Regular expressions are again the tool used to isolate the relevant tokens. Tokens scanned for include the standard nomenclature building blocks of the CIP and *D/L* systems (*R, S, D, L, E, Z, cis, trans*) as well as some auxiliary commonly used indicators (*syn, anti, endo, exo, threo, erythro, meso*). Furthermore, an attempt is made to identify optical rotation tokens ((+) and (−)). Special care is taken to detect and discard mixtures. Besides scanning for explicit words such as "mixture", sequences of two consecutive (i.e., separated only by white space, dash or slash) stereo tokens of opposite sign in configurations such as "*endo-exo*" are assumed to describe mixtures. In case of mixtures, no attempt is made for an assignment. Otherwise, the basic assignment phase described in 2.2.3 commences after the counts of every indicator token have been established. A sample of names correctly identified as mixtures is found in Table 9.

There exists a certain class of compounds where this joined matching-pair approach does not detect the basic enantiomeric mixture character of a catalog compound. The prototype of this kind of compound is cyclobutane-1,2-dicarboxylic acid. The catalog contains the acid labeled *trans* and the corresponding *cis* anhydride. The problem is shown in Figure 3. Actually, this compound pair corresponds to *meso* and *DL* tartaric acid derivatives. The tartaric acid derivatives, however, have the advantage of being traditionally labeled in a way which readily identifies the latter as an enantiomer mixture and the former as a compound where it is allowed to randomly select one of the stereocenters as *R* and the other as *S*. In the case of such ring compounds, the token scan result is modified if a *cis* or *trans* marker was detected without stereogenic double bonds being present, and instead two equivalent potential atomic stereocenters exist and at the same time no atom stereochemistry tokens are found. In this case the recognized *cis/trans* marker is deleted from the internal list, and a synthetic *D/L* or *meso* token is introduced.

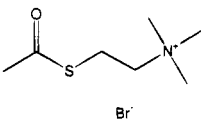
The stereo indicator token distribution in the Janssen Chimica catalog is shown in Table 10. This table counts every indicator type only once per molecule, so in case of "*L*-alanyl-*L*-alanine" only a single entry is added to the *L* token count. In contrast, a count of two *L* tokens is passed

Table 10. Indicator Token Distribution

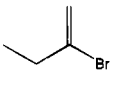
token	structure count	token	structure count
R	15	*E*	1
S	37	*Z*	0
D	133	*cis*	32
L	194	*trans*	80
erythro	1	*meso*	6
threo	4	opt. rotation (−)/(+)	148
endo	5	mixtures	77
exo	4		



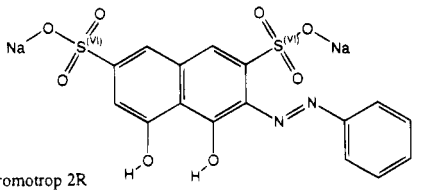
chloroform-D 98 atom% D



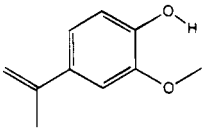
S-acetylthiocholine-bromide



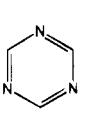
2-bromo-2-butene, cis-trans mixture



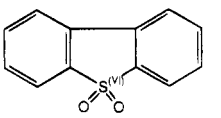
chromotrop 2R



isoeugenol, cis/trans mixture



S-triazine 98%



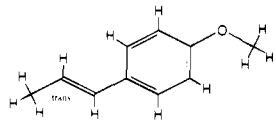
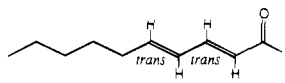
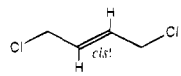
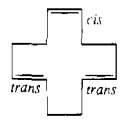
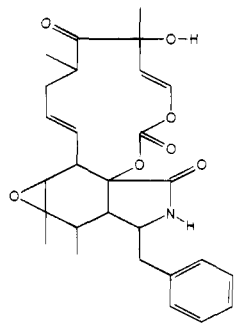
dibenzothiophene-S,S-dioxide

Figure 4. Example compounds whose names have spurious stereochemistry indicator tokens.

to the assignment module. This graph shows some interesting but not completely unexpected imbalances. First, it becomes very obvious that *D/L* nomenclature still dominates the naming of commercial products, despite its shortcomings and tendency to generate confusion.[30] The very few cases with CIP nomenclature are names of compounds not isolated from natural sources or produced by fermentation such as "(*R*)-(+)-1-phenylethylamine". Only one *E* descriptor was found, which turned out to be spurious, "cytochalasine E", and not a single *Z* counterpart. While there are 1771 compounds in the database which can exhibit stereochemistry, only 103 were marked as mixtures, and 504 nonmixtures had some kind of stereochemistry token. This leaves more than 1000 compounds (about 14%) with potential stereochemistry the catalog is ignorant of.

In a small (55) number of cases tokens were found, but no corresponding potential stereocenters, stereobonds, or stereoplanes could be identified. The most common reasons for failures of this type are sulfur atoms whose substituent is marked with an *S* such as in "S-acetylthiocholinbromide". Other reasons are deuterium markers misinterpreted as *D* stereo markers and fragments of dye names such as in "chromotrope 2R". Coding errors are also present, as represented by "2-bromo-2-butene *cis/trans* mixture" which is coded as 2-bromo-1-butene. The latter compound does not form *cis/trans* isomers. In the case of *cis/trans*-isoeugenol the *cis/trans* token presumably refers to a nonstandard naming convention concerning the relative position of the methoxy and hydroxy substituents to the isopropenyl group at the opposite side of the phenyl ring. If no possibility for stereochemistry could be found, these

Table 11. Examples of Bond Stereochemistry Assignments

name	method	plot
trans-4-amino-cinnamonic acid hydrochloride 97% mostly trans (Purity was stripped in name cleaning process, so only a single 'trans' remains)	full assignment	
trans, trans-2,4-decadienal technical grade	full assignment	
cis-1,4-dichloro-2-butene 95+% (Ignore the plot coordinates!)	full assignment	
trans, trans, cis-cyclododeca-1,5,9-triene 98%+	equivalent assignment	
cytochalasine E ('E' was misinterpreted as bond indicator - but there is more than one candidate in the big ring)	failed	

tokens are considered spurious, and no attempt for assignment is made. Figure 4 shows some examples.

2.2.3. Basic Assignment of Stereochemistry. This phase is entered after the analysis of potential stereochemistry and the token scan. Assignment for bond and atom stereochemistry proceeds independently. The current data structure of the WODCA system does not allow use of ring plane stereochemistry, so it is currently discarded immediately after detection.

The algorithms employed for stereobonds comprise only two rules:

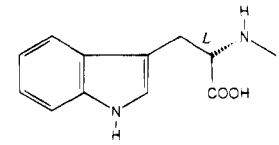
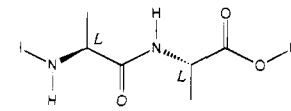
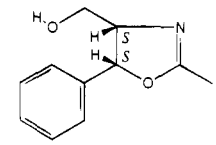
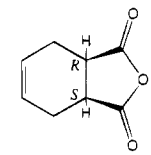
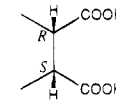
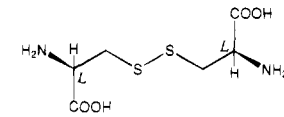
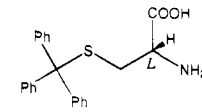
- If only a single type of bond stereo descriptor is found and its count matches the number of potential stereobonds, simple assignment is possible.
- If all potential stereobonds are topologically equivalent and the combined number of *cis/trans* or *E/Z* tokens is the same as the number of the potential stereobonds and no token from the other convention is present, assign the found set of descriptors arbitrarily.

With this simple rule set, 67 of the 69 compounds with potential bond stereochemistry and a stereo token could be assigned. This corresponds to a success rate of 97%. Examples for the bond assignment are found in Table 11.

The task of atom stereochemistry assignment turns out to be more demanding. The following four basic rules stand at the beginning of the process:

- If only a single type of atom stereo descriptor is found and its count matches the number of potential stereocenters, simple assignment is possible.

Table 12. Examples of Successful Simple Atom Stereochemistry Assignments

name	method	plot
L-abrine 99%	full assignment	
L-alanyl-*L*-alanine	full assignment	
(4*S*,5*S*)-(-)-4-(hydroxymethyl)-2-methyl-5-phenyl-2-oxazoline 99%	full assignment	
cis-4-cyclohexene-1,2-dicarboxylic acid anhydride	pseudo *R/S* via ring bond (this is some kind of meso compound)	
1735690 608-40-2 *meso*	*meso* assignment	
L-cystine 99%	equivalent atom assignment	
S-trityl-*L*-cystein 97%	sulfur subtraction and subsequent full assignment	

- If a single atom stereo token is found and all stereogenic atoms are topologically equivalent, assign the same descriptor to all atoms.
- If the compound contains a sulfur atom, a single *S* token was found and a single other type of token from the *R*, *D*, and *L* set is present, delete the *S* token and reapply the above rules from the beginning. This rule assumes that the *S* token is a substituent point descriptor, linking a group to the sulfur atom.
- If the *meso* marker was found and there is a pair of topologically equivalent stereogenic atoms, assign an *R* descriptor to one of the atoms and an *S* descriptor to the other.

These four rules work surprisingly well. From the 387 compounds which may exhibit atomic stereochemistry and possess a marker, 231 (59.7%) can be assigned without problems. Representative examples are shown in Table 12. Close to 60% successful assignments is a good start, but the

regularities in the unresolved compounds permit another step which yields again a significant improvement.

2.2.4. Fragment Libraries. The success rate can once more be improved by introducing a small fragment library. A two-step matching process is involved, and the library fragments are scanned twice.

In the first cycle, a naming substring associated to each fragment library molecule is searched in the catalog name. If this token is found and a substructure match between the library fragment and the catalog compound is successful, the search ends.

If no hits were obtained in the first cycle with any associated name fragment of the library, a substructure match without prior name check is performed in the second cycle. In the case of a substructure match, some information can be injected into the stereo specification process even if the fragment will probably not match with its full stereochemical feature set. This will be explained below.

An analysis of the molecules which could not be assigned with the simple methods presented before reveals a common problem. The number of stereocenters in these compounds is much higher than the number of indicator tokens. This is mostly a result of implicit stereochemical conventions expressed in trivial names, such as "*L*-glucose". Glucose has four stereocenters in the open chain form and the pyranose from adds another (the α and β anomers), but only the orientation of the stereocenter with the lowest label number is specified with the *L* descriptor. The rest of the configuration is implicitly contained in the trivial name and follows the configuration of the first center. As for the second major reason for failure, the configuration of a number of stereocenters is directly coupled to each other because they are part of a rigid ring structure. The most important example for this case is camphor with three stereocenters. This molecule has only two stereoisomers because it is impossible to invert the bridgehead carbon atoms independently.

The fragment library contains substructures and attached fragments of trivial names, for example, a glucose fragment with its full stereochemical specification and its associated "gluc[ou]" substring. The term "gluco" is used instead of "glucose" in order to capture slight naming variations such as "glucosyl". Some fragments have more than one attached name substring. If the name of a checked compound contains such a substring and a substructure match between the fragment from the library and the checked structure is successful, a stereo assignment is possible. If only a single stereo descriptor token has been found, meaning that the compound contains only a single fragment associated to a trivial name substring, the substructure match overlap area can be resolved fully from the coded stereochemistry of the fragment library molecule even if it contains more than a single stereogenic atom. So it is possible to assign the four stereocenters of glucose with their correct CIP configurations. They are specified in the fragment library for one enantiomer (*D* or *L*) and are all inverted if the stereo token of the library fragment is not identical to the stereo token of the query structure, which again is surmised from the indicator token lists generated in the name scanning phase. If the associated name substring was not found in the catalog compound name but a substructure match in the second cycle is successful, some limited partial assignment may still be possible if a *D*/*L* marker has been found. Since the primary atom of the *D*/*L* nomenclature is in the same position in a family of

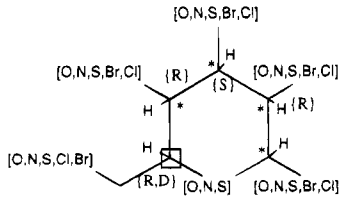
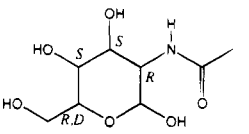
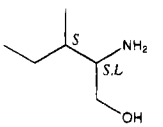
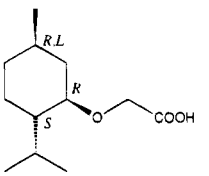


Figure 5. Example substructure fragments used in the library-based stereochemistry assignment process.

Table 13. Examples of Stereochemistry Assignments with the Aid of a Fragment Library

name	method	plot
N-acetyl-*D*-galactosamine	galactose fragment	
L(+)-2-amino-3-methylpentanole	isoleucine fragment	
L-(-)-menthoxyacetic acid	menthol fragment	

structurally related compounds (sugars, amino acids), this single point, but not the other stereocenters in the match overlap area, can be assigned with reasonable confidence. The practical value of this partial assignment may be dubious, but this is still an information gain. In the case of multiple tokens, care must be taken to assert that the match areas of candidate substructures are unique, mutually exclusive, and nonoverlapping. Only in this case, multiple assignment is without risk. However, the catalog data set did not contain such compounds.

Fortunately, the number of commonly used fragments is rather small. With only 10 fragments the rate of successful assignment for atom stereochemistry could be boosted from 59.7% to 72.9% (282/387) without loose (no name correspondence) matches. If loose matches without the name fragment correspondence are allowed, the result is even better with 297/387 (76.7%). The ten fragments comprised the following: glucose, mannose, galactose, arabinose, lyxose and fucose (pyranose form), ribose (furanose), isoleucine, and the camphor and menthol skeletons. These fragments were chosen for the study by visual inspection and counting of the unassigned compounds. The camphor fragment is an example of a match specification associated with two name fragments: "camph" and "born" for the camphor and borneol derivative families. The fragments are generally formulated with atom search lists in strategic positions instead of fixed heteroatoms, so thiosugars, sugar alcohols and acids, brominated derivatives, etc. will still match in the substructure. Additionally, sometimes open bond orders, for example, in the camphor skeleton in order to allow matching with both the borneol and camphor series, are used. An example of such a substructure is shown in more detail in Figure 5. It shows the glucose pyranose search fragment. The boxed

Table 14. Stereochemistry Assignment Results

total number of compds	8464
compds with potential simple atom stereochemistry	1455 (17.2%)
compds with potential simple bond stereochemistry	306 (3.6%)
compds with potential allenic/ring plane stereochemistry	34
compds with both potential simple atom and bond stereochemistry	22
compds with both potential simple atom and allenic/ring plane stereochemistry	2
compds with both potential simple bond and allenic/ring plane stereochemistry	0
total number of compds with any potential stereochemistry	1771 (20.9%)
compd names with identified stereo tokens	607 (7.2%)
compd names with mixture tokens	103
mixtures surmised because of *cis/trans* ring nomenclature	26
compd names with stereo tokens but not a mixture	504 (6.0%)
compd names with optical rotation, but no other descriptor	28
compd names with spurious indicators	55
compd names with indicators and potential simple atom or ring plane stereochemistry	464
compd name with indicators and potential simple bond stereochemistry	82
compd name with indicators, potential simple atom or ring plane stereochemistry and not a mixture	387
compd name with indicators, potential simple bond stereochemistry and not a mixture	69
successfully assigned simple atom stereochemistry/no fragments	231/387 (59.7%)
successfully assigned simple atom stereochemistry/with 10 fragments	297/387 (76.7%)
successfully assigned simple bond stereochemistry	67/69 (97%)
compds with simultaneous successful bond and atom stereochemistry assignment	0
total success rate (of principally resolvable compds)	364/456 (79.8%)

atom is the anchor for the *D/L* nomenclature. The starred atoms are additional locations of potential stereocenters. Matching atoms are assigned the *R* and *S* CIP descriptors attached in curly braces to the corresponding atom in the fragment. They are inverted if this fragment is matched on *L* glucose derivatives. Note that potential stereochemistry at atom 6 is covered with an asterisk, but by default no assignment for α or β anomers is carried out. The name of this fragment is "D D-Glucose GLUC[OU]", which is parsed as defining the *D* enantiomer for compounds with "GLUCO" or "GLUCU" in their names.

A few assignments made possible by the substructure library are listed in Table 13.

For these ten fragments, the average hit number is higher than five. The next eight fragments (xylose, rhamnose, fructose, proline, limonene, ephedrine, pinene/myrthene, and fenchone) which occur multiple times in the last 80 unassigned molecules are good for about 25 additional matches. This is an average of three hits per fragment. After that the possibilities for fragment clustering rapidly diminish because of the lack of common substructures. Table 14 gives a final summary about the catalogs statistics and the performance of the stereochemistry assignment algorithms described in this paper.

3. CONCLUSIONS

It has been shown that the information content of a plain connectivity matrix can be significantly enhanced by a combination of name string parsing, token context analysis, and connectivity relationship examination. Names used in this study are the catalog names of a major fine chemicals supplier which contain a variety of additional information besides the plain structure description. In the case of compounds of the complexity typically found in such catalogs, isotopic labeling positions can be assigned with a success rate approaching 100%. In the more complicated case of stereochemistry, a large subset of the compounds with defined stereochemistry can be resolved correctly by simple principles once the location of the potential stereocenters and stereogenic bonds has been determined. At this level, the success rate for the correct identification of the stereocenters and stereobonds is about 60% for atoms and 95% for bonds. By the introduction of a small library of commonly occurring fragments and with the application of substructure matching techniques the performance for atom assignment can be increased up to 75 to 80% correct resolution. Valuable information is thus extracted from unedited compound names which contain a lot of noise not related to the pure structure description. The methods described in this paper proceed without actually contemplating on the complicated structural implications of the ambiguous chemical naming conventions but focus on the extraction of clearly defined tokens which can be isolated from name strings with a high noise level.

ACKNOWLEDGMENT

Wolf-D. Ihlenfeldt gratefully acknowledges financial support from Studienstiftung des deutschen Volkes (scholarship) and Deutsche Forschungsgemeinschaft. We thank Janssen Chimica, Belgium (now Acros Organica Division), for providing us with their catalog data.

APPENDIX

Purity in Compound Names. Purity was mostly plainly coded as "x%", but other indicators also had to be taken into account. Furthermore, the inherent meaning of "purity" is not constant in the data set: An isotope labeled compound with "99.5 atom% deuterium" implies another concept of purity than "*trans*-compound 99%" or "*trans*-compound 70% *trans*". In the case of solutions, suspensions, or stabilizer quantity specifications, care must be taken to isolate true compound-specific information from preparation-related data. To add another complication, some keywords give a vague idea about the quality, but without precise numbers, such as "hochrein" (German for high purity grade), "techn." or "Eur. Pharm.". Some representative names found in the catalog which display purity information in one form or another are listed in Table 15.

Cleanup of Compound Names. In the cleanup step, the names are modified to conform to a standard pattern with upper case, normalized spacing, corrected spelling, and suppressed hyphenation. Although the names consisted of a single string, many of the entries were hyphenated between

Table 15. Plenty of Ways To Express Compd Purity and Concentration

L(+)-ascorbic acid Eur. Pharm.
L-(+)-arginine 98+%
benzene-D6 99,5 atom% D
benzene (free of thiophene)
9-BBN, 0,5M solution in hexane
2-benzoyl-5-norbornene techn. mixture of endo and exo
benzylchloride, 99.5+% stabilized with 0.25% propylenoxide
2.6-dichlorindophenol sodium salt hydrate P.A.
borontrifluoride-propanol-complex 14% BF3 W/V in propanol
4-bromo-o-xylene tech. 75-80% (contains 20-25% 3-bromo-o-xylene)
1,3-cyclopentandiol 98%, cis-trans isomer mixture (>75% cis)
cyclodecene tech. mostly trans
5-norbornene-2-aldehyde 95+%, mixture exo/endo 4/1
1-butene-3-on (3-2), ca 95%(stab with ca. 0.5% CH3CN, 0.5% CH3COOH, 0.5% H.Q., 3%[31]

column 38 and 42 such as in 2-ethyl-3-methyl-2-cyclohexene-1-on-4-*car- bon*saeuremethylester.

The cleanup step was performed with a collection of some 80 regular expression substitution specifications (*regsubs*). This number counts every instance of a keyword as a separate expression, but the true number of compiled *regexps* is lower because they were implemented with alternative branches in summary expressions. Auxiliary information not related to the compound name itself, to isotope labeling, or to stereochemistry is purposefully discarded in this process. The discarded part includes the purity information mentioned above. Of course the set of *regsubs* heavily depends on the language and coding style employed in a specific catalog. It is not directly portable to other naming schemes and company conventions. However, the point is that it is relatively easy to transform the compound name, which has been abused to hold a plethora of auxiliary data, into a string which contains practically only core information, because repeating patterns and conventions can be identified. The algorithm employs a mixture of deletion operators triggered by keywords, substitution expressions for normalizations, and general alignment rules. Deletion operators typically operate from a keyword plus leading general separator characters right to the end of the name. The total number of purely textual keywords for the deletion *regsubs* is about 50. Basically, all filter functions are applied sequentially in a carefully chosen order to every name. An exception are compounds with missing ordering information, where the name is not the textual structure representation but the catalog number. Here the expressions which recognize numbers as purity data are skipped because some of the catalog order numbers are lexically too close to purity specifications. Within a data set, the naming peculiarities tend to be uniform. After closely supervising and tuning the clean-up processing of the first 500 entries very few changes in the *regsub* formulas were needed to obtain satisfactory results for the rest of the dataset. The necessary human monitoring in the setup phase, which solely consists of the comparison of pairs of strings before and after processing, is much simpler than interactions requiring the visual matching of substrings in the name to structural molecule features.

REFERENCES AND NOTES

(1) International Union of Pure and Applied Chemistry. *Nomenclature of Organic Chemistry*; Pergamon: Oxford, U.K. 1979; Sections A–F and H.

(2) Wisniewski, J. L. AUTONOM: System for Computer Translation of Structural Diagrams into IUPAC-Compatible Names. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 324–332.

(3) McDaniel, J. R.; Balmuth, J. R. Kekulé: OCR—Optical Chemical (Structure) Recognition. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 373–378.

(4) Contreras, L. M.; Allendes, C.; Alvarez, L. T.; Rozas, R. Computational Perception and Recognition of Digitized Molecular Structures. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 302–307.

(5) Vander Stouw, G. G.; Elliot, P. M.; Isenberg, A. C. Automated Conversion of Chemical Substance Names to Atom Bond Connection Tables. *J. Chem. Doc.* **1974**, *14*, 187–193.

(6) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 1. Introduction and Background to a Grammar-Based Approach. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 101–106.

(7) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 2. Development of a Formal Grammar. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 106–112.

(8) Cooke-Fox, D. I.; Kirby, G. H.; Rayner, J. D. Computer Translation of IUPAC Systematic Organic Chemical Nomenclature. 3. Syntax Analysis and Semantic Processing. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 112–118.

(9) Conrow, K. Computer Generation of Baeyer System Names of Saturated, Bridged, Bicyclic, Tricyclic and Tetracyclic Hydrocarbons. *J. Chem. Doc.* **1966**, *6*, 206–212.

(10) Van Binnendyk, D.; MacKay, A. C. Computer-Assisted Generation of IUPAC Names of Polycyclic Bridged Ring Systems. *Can. J. Chem.* **1973**, *51*, 718–723.

(11) Mockus, J.; Isenberg, A. C.; Vander Stouw, G. G. Algorithmic Generation of Chemical Abstracts Index Names. 1. General Design. *J. Chem. Inf. Comput. Sci.* **1981**, *21*, 183–195.

(12) Stillwell, R. N. Computer Translation of Systematic Chemical Nomenclature to Structural Formulas. Steroids. *J. Chem. Doc.* **1973**, *13*, 107–109.

(13) Gasteiger, J.; Ihlenfeldt, W. D.; Röse, P. A Collection of Computer Methods for Synthesis Design and Reaction Prediction. *Recl. Trav. Chim. Pays-Bas* **1992**, *111*, 270–290.

(14) Ihlenfeldt, W. D.; Gasteiger, J. Computer-Assisted Planning of Organic Syntheses: The Second Generation. *Angew. Chem.* and *Angew. Chem., Int. Ed. Engl.*, in press.

(15) Among the obstacles was the assignment of charges: nominal atomic charges were coded with isolated plus and minus characters on the plot area, without connection to any atom. The geometrically closest atom typically was not the atom which actually owned the charge symbol, so a number of distance and element-weighted heuristics had to be developed to solve the charge assignment problem.

(16) Ihlenfeldt, W. D.; Takahashi, Y.; Abe, H.; Sasaki, S. Computation and Management of Chemical Properties in CACTVS: An Extensible Networked Approach toward Modularity and Compatibility. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 109–116.

(17) Ihlenfeldt, W. D.; Gasteiger, J. Hash Codes for the Identification and Classification of Molecular Structure Elements. *J. Comp. Chem.* **1994**, *15*, 793–813.

(18) Hanessian, S.; Franco, J.; Gagnon, G.; Larameé, D.; Larouche, B. Computer-Assisted Analysis and Perception of Stereochemical Features in Organic Molecules Using the CHIRON Program. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 413–425.

(19) Mata, P.; Lobo, A.; Marshall, C.; Johnson, A. P. Implementation of the Cahn-Ingold-Prelog System for Stereochemical Perception in the LHASA Program. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 491–504.

(20) Gann, L.; Gasteiger, J. Eine Verarbeitung der R,S- und E,Z-Nomenklatur zur Spezifikation der Stereochemie. In *Software-Entwicklung in der Chemie 1*; Gasteiger, J., Ed.; Springer: Berlin, Germany, 1987; pp 17–33.

(21) Ugi, I.; Gruber, B.; Stein, N.; Demharter, A. Set-Valued Maps as a Mathematical Basis of Computer Assistance in Stereochemistry. *J. Chem. Inf. Comput. Sci.* **1990**, *30*, 485–489.

(22) Peishoff, C. E.; Jorgensen, W. L. Computer-Assisted Mechanistic Evaluation of Organic Reactions. 10. *Stereochemistry. J. Org. Chem.* **1985**, *50*, 3174–84.

(23) Bersohn, M. The Computer Derivation of Stereochemical Relations from the Chirality of Ring Atoms. *J. Chem. Soc. Perkin Trans. 1* **1979**, 1975–1977.

(24) Davis, H. W. *Computer Representation of the Stereochemistry of Organic Molecules: With Application to the Problem of Discovery of Organic Synthesis by Computer*; Birkhaeuser: Basel, Switzerland, 1974.

(25) Wipke, W. T.; Dyott, T. M. Simulation and Evaluation of Chemical Synthesis. Computer Representation and Manipulation of Stereochemistry. *J. Am. Chem. Soc.* **1974**, *96*, 4825–34.

(26) Cahn, R. S.; Ingold, C.; Prelog, V. Spezifikation der molekularen Chiralität. *Angew. Chem.* **1966**, *78*, 413–447; *Angew. Chem., Int. Ed. Engl.* **1966**, *5*, 385–419.

(27) Prelog, V.; Helmchen, G. Bases of the CIP system and proposal for a revision. *Angew. Chem.* **1982**, *94*, 614−631; *Angew. Chem., Int. Ed. Engl.* **1982**, *21*, 567−654.
(28) Röse, P.; Gasteiger, J. Automated Derivation of Reaction Rules for the EROS 6.0 System for Reaction Prediction. *Anal. Chim. Acta* **1990**, *235*, 163−168.
(29) The ESSSR is the Extended Smallest Set of Smallest Rings. It contains all SSSR rings plus every smallest ring with three adjacent atoms all in the SSSR which is not in the SSSR. In the case of cubane, this set consists of the six four-membered rings on the cube faces instead of the five SSSR rings. In the case of norbornane, the ESSSR includes the six-membered ring.
(30) Slocum, D. W.; Sugarman, D.; Tucker, S. P. The Two Faces of D and L Nomenclature. *J. Chem. Ed.* **1971**, *48*, 597−600.
(31) This is the end of the name. The nature of the 3% compound remains obscure.

CI950240B