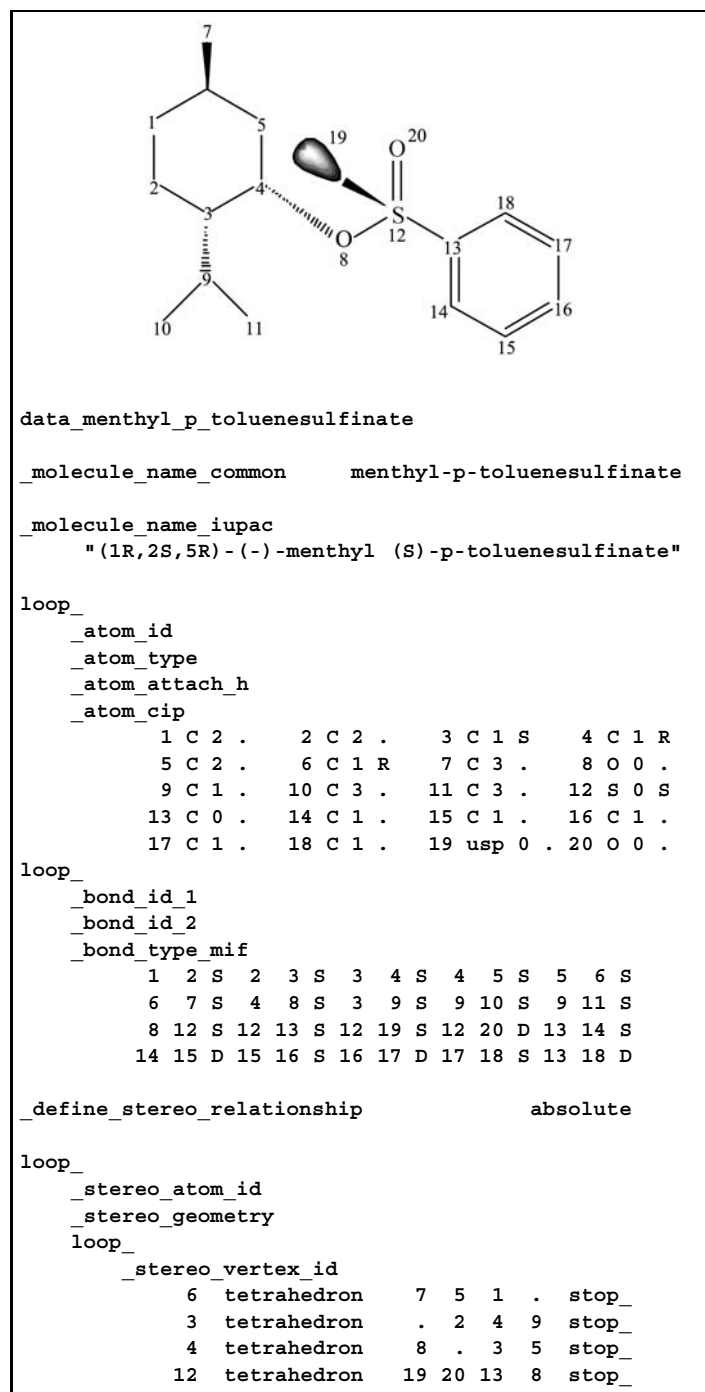


## 2.4. SPECIFICATION OF THE MOLECULAR INFORMATION FILE (MIF)

Fig. 2.4.8.3. Stereochemical data for menthyl-*p*-toluenesulfinate.

'sequenced data' (via the code `seq`). This permits a value string to contain alternative 'values' satisfying the following constructs: (a) the value string  $v_1, v_2, v_3$  signals that a data item must have the value  $v_1$  or  $v_2$  or  $v_3$ , and (b) the value string  $v_1:v_2$  signals that a data item must have a value in the range  $v_1$  to  $v_2$ . Combinations of these constructions are permitted. All values must comply with the requirements defined by the attributes `_enumeration` and `_enumeration_range`.

An example of a substructural query in a MIF is shown in Fig. 2.4.9.1 for a conjugated ketone or thioketone fragment. Points of permitted variability of atom properties occur at atom 1, an  $sp^3$  carbon atom that must have at least one attached hydrogen atom, and at atom 5, which can be S or O. The conjugated multiple C—C bond (3–4) is defined to be either localized double, delocalized double or aromatic using CCDC bonding conventions. Query coding of this type should be readily generated from most

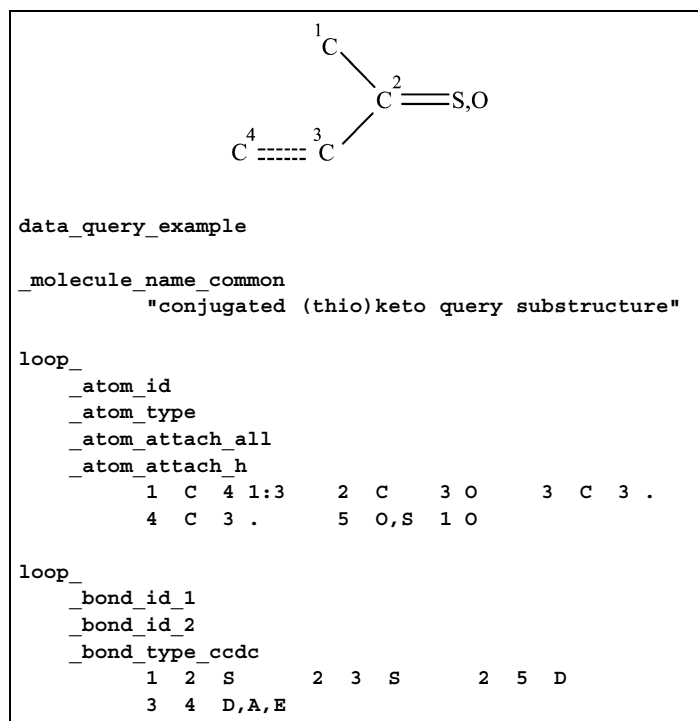


Fig. 2.4.9.1. Query substructure for conjugated ketones or thioketones. Atom C1 is  $sp^3$  hybridized (total number of attached hydrogen and non-hydrogen atoms = 4) and carries at least one hydrogen atom. Bond C3—C4 may be localized double (D), aromatic (A) or delocalized double (E) in CCDC conventions.

graphical 2D search interfaces or be readable directly by a variety of 2D substructure search programs.

## 2.4.10. Conclusion

The present proliferation of formats for chemical applications tends to inhibit and complicate the exchange and use of chemical data. Many widely used chemical formats have a finite half-life because they are inflexible and not readily extensible. Others offer universality [e.g. Abstract Syntax Notation 1 (ISO, 2002a,b); Dalby *et al.* (1992); see <http://www.daylight.com/smiles/>] but lack visual simplicity, generality or machine readability. Nevertheless, the Molecular Information File approach has these properties but needs significantly more development to be a viable exchange approach for mainstream chemistry. The MIF dictionary enables chemical data items to be defined at high precision and this offers real benefits for the creation of a domain ontology in this field.

Herein we have outlined the basic MIF approach and provided definitions for an initial core of data items that are fundamental for the representation of 2D and 3D chemical structures and 2D substructures. These core data items cover most of the basic chemical data-exchange requirements of molecular modelling and database applications, but are clearly only a first step towards the level of chemical data exchange needed. Future MIF developments in applications software and, particularly, in data definitions are expected to encompass other aspects of chemistry. These developments will need the collaborative involvement and support of subject specialists from both academia and industry.

## References

- Allen, F. H. (2002). *The Cambridge Structural Database: a quarter of a million crystal structures and rising*. *Acta Cryst.* **B58**, 380–388.
- Allen, F. H., Barnard, J. M., Cook, A. P. F. & Hall, S. R. (1995). *The Molecular Information File (MIF): core specifications of a new standard format for chemical data*. *J. Chem. Inf. Comput. Sci.* **35**, 412–427.

## 2. CONCEPTS AND SPECIFICATIONS

- Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *The development of versions 3 and 4 of the Cambridge Structural Database System*. *J. Chem. Inf. Comput. Sci.* **31**, 187–204.
- Allen, F. H. & Glusker, J. P. (2002). *Preface*. *Acta Cryst.* **B58**, Part 3.
- Allen, F. H. & Rogers, D. (1970). *X-ray studies of terpenoid derivatives, Part III. A re-determination of the crystal structure of (+)-3-bromocamphor: the absolute configuration of (+)-camphor*. *J. Chem. Soc. B*, pp. 632–636.
- ASTM (1994). *Standard specification for the content of computerized chemical structural information files or data sets*. ASTM Standard E 1586-93. American Society for Testing and Materials, Philadelphia, PA, USA.
- Barnard, J. M. (1990). *Draft specification for revised version of the Standard Molecular Data (SMD) format*. *J. Chem. Inf. Comput. Sci.* **30**, 81–96.
- Barnard, J. M. & Cook, A. P. F. (1992). *The Molecular Information File (MIF): a standard format for molecular information*. Report. Chemical Structure Association, London, England.
- Barnard, J. M., Cook, A. P. F. & Rohde, B. (1990). *Beyond the structure diagram*, edited by D. Bawden & E. Mitchell, pp. 29–41. Chichester: Ellis Horwood.
- Bebak, H., Buse, C., Donner, W. T., Hoever, P., Jacob, H., Klaus, H., Pesch, J., Roemelt, J., Schilling, P., Woost, B. & Zirz, C. (1989). *The Standard Molecular Data format (SMD format) as an integration tool in computer chemistry*. *J. Chem. Inf. Comput. Sci.* **29**, 1–5.
- Brown, I. D. (1988). *Standard Crystallographic File Structure-87*. *Acta Cryst.* **A44**, 232.
- Cahn, R. S., Ingold, C. K. & Prelog, V. (1966). *Specification of molecular chirality*. *Angew. Chem. Intl Ed. Engl.* **5**, 385–415.
- Dalby, A., Nourse, J. G., Hounshell, W. D., Gushurst, A. K. I., Grier, D. L., Leland, B. A. & Laufer, J. (1992). *Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited*. *J. Chem. Inf. Comput. Sci.* **32**, 244–255.
- Feldmann, R. J., Milne, G. W. A., Heller, S. R., Fein, A., Miller, J. A. & Koch, B. (1977). *An interactive substructure search system*. *J. Chem. Inf. Comput. Sci.* **17**, 157–163.
- Hall, S. R. (1991). *The STAR File: a new format for electronic data transfer and archiving*. *J. Chem. Inf. Comput. Sci.* **31**, 326–333.
- Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *The Crystallographic Information File (CIF): a new standard archive file for crystallography*. *Acta Cryst.* **A47**, 655–685.
- Hall, S. R. & Cook, A. P. F. (1995). *STAR Dictionary Definition Language: initial specification*. *J. Chem. Inf. Comput. Sci.* **35**, 819–825.
- Hall, S. R. & Spadaccini, N. (1994). *The STAR File: detailed specifications*. *J. Chem. Inf. Comput. Sci.* **34**, 505–508.
- Harary, F. (1972). *Graph theory*, 3rd ed. London: Addison-Wesley.
- ISO (2002a). ISO/IEC 8824-1. *Abstract Syntax Notation One (ASN.1). Specification of basic notation*. Geneva: International Organization for Standardization.
- ISO (2002b). ISO/IEC 8825-1. *ASN.1 encoding rules. Specification of Basic Encoding Rules (BER), Canonical Encoding Rules (CER) and Distinguished Encoding Rules (DER)*. Geneva: International Organization for Standardization.
- Mockus, J. & Stobaugh, R. E. (1980). *The Chemical Abstracts Service Chemical Registry System. VII. Tautomerism and alternating bonds*. *J. Chem. Inf. Comput. Sci.* **20**, 18–22.
- Prelog, V. & Helmchen, G. (1982). *Basic principles of the CIP-system and proposals for a revision*. *Angew. Chem. Intl Ed. Engl.* **21**, 567–583.
- Wipke, W. T. & Dyott, T. M. (1975). *Use of ring assemblies in ring perception algorithm*. *J. Chem. Inf. Comput. Sci.* **15**, 140–147.