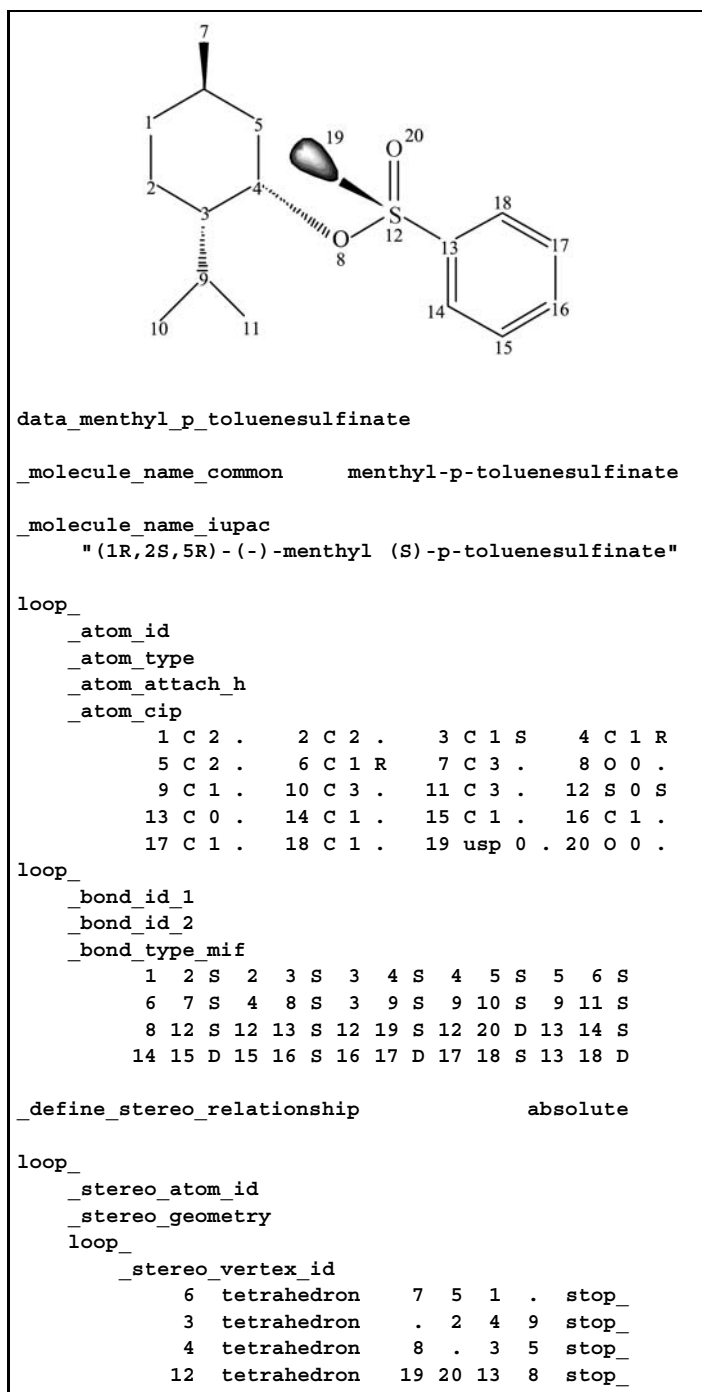


2.4. SPECIFICATION OF THE MOLECULAR INFORMATION FILE (MIF)

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

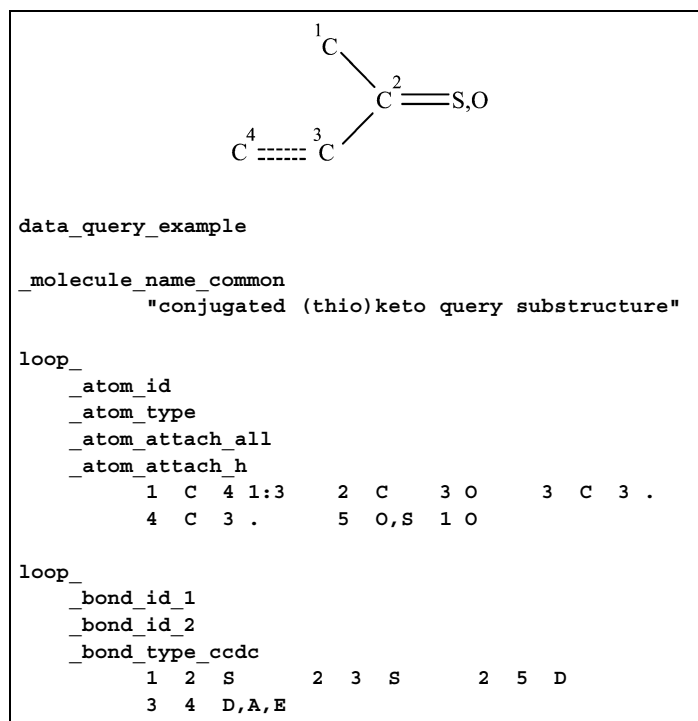


Fig. 2.4.9.1. Query substructure for conjugated ketones or thioketones. Atom C1 is sp^3 hybridized (total number of attached hydrogen and non-hydrogen atoms = 4) and carries at least one hydrogen atom. Bond C3—C4 may be localized double (D), aromatic (A) or delocalized double (E) in CCDC conventions.

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, 2002*a,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.