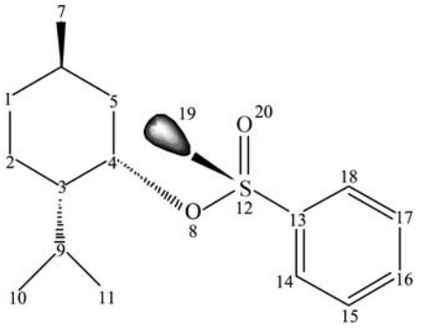
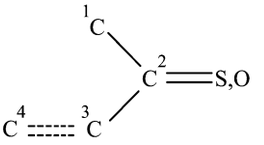


```
data_menthyl_p_toluenesulfinate

_molecule_name_common     menthyl-p-toluenesulfinate

_molecule_name_iupac
     "(1R,2S,5R)-(-)-menthyl (S)-p-toluenesulfinate"

loop_
    _atom_id
    _atom_type
    _atom_attach_h
    _atom_cip
           1 C 2 .    2 C 2 .    3 C 1 S    4 C 1 R
           5 C 2 .    6 C 1 R    7 C 3 .    8 O 0 .
           9 C 1 .   10 C 3 .   11 C 3 .   12 S 0 S
          13 C 0 .   14 C 1 .   15 C 1 .   16 C 1 .
          17 C 1 .   18 C 1 .   19 usp 0 . 20 O 0 .
loop_
    _bond_id_1
    _bond_id_2
    _bond_type_mif
         1  2 S  2  3 S  3  4 S  4  5 S  5  6 S
         6  7 S  4  8 S  3  9 S  9 10 S  9 11 S
         8 12 S 12 13 S 12 19 S 12 20 D 13 14 S
        14 15 D 15 16 S 16 17 D 17 18 S 13 18 D

_define_stereo_relationship             absolute

loop_
    _stereo_atom_id
    _stereo_geometry
    loop_
        _stereo_vertex_id
             6  tetrahedron    7  5  1  .  stop_
             3  tetrahedron    .  2  4  9  stop_
             4  tetrahedron    8  .  3  5  stop_
            12  tetrahedron   19 20 13  8  stop_
```

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 $v1, v2, v3$ signals that a data item must have the value $v1$ or $v2$ or $v3$, and (*b*) the value string $v1{:}v2$ signals that a data item must have a value in the range $v1$ to $v2$. 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



```
data_query_example

_molecule_name_common
         "conjugated (thio)keto query substructure"

loop_
    _atom_id
    _atom_type
    _atom_attach_all
    _atom_attach_h
          1  C  4 1:3    2  C    3 O     3  C  3 .
          4  C  3 .      5  O,S  1 O

loop_
    _bond_id_1
    _bond_id_2
    _bond_type_ccdc
          1  2  S      2  3  S      2  5  D
          3  4  D,A,E
```

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.* B**58**, 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.