

## WHITEPAPER

# THE KEYS TO UNDERSTANDING MDL KEYSSET TECHNOLOGY

---

## INTRODUCTION

Structure-based keys used in MDL technology (now Accelrys) are defined by the contents of two ASCII files, `enkfil.dat` and `eksfil.dat`. The `enkfil.dat` file defines which structural features will set which keybits. The `eksfil.dat` file defines a subset of the `enkfil.dat`-defined keybits, which will be used as a “subset keyset”. In default installations the `enkfil.dat` file defines the MDL 960-bit keyset, and the `eksfil.dat` file defines the MDL 166-bit keyset. There are two additional key files which are encountered in MDL software. They are the `bnkfil.dat` and `bksfil.dat` files, and they are simply binary versions of the `enkfil.dat` and `eksfil.dat` files.

Reading these files is straightforward, and allows anyone to understand what structural features set which keybits. Additionally, because the identity of the keyset and subset-keyset are defined by these files, a user can easily define their own keyset and/or subset keyset. Such custom keysets can be expected to outperform the default keyset for focused or specialized tasks.

## TERMINOLOGY

In this paper we will use the following definitions to refer to keyset components. “Descriptor” refers to a molecular feature. Descriptors can be encoded into binary “keybits.” There can be a one-to-one, a many-to-one or a many-to-many relationship between descriptors and keybits. An ordered collection of keybits constitutes a “keyset.” “Keys” is context-sensitive and is used to refer either to keybits or keysets. Additionally, Accelrys has specific definitions for a number of general chemistry terms. “Heteroatom” is used to refer to any non-C, non-H atom, and is abbreviated “Q”. “Halogens” consist of F, Cl, Br, and I. “Other” atoms include any atoms other than H, C, N, O, Si, P, S, F, Cl, Br, and I, and is abbreviated “Z”. “Aromatic” refers to bonds, which are either “Kekule aromatic” or “arom5”. “Kekule aromatic” bonds are those in a six-membered ring system with alternating double and single bonds, or the perimeter bonds of an azulene. “Arom5” bonds are those in a five-membered ring with two double bonds and a “heteroatom” or C- at the apex of the ring.



The fields are defined, in Fortran notation, as (I3, I3, I3, I4, I3, I3, I5, I5, I5). Text after the fixed-width fields is presently ignored.

## N1, N2 AND N3—BASIC DESCRIPTOR ENCODING

The largest block of descriptors makes use of algorithmically calculated atom-based properties. Specifically we perceive 30 properties of type A, as listed in Appendix II, and 26 properties of type P (most of which are the same as type A properties), as listed in Appendix III. Additionally, we perceive 32 one-atom environments, as listed in Appendix VI.

### N1 = 0

These descriptors consist of one or two properties of type A (see Appendix II) located on a single atom. The properties are encoded as N2 and N3, with  $N2 \leq N3$ .

### $1 \leq N1 \leq 4$

These descriptors consist of an atom with property P (see Appendix III), separated by one to four bonds from a second atom with a property P'. N1 encodes the number of bonds between the atoms, and N2 and N3 encode the appropriate values of P and P'. Once again,  $N2 \leq N3$ .

### N1 = 5

These descriptors include both custom and Sgroup features (see Appendix IV). The first sub-block of these descriptors encodes a series of properties, which are used in the 166-bit MDL keyset. The next sub-block of 256 descriptors encodes atom types of 1-256. Atom types between 1 and 103 correspond directly with periodic table elements. The range 104-256 allows encoding of custom atom types (although the ptable limits one to only 200 symbols). A final sub-block of descriptors encodes a variety of MDL Sgroup properties.

### N1 = 6

These descriptors encode one of 264 atom-bond-atom combinations (see Appendix V). Atoms include C, N, O, Si, P, S, F, Cl, Br, I, and "other" (Z), with bond types of "single" (-), "double" (=), "triple" (#) and "Kekule aromatic" (%).

### N1 = 7

These descriptors consist of an atom with property A (see Appendix II) located in the center of a particular atom environment (see Appendix VI). In this case N2 encodes the atom environment and N3 encodes the property A.

## N4—THE OCCURRENCE COUNT

In converting these descriptors to keybits, we also need to define an occurrence count, which is encoded in N4. The keybit will be set for N4 or more occurrences of a descriptor. This allows us to set keybits for occurrence counts of "1 or more" up to "999 or more".

## N5 THROUGH N9—KEYBIT ENCODING

The last five numbers in the keybit definition are used to control which keybit(s) is (are) set by the descriptor defined by the first four numbers. N5 is used to specify the number of keybits which are set, which is in the range of one to three. N6 is a flag indicating whether or not the keybit(s) set are also set by other descriptors [(1) = yes; (0) = no]. The final three numbers, N7, N8, and N9, identify the keybits, with '0' used for padding.

## FINAL REMARKS

A number of properties are incompatible, resulting in the fact that fewer descriptors are chemically possible than are mathematically allowed. Additionally, a number of descriptors can be encoded two different ways, most frequently by reversing N2 and N3. The present MDL algorithm will only set keybits for N2 less than or equal to N3 in these instances. As a result we can encode 3,234 different descriptors encoding occurrence counts of "1 or more". Since these descriptors can have occurrence counts of "1 or more" to "999 or more", we have the ability, in theory, to produce in excess of three million distinct descriptors, which can be combined into innumerable keysets.

## EXAMPLES

Returning to the default enkfil.dat file (pg. 1), we read the first line of the keybit definition section:

```
0 0 3 3 1 0 872 0 0
```

N1 = 0, so we identify this as a descriptor encoding one or two properties located on a single atom. From Appendix III, we see that N2 = 0 is a null property, and N3 = 3 corresponds to an atom with multiple, non-aromatic bonds.

N4 = 3 corresponds to an occurrence count of three or more.

Turning to N5 through N9, we see that only one keybit is set (N5 = 1), no other descriptor sets this keybit (N6 = 0), and the keybit set is 872 (N7 = 872).

If we look at the fourth descriptor:

```
0 0 4 2 2 1 728 712 0
```

we see that it encodes an atom (N1 = 0) with at least four neighbors (N3 = 4), and an occurrence count of 2 or more (N4 = 2). We see that this descriptor sets two keys (N5 = 2), and that the keybit(s) can be set by other descriptors (N6 = 1). In fact, if we search the enkfil.dat

file for 728 and 712, we find the following lines:

```
1 8 15 2 3 1 728 588 612
```

```
1 16 18 3 2 1 685 728 0
```

```
3 8 8 3 1 1 728 0 0
```

and

```
1 8 18 3 3 1 712 760 808
1 17 23 3 2 1 712 804 0
3 2 2 3 1 1 712 0 0
3 3 18 3 2 1 712 713 0
```

From the preceding section, we can identify all of these descriptors as corresponding to properties on two atoms, separated by one-four bonds ( $1 \leq N1 \leq 4$ ).

Concretely, the first descriptor corresponds to a carbon atom with at least two single bonds and at least two hydrogens ( $N2 = 8$ ), one bond away ( $N1 = 1$ ) from an atom at a ring/chain boundary, with the path between the atoms passing through the chain bond ( $N3 = 15$ ). The occurrence count is two or more ( $N4 = 2$ ), and this descriptor sets three keybits ( $N5 = 3$ ), which can be set by other descriptors ( $N6 = 1$ ). The keybits set are 728, 588 and 612 ( $N7-9$ ).

Similarly, the last descriptor corresponds to an atom with one or more multiple bonds ( $N2 = 3$ ) separated by three bonds ( $N1 = 3$ ) from an atom at a ring/chain boundary, with the path passing through the ring bond ( $N3 = 18$ ). The occurrence count is three ( $N4 = 3$ ), two keybits are set ( $N5 = 2$ ), and the keybits can be set by other descriptors ( $N6 = 1$ ). The keybits set are 712 and 713 ( $N7-8$ ).

Continuing to read the enkfil.dat file, we come to a section containing descriptors with  $N1 = 5$ :

```
....
5 0 1 1 1 0 138 0 0
5 0 2 1 1 0 1 0 0
5 0 3 1 1 0 116 0 0
5 0 4 1 1 0 918 0 0
```

....

The identity of these descriptors can be read directly from Appendix IV. Thus, the first four entries in this section correspond to "charge", "isotope", "other atom" (any atom other than H, C, N, O, Si, P, S, F, Cl, Br and I) and "methyl". Occurrence counts are one or more, and each descriptor sets a single keybit, which is not set by any other descriptor in this keyset.

The next section of the enkfil.dat file contains descriptors with  $N1 = 6$ :

```
....
6 018 1 1 0 908 0 0
6 018 2 1 0 853 0 0
6 018 3 1 0 585 0 0
6 019 1 1 0 907 0 0
6 019 2 1 0 831 0 0
6 019 3 1 1 713 0 0
```

....

These correspond to a set of atom-bond-atom properties, which are listed in Appendix V. We see that the first three descriptors are C-N bonds ( $N_2 = 0$ ,  $N_3 = 18$ ), with occurrence counts of one or more, two or more and three or more. These three descriptors each set a single keybit, which is not set by another descriptor.

Similarly, the next three descriptors encode different occurrence counts for C-O bonds, with the minor complication that the last of these descriptors sets keybit 713, which can be set by another descriptor (one of which we have already encountered).

The final section of descriptors has  $N_1 = 7$ :

```
7 0 3 1 1 0 6 1 9 0 0
```

and corresponds to atoms with a property in a certain chemical environment. The properties are listed in Appendix II, and the environments are listed in Appendix VI. Thus, we see that the atomic environment is a carbon, with at least two carbon neighbors ( $N_2 = 0$ ). The atomic property is that the carbon has at least one multiple bond (double or triple) ( $N_3 = 3$ ). Again, the occurrence count is one or more, one key is set and no other descriptor sets this key.

## WEIGHTS

Keybit weights are defined in the `enkfil.dat` file following the keybit definition termination line:

```
1 0 0 0 0 0 0 0 0 0
```

```
1 0 0
```

```
1 0 0
```

```
1 0 0
```

```
....
```

Weights are stored one per line in I5 format; their order corresponds to the keybit order, i.e., the first weight corresponds to keybit one, the second to keybit two, and the 960th to keybit 960. These weights are used in similarity calculations.

## EKSFIL.DAT—DEFINING A SUBSET

The `eksfil.dat` file is used to define a subset of the keybits defined by the `enkfil.dat` file, which will be used as a subset keyset. In default installations of MDL software, the `eksfil.dat` file defines the familiar 166-bit keyset, which is a subset of the 960-bit keyset. The default `eksfil.dat` file starts:

```
3 25 85 166 0 0 0 0 0 0
```

```
1 ISOTOPE
```

```
2 103 < ATOMIC NO. < 256
```

```
29 GROUP IVA,VA,VIA PERIODS 4-6 (GE..)
```

```
4 ACTINIDE
```

6 GROUP IIIB,IVB (SC.)  
 7 LANTHANIDE  
 8 GROUP VB,VIB,VIIB (V.)  
 9 QAAA@1

....

The first line is a header, which contains the month, day and year followed by the number of subset keybits defined. The other fields in the header line are presently ignored. The header fields are defined, in Fortran notation, as (10I8). The next lines define, in order, the mapping of a subset keybit, given by the line number, to the enkfil.dat-defined keybit. The keybits are listed one per line as an I5 formatted integer. The rest of the line is ignored and traditionally holds a comment relating to features, which set that particular keybit. MDL Line Notation (formerly called MDL or Substructure-search Query Language) is frequently used. The notation "&..." is sprinkled throughout these comments and highlights keybits, which are set by more than one descriptor.

## CONCLUSIONS

In MDL software the keyset and subset keyset used (960-bit and 166-bit for default installations) are defined by the enkfil.dat and eksfil.dat files. Both have a simple structure and can be read to discern the structural features responsible for setting each keybit.

## OTHER RESOURCES

Approaches to constructing custom keysets can be found in J. L. Durant, B. A. Leland, D. R. Henry and J. G. Nourse, "Reoptimization of MDL Keys for Use in Drug Discovery", *J. Chem. Inf. Comput. Sci.*, 42, 1273-1280 (2002).

A discussion of MDL Line Notation (also known as MDL or Substructure-search Query Language) can be found in the Accelrys Cheshire online documentation. From the index choose "MDL Line Notation, create mol or rxn." There is also a description in the ISIS/Host Administrators Guide, Version 5.0, in Chapter 7 "Customizing Chemistry". The appropriate section is entitled "Substructure-Search Query Language."

A discussion of the use of custom enkfil.dat and eksfil.dat files can be found in the Cheshire online documentation. From the index choose "SSKeys, using customized files." Chapters 5 and 6 of the ISIS/Host Administrators Guide, Version 5.0 also have information on loading of enkfil.dat and eksfil.dat files.

To learn more about Accelrys cheminformatics, go to [accelrys.com/lea](http://accelrys.com/lea)

**Appendix I: Basic Descriptor Encoding Overview**

N1	N2	N3	Brief Description	Appendix
0	0	A	atom with one property	II
0	A	A'	atom with two properties	II
1	P	P'	one bond path between atom with property P and another atom with property P'	III
2	P	P'	two bond...	III
3	P	P'	three bond...	III
4	P	P'	four bond...	III
5	n	n	custom and Sgroup	IV
6	n	n	atom-bond-atom combinations	V
7	X	A	extended atom with A property at center	VI, II

**Appendix II: Single-atom Atom-based Properties**

N	A
0	null
1	atom with at least three neighbors
2	heteroatom
3	atom involved in some multiple bonds, not aromatic
4	atom with at least four neighbors
5	atom with at least two heteroatom neighbors
6	atom with at least three heteroatom neighbors
7	heteroatom with at least one hydrogen attached
8	carbon with at least two single bonds and at least two hydrogens attached
9	carbon atom in a C=C double bond
10	atom has at least two single bonds
11	atom has at least three single bonds
12	atom is in at least two different six-membered rings
13	not used
14	atom has more than two ring bonds
15	atom is at a ring/chain boundary
16	central atom is at an aromatic/non-aromatic boundary
17	atom with more than one chain bond
18	atom is in a ring
19	aromatic atom
20	atom is a heteroatom in a ring.
21	rare properties: atom with five or more neighbors, atom in four or more rings, or atom types other than H, C, N, O, S, F, Cl, Br, or I.
22	rare properties: atom has a charge, is an isotope, has two or more multiple bonds, or has a triple bond
23	nitrogen
24	sulfur
25	oxygen
26	atom in a three-membered ring
27	atom in a four-membered ring
28	atom in a five-membered ring
29	atom in a six-membered ring
30	atom has two neighbors, each with three or more neighbors (including the central atom).
31	atom has two hydrocarbon (CH <sub>2</sub> ) neighbors

**Appendix III: Atom-based Properties**

N	P
0	null
1	atom with at least three neighbors
2	heteroatom
3	atom involved in one or more multiple bonds, not aromatic
4	atom with at least four neighbors
5	atom with at least two heteroatom neighbors
6	atom with at least three heteroatom neighbors
7	heteroatom with at least one hydrogen attached
8	carbon with at least two single bonds and at least two hydrogens attached
9	carbon with at least one single bond and at least three hydrogens attached
10	halogen
11	atom has at least three single bonds
12	atom is in at least two different six-membered rings
13	not used
14	atom has more than two ring bonds
15	atom is at a ring/chain boundary. When a comparison is done with another atom, the path passes through the chain bond.
16	atom is at an aromatic/non-aromatic boundary. When a comparison is done with another atom, the path passes through the aromatic bond
17	atom with more than one chain bond
18	atom is at a ring/chain boundary. When a comparison is done with another atom, the path passes through the ring bond
19	atom is at an aromatic/non-aromatic boundary. When a comparison is done with another atom, the path passes through the non-aromatic bond.
20	atom is a heteroatom in a ring.
21	rare properties: atom with five or more neighbors, atom in four or more rings or atom types other than H, C, N, O, S, F, Cl, Br, or I.
22	rare properties: atom has a charge, is an isotope, has two or more multiple bonds or has a triple bond
23	nitrogen
24	sulfur
25	oxygen
26	not used
27	not used
28	not used
29	not used
30	atom has two neighbors, each with three or more neighbors (including the central atom).
31	atom has two hydrocarbon (CH <sub>2</sub> ) neighbors

**Appendix IV: Custom and Sgroup Properties**

N1	N2	N3	Property
5	0	1	charge (in structure somewhere)
5	0	2	isotope
5	0	3	“other” atom type
5	0	4	CH <sub>3</sub>
5	0	5	halogen
5	0	6	NH <sub>2</sub>
5	0	7	five-membered ring
5	0	8	six-membered ring
5	0	9	Kekule-aromatic ring
5	0	10	seven-membered ring
5	0	11	eight-membered ring
5	0	12	103 < atomtype < 256
5	0	13	more than one fragment
5	1	1-31	atomtypes 1-31
5	2	0-31	atomtypes 31-63
5	3	0-31	atomtypes 64-95
5	4	0-31	atomtypes 96-127
5	5	0-31	atomtypes 128-159
5	6	0-31	atomtypes 160-191
5	7	0-31	atomtypes 192-223
5	8	0-31	atomtypes 224-255
5	9	1	atomtype 256
5	11	1	Component Sgroup type
5	11	2	RU Sgroup type
5	11	3	Monomer Sgroup type
5	11	4	Copolymer Sgroup type
5	11	5	Alternating copolymer subtype
5	11	6	Random copolymer subtype
5	11	7	Block copolymer subtype
5	11	8	Graft Sgroup type
5	11	9	Formulation Sgroup type
5	11	10	Mixture Sgroup type
5	11	11	Crosslink Sgroup type
5	11	12	Modification Sgroup type
5	11	13	Any polymer Sgroup type
5	11	14	Data Sgroup type
5 11 15 thru 5 13 5			Data Sgroup field number
5	13	6	Mer Sgroup type

Appendix V: Atom-Bond-Atom Properties<sup>1</sup>

N2	N3		N2	N3		N2	N3		N2	N3		N2	N3	
0	17	C-C	4	8	Br-Br	11	6	P=P	18	4	S#S	25	2	N%N
0	18	C-N	4	9	Br-Si	11	7	P=F	18	5	S#Cl	25	3	N%O
0	19	C-O	4	10	Br-I	11	8	P=Br	18	6	S#P	25	4	N%S
0	20	C-S	4	15	Br-Y	11	9	P=Si	18	7	S#F	25	5	N%Cl
0	21	C-Cl	4	25	Si-Si	11	10	P=I	18	8	S#Br	25	6	N%P
0	22	C-P	4	26	Si-I	11	15	P=Z	18	9	S#Si	25	7	N%F
0	23	C-F	4	31	Si-Z	11	23	F=F	18	10	S#I	25	8	N%Br
0	24	C-Br	5	10	I-I	11	24	F=Br	18	15	S#Z	25	9	N%Si
0	25	C-Si	5	15	I-Z	11	25	F=Si	18	21	Cl#Cl	25	10	N%I
0	26	C-I	7	31	Z-Z	11	26	F=I	18	22	Cl#P	25	15	N%Z
0	31	C-Z	8	17	C=C	11	31	F=Z	18	23	Cl#F	25	19	O%O
1	2	N-N	8	18	C=N	12	8	Br=Br	18	24	Cl#Br	25	20	O%S
1	3	N-O	8	19	C=O	12	9	Br=Si	18	25	Cl#Si	25	21	O%Cl
1	4	N-S	8	20	C=S	12	10	Br=I	18	26	Cl#I	25	22	O%P
1	5	N-Cl	8	21	C=Cl	12	15	Br=Z	18	31	Cl#Z	25	23	O%F
1	6	N-P	8	22	C=P	12	25	Si=Si	19	6	P#P	25	24	O%Br
1	7	N-F	8	23	C=F	12	26	Si=I	19	7	P#F	25	25	O%Si
1	8	N-Br	8	24	C=Br	12	31	Si=Z	19	8	P#Br	25	26	O%I
1	9	N-Si	8	25	C=Si	13	10	I=I	19	9	P#Si	25	31	O%Z
1	10	N-I	8	26	C=I	13	15	I=Z	19	10	P#I	26	4	S%S
1	15	N-Z	8	31	C=Z	15	31	Z=Z	19	15	P#Z	26	5	S%Cl
1	19	O-O	9	2	N=N	16	17	C#C	19	23	F#F	26	6	S%P
1	20	O-S	9	3	N=O	16	18	C#N	19	24	F#Br	26	7	S%F
1	21	O-Cl	9	4	N=S	16	19	C#O	19	25	F#Si	26	8	S%Br
1	22	O-P	9	5	N=Cl	16	20	C#S	19	26	F#I	26	9	S%Si
1	23	O-F	9	6	N=P	16	21	C#Cl	19	31	F#Z	26	10	S%I
1	24	O-Br	9	7	N=F	16	22	C#P	20	8	Br#Br	26	15	S%Z
1	25	O-Si	9	8	N=Br	16	23	C#F	20	9	Br#Si	26	21	Cl%Cl
1	26	O-I	9	9	N=Si	16	24	C#Br	20	10	Br#I	26	22	Cl%P
1	31	O-Z	9	10	N=I	16	25	C#Si	20	15	Br#Z	26	23	Cl%F
2	4	S-S	9	15	N=Z	16	26	C#I	20	25	Si#Si	26	24	Cl%Br
2	5	S-Cl	9	19	O=O	16	31	C#Z	20	26	Si#I	26	25	Cl%Si
2	6	S-P	9	20	O=S	17	2	N#N	20	31	Si#Z	26	26	Cl%I
2	7	S-F	9	21	O=Cl	17	3	N#O	21	10	I#I	26	31	Cl%Z
2	8	S-Br	9	22	O=P	17	4	N#S	21	15	I#Z	27	6	P%P
2	9	S-Si	9	23	O=F	17	5	N#Cl	23	31	Z#Z	27	7	P%F
2	10	S-I	9	24	O=Br	17	6	N#P	24	17	C%Cl	27	8	P%Br
2	15	S-Z	9	25	O=Si	17	7	N#F	24	18	C%N	27	9	P%Si
2	21	Cl-Cl	9	26	O=I	17	8	N#Br	24	19	C%O	27	10	P%I
2	22	Cl-P	9	31	O=Z	17	9	N#Si	24	20	C%S	27	15	P%Z
2	23	Cl-F	10	4	S=S	17	10	N#I	24	21	C%Cl	27	23	F%F
2	24	Cl-Br	10	5	S=Cl	17	15	N#Z	24	22	C%P	27	24	F%Br
2	25	Cl-Si	10	6	S=P	17	19	O#O	24	23	C%F	27	25	F%Si
2	26	Cl-I	10	7	S=F	17	20	O#S	24	24	C%Br	27	26	F%I
2	31	Cl-Z	10	8	S=Br	17	21	O#Cl	24	25	C%Si	27	31	F%Z
3	6	P-P	10	9	S=Si	17	22	O#P	24	26	C%I	28	8	Br%Br
3	7	P-F	10	10	S=I	17	23	O#F	24	31	C%Z	28	9	Br%Si
3	8	P-Br	10	15	S=Z	17	24	O#Br				28	10	Br%I
3	9	P-Si	10	21	Cl=Cl	17	25	O#Si				28	15	Br%Z
3	10	P-I	10	22	Cl=P	17	26	O#I				28	25	Si%Si
3	15	P-Z	10	23	Cl=F	17	31	O#Z				28	26	Si%I
3	23	F-F	10	24	Cl=Br							28	31	Si%Z
3	24	F-Br	10	25	Cl=Si							29	10	I%I
3	25	F-Si	10	26	Cl=I							29	15	I%Z
3	26	F-I	10	31	Cl=Z							31	31	Z%Z
3	31	F-Z												

1. "Z" is any atom other than C, N, O, Si, P, S, F, Cl, Br, I; bond types are "single" (-), "double" (=), "triple" (#) and "ring" (%)

**Appendix VI: Atomic Environments**

n	Atom Environment <sup>2</sup>
0	C(CC)
1	C(CCC)
2	C(CN)
3	C(CCN)
4	C(NN)
5	C(CNN)
6	C(NNN)
7	C(CO)
8	C(CCO)
9	C(NO)
10	C(CNO)
11	C(NNO)
12	C(OO)
13	C(COO)
14	C(NOO)
15	C(OOO)
16	Q(CC)
17	Q(CCC)
18	Q(CN)
19	Q(CCN)
20	Q(NN)
21	Q(CNN)
22	Q(NNN)
23	Q(CO)
24	Q(CCO)
25	Q(NO)
26	Q(CNO)
27	Q(NNO)
28	Q(OO)
29	Q(COO)
30	Q(NOO)
31	Q(OOO)

2. The first symbol is the central atom, with atoms bonded to the central atom listed in parentheses. "Q" is any non-C, non-H atom. If only two atoms are in parentheses, there is no implication concerning the other atoms bonded to the central atom.