Quick Guide to Creating a Structure-Data File (SD File) for Electronic Common Technical Document (eCTD) Submissions

Disclaimer

The mention of commercial products, their sources, or their use in connection with material reported herein is not to be construed as either an actual or implied endorsement of such products by the Department of Health and Human Services.

Introduction

The pharmaceutical industry may now submit structures to the Agency through the Electronic Common Technical Document (eCTD) as a single Structure-Data File (SD File) with standardized data fields meeting the Agency's cheminformatics and review needs. SD Files are acceptable in all of Module 3. This guide is designed to help you quickly create an SD File. It is not intended to be a comprehensive review of the SD File format. Extensive details on the file format can be found on the internet[1]. For general questions on formatting or submitting SD Files, please contact FDA's Global Substance Registration System (GSRS) at FDA-SRS@fda.hhs.gov.

What is an SD File?

- An SD File is a chemical structure-data file format that can associate data with one or more chemical structures.
- An SD File is also commonly called an ".sdf", "SD File", or just an "SDF."
- SD Files have an ".sdf" file extension.

SD Files encode chemical structure data using the molfile connection table format. This format depicts chemical structures using a block of text that lists the atoms, bonds, connectivity, and coordinates.

```
Acetamide


  4  3  0  0  0  0  0  0  0  0999 V2000
    0.0000   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.6188    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -0.6188    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.6188    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
M  END
>  <Notes>
Applicants can include text
$$$$
```

Figure 1. Example text contents of an SD File

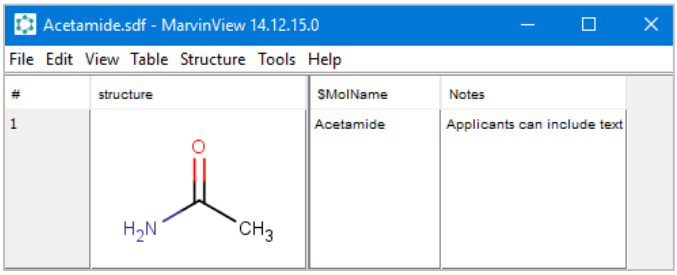


Figure 2. Chemical software can interpret an SD File and translate the contained content into a graphical chemical structure and data table.

Guidelines for creating an SD File for submissions

- Please use V2000 format.
- The file name should include the application type and number (i.e., MF 012345, ANDA 012345) and name of the Active Pharmaceutical Ingredient (API).
- Include other columns/data items linked to each chemical structure.
    - Name(s) of the chemical as referenced in the application. This data item may be called *NAME.* Unique common names or names found in monographs may be used.
    - Application type and number
        - The data item may be called *APPLICATION_NUMBER.*
        - Include the application type prefix (e.g., MF, ANDA, NDA, IND, or BLA) before the number (e.g., MF-012345).
    - Unique identifier for cross-references
        - The data item may be called *ID.*
        - Include an identifier string which is unique across the SD File and can be used for referencing structures in the file and other documents (e.g., Structure 1, Structure 2, etc.).
    - Role of each chemical
        - The data item may be called *ROLE.*
        - The contents could be, for example: active ingredient, process impurity, intermediate, degradant, metabolite, and/or starting material.
        - Other appropriate descriptions are allowed if necessary.
    - Unique Ingredient Identifiers (UNII), if available
        - Available at https://precision.fda.gov/uniisearch and https://gsrs.ncats.nih.gov/ginas/app/beta/. The data item should be called *UNII.*
    - CAS number, if available
        - Available from SciFinder - https://scifinder.cas.org. The data item should be called *CAS.*
    - Notes, if needed
        - The data item may be called *NOTES*.
        - Notes specify/qualify something about the substance that is not easily discernable from the structure alone.
        - For example, specify that the chemical substance has axial chirality and is the "*R*" stereoisomer.

Place SD File(s) in appropriate section of eCTD

- Create one SD file, if possible and depending on needs of submission, and include chemical structures of the following:
    - Drug substances
    - Drug products
    - Starting materials
    - Intermediates
    - Impurities
    - Leachables exceeding the analytical evaluation threshold (AET) (calculated from the safety concern threshold (SCT))
- Within Module 3 there are several sections which may contain information relating to drug substance (3.2.S) or drug product (3.2.P). Place SD File(s) in appropriate sections to aid in an efficient review.
- **Impurities evaluated by (Q)SAR to support selection of a surrogate molecule or predict potential mutagenicity should be placed in Module 3 sections 3.2.S.3.2 for drug substance and 3.2.P.5.5 for drug products.**
- Other SD Files may go in, but are not limited to:
    - Drug Substance
        - 3.2.S.2.6 Manufacturing process development
        - 3.2.S.3.2 Impurities
        - 3.2.S.6 Container closure system
    - Drug Product
        - 3.2.P.2.3 Manufacturing process development
        - 3.2.P.2.4 Container closure system
        - 3.2.P.3.3 Description of manufacturing process and process controls
        - 3.2.P.5 Control of drug product
        - 3.2.P.5.5 Impurities
        - 3.2.P.7 Container closure system
        - 3.2.P.8 Stability
    - **3.2.R Supportive files placed here for any of the sections above should be properly hyperlinked.** One SD File may be suitable when substances pertain to multiple sections.
- Nonclinical study reports submitted in Module 4 may include hyperlink(s) to SD Files submitted in Module 3 to assist cross-references.

Software to create an SD File

Freely available and/or commercial software can be used to help you create an SD File. The following are example software tools that can aid in creating or editing an SD File. Note that this list is not exhaustive and many other tools may be used to create and edit.

- Open Babel
- DataWarrior
- KNIME SDF Writer
- ChemAxon Marvin
- Microsoft Excel with ChemDraw Add-in from ChemOffice
- Instem Leadscope SDF Editor
- Molecular Operating Environment
- RDKit
- CDK

Sample instructions to create an SD File using DataWarrior or Instem Leadscope SDF Editor

DataWarrior Example

The following provides sample steps for creating an SD File using the DataWarrior software tool.

1. Open DataWarrior Software.
2. Click File →New.
3. Choose “Column Type” of “structure” and click “Add Column”.
4. For each other data item (e.g., ID, NAME, ROLE, APPLICATION_NUMBER, UNII, etc.), select “Column Type” of “Text”, enter the name of the column, and click “Add Column”
5. Once all columns are defined, click “OK”.
6. A blank form with the data items will be displayed. Under the “structure” section, double-click the structure box to draw a chemical structure. Alternatively, you can copy a known SMILES string or molfile contents (sometimes called “MOL Text”) to the clipboard and paste it in the structure box by selecting “Paste Structure or Name” from the context (right-click) menu.
7. Populate each relevant data item by entering the values. You may add new data item columns by clicking Data->Add Empty Columns.
8. When finished with one structure entry, click “New Row” to begin entering the next record. Repeat steps 6-8 until all structures are entered.
9. Go to File → Save Special… → SD-File
    - File name: SampleSDF_DMF_012345_API_Name.sdf, click “Save”
    - Structure Column: Structure
    - SD-file version: **Version 2**
    - Atom Coordinates: 2D
    - Compound name column: <automatic>.
10. Click “OK”.

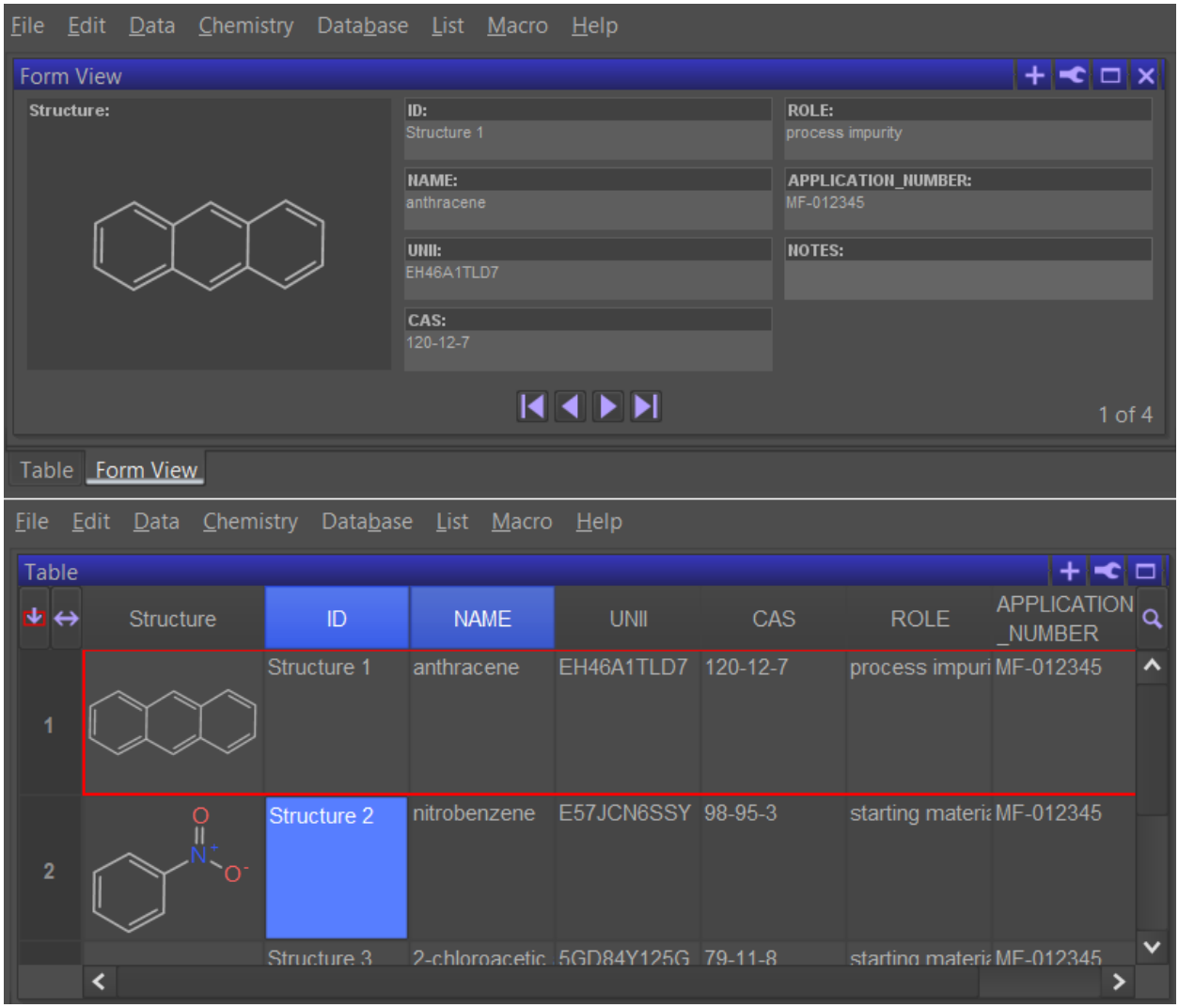


Figure 3. Screenshot of chemical structures and associated data in an SD File in DATAWarrior

Leadscope Example

The following provides sample steps for creating an SD file using Instem Leadscope SDF Editor v1.0.3-2.

1. Open Leadscope SDF Editor software.
2. Go to File → Draw New Structure.
3. In the structure editor window, draw the structure. At the bottom, there is a text field for the Name. Enter the Name of the substance here. Click Ok button on bottom right-hand side. Structure and Name will appear in the SD file table.
4. Repeat step 3 for additional chemicals.
5. Go to Edit → Add property → Type name of property (e.g., NAME). Repeat for additional columns.
6. To add data, double-click a cell and type associated data. Press Enter.
7. Go to File → Save As →
    - File name: SampleSDF_DMF_012345_API_Name.sdf
    - Files of type: SD File (.sdf).
8. Click Save.

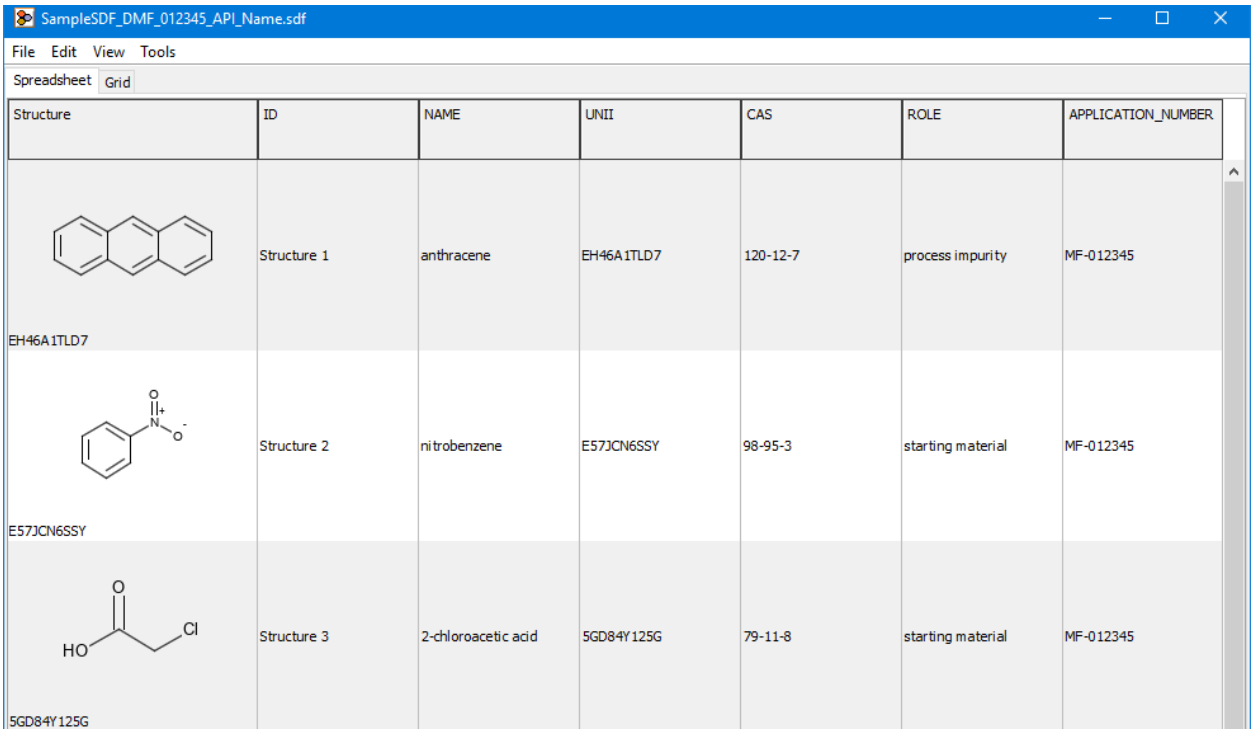
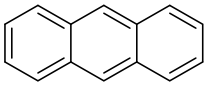
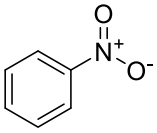
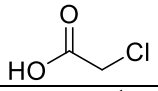
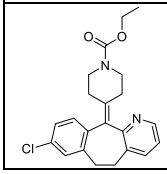


Figure 4. Screenshot of drug substances and associated data in an SD file in Leadscope

Sample SD File with columns/data items and associated data when viewed using an SD File viewer.

An example of the raw ASCII text of the SD File is available in the Appendix.

Structure	ID	NAME	UNII	CAS	ROLE	APPLICATION_NUMBER
	Structure 1	anthracene	EH46A1TLD7	120-12-7	process impurity	MF-012345
	Structure 2	nitrobenzene	E57JCN6SSY	98-95-3	starting material	MF-012345
	Structure 3	2-chloroacetic acid	5GD84Y125G	79-11-8	starting material	MF-012345
	Structure 4	loratadine	7AJO3BO7QN	79794-75-5	active ingredient	MF-012345

Recommendations for special chemical structure scenarios

One SD file containing more than one chemical structure will be expected. For special chemical structure scenarios, such as but not limited to, salts, mixtures, 3-D configurations, isotopes, or stereochemistry, please use the guidelines below. For complex substances other than small molecules (e.g., proteins, nucleic acids) please contact FDA-SRS@fda.hhs.gov for a UNII.

- Do
  - **DO** Include only one structure per (MOL) record.
  - **DO** Include stoichiometry explicitly if known for salts and solvates.
  - **DO** For substances that have a UNII and/or CAS registry number but have no appropriate single simple chemical structure, the structure component of the record can be left blank while the UNII and/or CAS data item are filled.
- Do Not
  - **Do Not** Include query attributes for atoms or bonds – e.g., do not include X's, star atoms, etc. The double-either bond is allowed when necessary.
  - **Do Not** Include "Sgroups" (e.g., monomer brackets, SRU designations brackets, data brackets, etc.) within the molfile. The one exception is that "multiple group" Sgroup bracket types (i.e., "MUL" groups) can be used to represent stoichiometry of salts and hydrates.
  - **Do Not** Include "super atoms" where shorthand notations like "NO2" and "Ph" are used to specify functional groups. Every atom should be specified in isolation without aliases or groupings.
- Isotopes - Include specific isotopes only if structure is isotopically labelled or isotopically enriched. The "M ISO" attribute of the molfile should be used to encode this information when needed.
- Stereochemistry
  - Use stereo bonds (wedge and dash bonds) to indicate stereochemistry. Do not use other annotations/embedded text properties in the molfile to indicate stereochemistry information.
  - Include all known stereochemistry and indicate where stereochemistry is present. If stereochemistry is unknown, explicitly indicate in the "NOTES" SD File data item that the stereochemistry is unknown or a racemic mixture.
  - If double bond geometry is known and relevant, it should be depicted by appropriate atom coordinates. The implicit E or Z configuration of the depicted 2D coordinates will be assumed to be intentional.
  - If double bond geometry is unknown or is a mixture, the double-either bond should be used OR the atoms should be arranged in a colinear fashion which does not imply E/Z designation.
  - Stereo bonds should point to chiral atoms.
  - Use explicit hydrogens to specify stereochemistry on rings.
- Configurations
  - Depict explicit Kekulé forms of molecules using single and double bonds instead of using aromatic bond types.
  - Where possible, draw the prevalent tautomeric form of the compound.
  - 2D representations are both adequate and preferred to 3D representations.
- Mixtures, salts, or coordination complexes
  - Impurities and metabolites should be represented as free acids or bases. Do not represent as salts in impurities and metabolites.
  - Drug substances that are truly salts (e.g., Na^+, H^+, $NH4^+$, Cl^-) can be represented as salts. Amine salts other than quaternary amines should be

represented uncharged as hydrochlorides, hydrobromides etc. Metal salts can be charged but should always be charge balanced.
    - Where possible, each structure record should contain only one covalently connected entity. Multiple covalent entities within one record are allowed for the case of salts, hydrates, coordination complexes and inorganic substances.
    - A mixture of organic substances should be presented in separate records.
- Carbohydrates, steroids, cyclic structures
    - Use planar drawings for cyclic structures and ring systems. See GSRS cholesterol and tetracycline example.
    - Use stereo bonds (wedge and dash bonds) to indicate stereochemistry instead of Fischer, Haworth, and chair projections. See GSRS D-galactose example.
- Proteins and peptides
    - For complex proteins, peptides, and substances with unnatural amino acids that require additional data elements to describe or are difficult to represent with a single chemical structure, please contact FDA-SRS@fda.hhs.gov to obtain a UNII for the substance. See GSRS zendusortide and zendusortide structure example.
    - **Do Not** include collapsed amino-acid notations (single letter or 3 letter) in the structure. Instead have the amino acids expanded without groupings.
    - **Do Not** include aliases for atoms (e.g., “Me” for methyl).

Additional support

- FDA’s GSRS enables an efficient and accurate exchange of information on substances through their UNII, which can be generated prior to submissions and at any time in the regulatory life cycle.
    - To search for a substance by name, use FDA's GSRS UNII Search Service https://precision.fda.gov/uniisearch.
    - To search for a substance by chemical structure, use the public NIH GSRS database https://gsrs.ncats.nih.gov/ginas/app/beta/.
    - To obtain a UNII for a substance, please contact FDA-SRS@fda.hhs.gov.
    - For complex substances that require additional data elements to describe or are difficult to represent with a single chemical structure, please contact FDA-SRS@fda.hhs.gov to obtain a UNII for the substance.
- For general questions on formatting or submitting SD Files, please contact FDA-SRS@fda.hhs.gov.
- For trouble submitting SD Files related to DMF applications, please contact DMFOGD@fda.hhs.gov.
- If you have questions for CDER related to eCTD submissions, please contact the CDER Electronic Submission (ESUB) Support Team at esub@fda.hhs.gov.
- If you have questions for CBER related to eCTD submissions, please contact the CBER ESUB Support Team at esubprep@fda.hhs.gov.

References

1. SDfiles (multiple structures and optional data). CTFILE FORMATS *BIOVIA* Databases 2016. *Dassault* Systèmes © 2015. http://help.accelrysonline.com/ulm/onelab/1.0/content/ulm_pdfs/direct/reference/ctfileformats2016.pdf (accessed May 12, 2021), page 71.

Appendix: Sample ASCII text of an SD File

```
EH46A1TLD7


 14 16  0  0  0  0  0  0  0  0999 V2000
   -2.1434    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.1434   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4289    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7145    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289   -0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.1434   -0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.1434    0.4125    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.4289    0.8250    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  2  3  1  0  0  0  0
  3  4  2  0  0  0  0
  4  5  1  0  0  0  0
  5  6  2  0  0  0  0
  6  1  1  0  0  0  0
  4  7  1  0  0  0  0
  7  8  2  0  0  0  0
  8  9  1  0  0  0  0
  9 10  2  0  0  0  0
 10  5  1  0  0  0  0
  8 11  1  0  0  0  0
 11 12  2  0  0  0  0
 12 13  1  0  0  0  0
 13 14  2  0  0  0  0
 14  9  1  0  0  0  0
M  END
>  <ID>
Structure 1

>  <NAME>
anthracene

>  <UNII>
EH46A1TLD7

>  <CAS>
120-12-7

>  <ROLE>
process impurity

>  <APPLICATION_NUMBER>
MF-012345

$$$$
E57JCN6SSY


  9  9  0  0  0  0  0  0  0  0999 V2000
   -1.4283    0.0007    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.4283   -0.8243    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7139   -1.2368    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0006   -0.8243    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0006    0.0007    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7139    0.4132    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7151    0.4132    0.0000 N   0  3  0  0  0  0  0  0  0  0  0  0
    0.7151    1.2368    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    1.4283    0.0014    0.0000 O   0  5  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  2  3  1  0  0  0  0
```

```
  3  4  2  0  0  0  0
  4  5  1  0  0  0  0
  5  6  2  0  0  0  0
  6  1  1  0  0  0  0
  5  7  1  0  0  0  0
  7  8  2  0  0  0  0
  7  9  1  0  0  0  0
M  CHG  2   7   1   9  -1
M  END
>  <ID>
Structure 2

>  <NAME>
nitrobenzene

>  <UNII>
E57JCN6SSY

>  <CAS>
98-95-3

>  <ROLE>
starting material

>  <APPLICATION_NUMBER>
MF-012345

$$$$
5GD84Y125G


  5  4  0  0  0  0  0  0  0  0999 V2000
   -0.3572   -0.2062    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.3572    0.6188    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -1.0717   -0.6188    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.3572   -0.6188    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.0717   -0.2062    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
  1  2  2  0  0  0  0
  1  3  1  0  0  0  0
  1  4  1  0  0  0  0
  4  5  1  0  0  0  0
M  END
>  <ID>
Structure 3

>  <NAME>
2-chloroacetic acid

>  <UNII>
5GD84Y125G

>  <CAS>
79-11-8

>  <ROLE>
starting material

>  <APPLICATION_NUMBER>
MF-012345

$$$$
7AJO3BO7QN


 27 30  0  0  0  0  0  0  0  0999 V2000
    2.5981   10.3100    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2990    9.5600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.2990    8.0600    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    7.3100    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.2990    8.0600    0.0000 O   0  0  0  0  0  0  0  0  0  0  0  0
   -0.0000    5.8100    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
```

```
    1.3000    5.0600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.3000    3.5600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    2.8100    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3000    3.5600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3000    5.0600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.0000    1.3100    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.3500    0.6600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    2.4500    1.6800    0.0000 N   0  0  0  0  0  0  0  0  0  0  0  0
    3.8800    1.2400    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    4.2200   -0.2200    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    3.1200   -1.2400    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    1.6900   -0.8000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
    0.7500   -1.9700    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -0.7500   -1.9700    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.6900   -0.8000    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -1.3500    0.6600    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -2.4500    1.6800    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -3.8800    1.2400    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -4.2200   -0.2200    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
   -5.6543   -0.6592    0.0000 Cl  0  0  0  0  0  0  0  0  0  0  0  0
   -3.1200   -1.2400    0.0000 C   0  0  0  0  0  0  0  0  0  0  0  0
  1  2  1  0  0  0  0
  2  3  1  0  0  0  0
  3  4  1  0  0  0  0
  4  5  2  0  0  0  0
  4  6  1  0  0  0  0
  6  7  1  0  0  0  0
  7  8  1  0  0  0  0
  8  9  1  0  0  0  0
  9 10  1  0  0  0  0
 10 11  1  0  0  0  0
  6 11  1  0  0  0  0
  9 12  2  0  0  0  0
 12 13  1  0  0  0  0
 13 14  2  0  0  0  0
 14 15  1  0  0  0  0
 15 16  2  0  0  0  0
 16 17  1  0  0  0  0
 17 18  2  0  0  0  0
 13 18  1  0  0  0  0
 18 19  1  0  0  0  0
 19 20  1  0  0  0  0
 20 21  1  0  0  0  0
 21 22  2  0  0  0  0
 12 22  1  0  0  0  0
 22 23  1  0  0  0  0
 23 24  2  0  0  0  0
 24 25  1  0  0  0  0
 25 26  1  0  0  0  0
 25 27  2  0  0  0  0
 21 27  1  0  0  0  0
M  END
>  <ID>
Structure 4

>  <NAME>
loratadine

>  <UNII>
7AJO3BO7QN

>  <CAS>
79794-75-5

>  <ROLE>
active ingredient

>  <APPLICATION_NUMBER>
MF-012345

$$$$
```