

[← CDD Vault](#) [Knowledgebase](#) [Community](#) [Requests](#) [New Request](#)[Sign in](#)[CDD Support](#) > [Getting Started](#) > [Training Guide](#)

2: Importing the first compound library

**Anna Spektor**

Today at 11:40

[Follow](#)

In this lesson we jump right into importing data. We will register a compound manually, then import a compound library from a file. We'll learn about handling of chemical structures and associated molecular information. We will learn how to prepare properly formatted import files in comma-delimited CSV or in SDF formats, and the import process itself. Many of CDDs' key elements introduced here in the context of compound libraries are also applicable to other types of data, and will reappear in subsequent lessons.

[Molecules and chemical registration](#)

[Molecule Definition](#)[Chemical Registration](#)[Hands-on Example](#)[Bulk import from file](#)[Import step 1: Choose data file](#)[Import step 2: Mapping import fields](#)[Import step 3: Validation Report](#)[Where are my compounds?](#)

Recently viewed articles

[Adding, Removing and Editing Vault Users](#)[Table of Contents with Hyperlinks](#)[Limitations](#)[Token-based Authentication](#)[Security and Access Control](#)

Related articles

[3: Importing Single-Point Screening Data](#)[Bulk registration of molecules from file](#)[Setting up a Dose Response Protocol](#)[Table of Contents with Hyperlinks](#)[Molecule Validation Rules during Registration](#)

CDD is primarily a structure-activity relationship (SAR) database, designed for optimal storage of small molecules and associated biological results, with much emphasis on chemical intelligence surrounding the structures, like substructure and similarity searches and chemical property prediction. Therefore, all data in the database is linked to, and pivoted around **molecules**. CDD provides extensive structure and name validation of newly imported libraries, so it is important to supply the best structure representations available, including stereochemistry. In some cases when unambiguous structures are not available, CDD is flexible as well. Let's look at how to create a molecule record, and then how to import an entire library from a file.

Molecules and chemical registration

Molecule Definition

A CDD molecule record typically represents the structure of a small drug-like compound that is being tested for pharmacological activity in screening assays. CDD provides pre-defined attributes, calculations, and tools that are important in this kind of research. However, a molecule record can store any other kind of testable entity, for which you wish to record data: such as fragments, peptides, antibodies, DNA fragments, or even patients.

As compounds are purchased or synthesized for testing, they need to be registered in the vault as molecule and batch records. There are two different registration modes available. By default, your CDD vault will come with a formal chemical registration system. A more flexible informal registration system is also available on request, and we will briefly go over it at the end.

Example

Let's manually register Niacin:

5g of Niacin (SMILES: OC(=O)C1=CC=CN=C1), purchased from Sigma-Aldrich. Sigma sent the following accompanying information: Sigma catalog number: N4126-5G, Sigma Lot number: 2936.29.1520, CAS Number: 59-67-6, purity >= 98% by HPLC.

In order to capture all of this information in CDD vault, we need to define a field that will store each piece of information. This is done using chemical registration, which is described below.

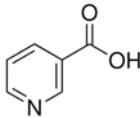
Chemical registration

The CDD compound registration system automatically generates unique molecule and batch IDs. Salts are automatically stripped from the core structure, and a unique numeric identifier is assigned to both the core structure and each new batch. Salt information is stored as a batch field, and the formula weight is calculated automatically based on the core structure and salt stoichiometry. Thus every unique core structure is assigned a unique molecule name, and there may be multiple batches with different salt forms for each molecule.

CDD also supports "structureless registration", which is useful when handling mixtures, natural products, or other cases where a defined structure is not possible. Note, that although importing compounds can be done in bulk, *updating structures at a later time must be done manually to ensure data integrity.*

A new compound record consists of two core parts: the **structure** of the compound (including the salt) and **batch-related attributes**. A brand new CDD vault will have a set of batch attributes already defined. These may be used as is, or configured specifically for your needs on the [Settings-> Vault-> Registration Rules](#) page.

Here is a list of all pre-defined molecule and batch attributes that can be populated, and how it can be applied to the Niacin example above. The list is sufficient to capture everything except the Sigma Lot number and CAS numbers. We need to create new fields for them.

Parameter	How supplied	Description	Example/Notes
Molecule Name	Auto-generated by CDD	Unique name and primary identifier of the molecule. This is generated based on the registration sequence.	DEMO-0000001 : for this example, prefix "DEMO" is used, and Niacin is the first compound in the vault.
Synonyms	optional	<i>Unique</i> name, and secondary identifier of the molecule. Synonyms are checked for uniqueness just like molecule names. This field can have multiple values.	We suggest to use this field only if the identifier is truly unique, such as common-, trade- or IUPAC names. E.g. "Niacin" is unique, an identifier such ABC-12345 might not be.
Structure	Required for defined molecules, but structureless registration is supported	May be supplied in any of the following formats: SMILES, MOL files, ChemAxon extended SMILES, IUPAC names, or drawn using CDDs' structure editor. A salt may be included in the structure. If no structure is supplied by user, none of the chemical properties can be calculated.	 <p>This is a structure image. You can draw it in the structure editor, or paste as a SMILES string: <chem>OC(=O)C1=CC=CN=C1</chem></p>
Batch Name	auto-generated	Batch names are auto-generated as a 3 digit number. Alternate batch naming scheme is possible prefix-based on salt form of the compound.	001 : Salt data are stored in a separate batch field.

automatica

Batch Salt	lly stripped from structure or entered as a separate field	Salt and solvate stripping occurs automatically, unless this information is imported via a separate text-based salt field. See this article for in-depth compound registration.	No salt, free base or acid: if there is no salt information present.
Formula Weight	auto-generated	The formula weight is automatically calculated based on the stoichiometry of core structure, salt and solvent.	123.109 g/mol
Date	optional, but we highly recommend you make it required.	Synthesis or purchase date. (An admin may rename the field appropriately.)	today's date: the default date when registering a compound. You can change this to reflect the right date.
Person	optional, but we highly recommend you make it required	Responsible chemist or another responsible contact. (An admin may rename the field appropriately.)	Your Name: your name is entered automatically when you register a new compound. You can change this to reflect the right owner.
Vendor	optional	Vendor name, if this is a purchased compound. (An admin may rename the field appropriately.)	Sigma-Aldrich
Note	optional	Any additional batch-specific information, such as initial amount, purity, or appearance comments. This field can have multiple values. (An admin may rename the field appropriately.)	purity >= 98% by HPLC; cat. no N4126-5G any information can be stored here, e.g. additional vendor batch information

Registration rules

All of the batch fields listed above can be managed on the *Settings->Vault->Registration* rules page by the vault administrator. This is an advanced feature that is described in detail [here](#). Let's add a new field now. Navigate to this page, and click "Add/edit batch fields". At the bottom of the list click "Add a batch field", and fill in the form as follows, then move the new field to the top of the list using the up/down icon, and click "Update batch fields". The new "External ID" field is now added to every new batch record, and we can use it to store the Sigma-Aldrich lot number in our niacin example.

Of course you may add multiple fields and name them as required, e.g. instead of "External Identifier" you might want to name it "Catalog Number" and a second field "Lot Number".

Name	Data Type	Must be Unique	This field	Move up/down icon
External Identifier	Text	<input checked="" type="checkbox"/>	is optional	⬆

User-defined fields

Field	Value
CAS#	59-67-6
CAT#	N4126-5G

User-defined fields allow you to customize a molecule record if you wish to store information about the molecule that does not fit into the pre-defined fields above. The first time you enter a user-defined field, you define both the field **label** and the field **value** (afterwards, you only enter values). In the Niacin CAS number example, the label is "CAS#" and the value is "59-67-6". Similarly, Sigma's catalog number could become a UDF instead of a synonym: label is "CAT#" and value is "N4126-5G". When you think about your vault configuration, consider how you wish to classify different attributes: does it make more sense simply as a note, a separate field, or will your future searching and sorting needs be better met if you to store it as a UDF.

Hands on example

Moving from theory to practice, register Niacin in CDD on your own using the table above. If you haven't logged into your CDD sandbox, do so now. Your sandbox vault will have a registration sequence based on the name of the vault.

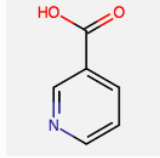
Directly from the **Explore Data** page, click on **Create a new** in the side-bar, and choose **Molecule**.

Fill in the initial form with the structure and all batch information, as shown in the table above. Structure can be supplied as SMILES, MOL files, ChemAxon extended SMILES (CXSMILES) or IUPAC names. We'll use the structure editor here, but feel free to try out entering the IUPAC name (pyridine-3-carboxylic acid) or SMILES string OC(=O)C1=CC=CN=C1

Click on "Launch the structure Editor" box, and let's draw Niacin.

If you are familiar with chemical structure editors such as ChemDraw, Marvin JS editor will look very similar. Use the structure template on the bottom, elements on the right, and bonds on the left.

When you are finished drawing, click "Use this structure" to send it to CDD's molecule form. Notice that now a structure preview is shown in the form, as well as the SMILES structure representation. Choose the correct project and click "Create Molecule".



Molecule Definition

Structure: niacin
Mrv1583109131614562D
9 9 0 0 0 0 999 V2000
0.0000 2.4750 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
0.7145 2.0625 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0
1.4289 2.4750 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0
0.7145 1.2375 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
1.4289 -0.0000 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 0.0000 0.0000 N 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 2 1 0 0 0 0
2 3 2 0 0 0 0
2 4 1 0 0 0 0
4 5 2 0 0 0 0
5 6 1 0 0 0 0
6 7 2 0 0 0 0
7 8 1 0 0 0 0
8 9 2 0 0 0 0
4 9 1 0 0 0 0
M END
|
SMILES, MOL or IUPAC

Batch Information

External Identifier:

Date:

Person:

Place:

Vendor:

Note:

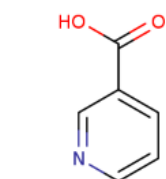
Project:

[Create Molecule](#) or [cancel](#)

As the core molecule record is created, you are taken to the Molecule Details page. This is where additional attributes can be added and/or edited. The page has an **Overview** tab, **Batches**, **Plates**, **Protocols** and **Files**. The record also has a left-hand **meta-data panel**, with a structure preview, link-outs to related molecules on CDD and [Chemspider](#), project meta-data and management, and the "Delete" button.

ASRD-0000628

Vault:
AS Registration with
Duplicates



[CDD-480](#) (See more data sources)

[Find molecules with this structure](#)

[View ChemSpider page](#)

[Add to a collection](#)

[Add a batch](#)

[Manage project access](#)

[Delete this molecule](#)

Showing data from 1 of 1 project

Owner: [Terry Gilliam](#)

Created: November 18, 2011

Updated: September 13, 2016

Overview Batches 1 Plates 0 Protocols 0 Collecti... 0 Projects 1

Files 1

Definition [Edit definition and structure](#)

Name: ASRD-0000628

Synonyms: Niacin and N4126-5G

SMILES · CXSMILES · InChI · InChIKey · IUPAC

Structure:

User-defined Fields [Edit user-defined fields](#)

CAS#: 59-67-6

Lipinski Properties ?

Molecular weight: 123.109 g/mol

log P: -0.17

H-bond donors: 1

H-bond acceptors: 3

Lipinski Rule of 5: Satisfied
4 of 4 within
desirable range

Additional Properties ?

Formula: C₆H₅NO₂

pK_a: 2.79 (Acidic)

Exact mass: 123.0320 g/mol

Heavy atom count: 9

Composition: C (58.54%),
H (4.09%),
N (11.38%),
O (25.99%)

On the **Overview** tab, **Definition** section, click on [Edit definition and structure](#) to add additional identifiers, e.g. the Sigma catalog number (though, again, in general you want to be certain about its uniqueness). Click [Add a synonym](#), and paste in the number: **N4126-5G**. Don't forget to "Save changes". Now niacin has two unique names by which it may be identified for either look-up during a search or as the molecule identifier when importing assay data. Notice the available structure formats once that the molecule is registered: **SMILES**, **CXSMILES**, **InChI**, **InChIKey**, **IUPAC**. You can export the structure in any of these formats, as well as MOL /SDF format.

While you're on the Overview tab, let's also add CAS number (CAS# 59-67-6) to the **User-defined Fields** section: click on [Edit user-defined fields](#), and then [Add a user-defined field](#). Then fill in both the field and value input boxes, and don't forget to save.

The last section to point out on the Overview tab are two panels of calculated properties: Lipinski's on the left, and additional physicochemical to the right. These are calculated using ChemAxon's chemistry engine, and are based on the supplied structure. You can learn more about these properties by clicking on the icons. Here's what your Niacin record could look like. There is one associated batch, 0 plates to which this molecule is assigned, and 0 tested protocols. There are also no files attached to this record.

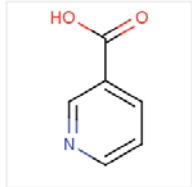
Alternate Compound Registration.

CDD also offers a flexible non-registration mode, where the user determines the naming convention. Batch tracking is not required, and salt stripping is not performed. The main advantage of this option is its flexibility, permitting structureless molecules, and variable naming rules for different subsets of structures (Note: structureless molecules are permitted in registration mode. This will be described in a later lesson). The user should be aware that this system will require more quality control (QC) upfront, before compounds are registered, since this system is more permissive than the default registration mode. Here is what niacin would look like if registered in this fashion:

- The molecule is simply named "Niacin", with no registration ID automatically assigned. You can choose to assign a numeric ID, but this will need to be generated by you.
- A batch is not automatically registered with the core molecule, but may be added after the core compound is registered. There are no required fields or naming conventions, so diligence is required from the user to maintain consistent data.
- There is no Salt field or formula weight, since salt stripping is not performed. Every new salt form of niacin will become a new molecule record. It will be up to you to strip salts, and store relevant information in the Batch Notes section.

Niacin

Vault:
Anna Spektor Sandbox



CDD-480 (See more data sources)

Find molecules with this structure
View ChemSpider page

Currently viewing data from 1 project (see which one)

Overview Batches 0 Plates 0 Protocols 0 Files 1

Definition [Edit definition and structure](#)

Name: Niacin

Synonyms: (no synonyms)

Description:

Structure: SMILES CXSMILES InChI InChIKey IUPAC

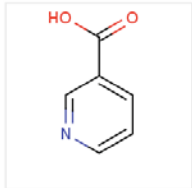
OC(=O)c1cccnc1

User-defined Fields [Edit user-defined fields](#)

CAS#: 59-67-6

Niacin

Vault:
Anna Spektor Sandbox



CDD-480 (See more data sources)

Find molecules with this structure
View ChemSpider page

Overview Batches 1 Plates 0 Protocols 0 Files 1

Batch [Add a Batch](#)

Batch	Details	Actions
Name: 2936.29.1520	Person:	Edit Delete
Date: 2011-10-14	Vendor: Sigma-Aldrich	
	Place:	
Notes: purity is >= 98% by HPLC; Initial amount= 5g		

Informal registration must be activated at the vault level by a CDD admin, so you will need to [send a request](#) to us, asking to switch it on. This is an important choice to make early in the evaluation, because the mode of registration should not be changed after compound libraries are imported. In most cases formal registration will be suitable for both academic groups and drug discovery companies.

Bulk import from file

File formatting

Now that you know how to create a single molecule, we move to the next step of importing compounds in bulk from a properly formatted text file. CDD accepts two file formats for import: [Comma Separated Values](#) format, or CSV, and [Structure Data File](#) format, or SDF. A third format, .zip, is intended for import of multiple files at once, e.g. in the case of associated picture, pdf or other data to which a csv (or sdf) file would have

a field with the corresponding file name included in that .zip file.

SDF files are usually supplied by compound vendors when purchasing a library, and can be generated by most standard cheminformatics tools. These files will be properly formatted by default and do not require any formatting.

We recommend in general using sdf format especially when importing structure containing stereochemistry information sdf handles this in general better and more specific than SMILES.

CSV files can be easily generated from Excel, and need to be formatted in a CDD-readable manner.

Here are the rules for creating a correct CSV file.

	A	B	C	D	E	F	G	H	I	J	K
1	Synonym	SMILES	Solvent (UDF)	Plate Name	Well	Plate Locatio	Batch Extern	Batch Date	Batch Person	Batch Vendo	Warning (Batch N
2	triclosan	Oc1cc(C)ccc	Ethanol	Targeting FA A02		Freezer 1A	KCG-09-001	8/17/09	R. Stuart	ChemX Inc.	This batch came v
3	3,4,5-trichlor	C1c1c(OC)c(C	Ethanol	Targeting FA B02		Freezer 1A	KCG-09-002	8/17/09	R. Stuart	ChemX Inc.	This batch came v
4	Pentachloror	C1c1c(OC)c(C	Ethanol	Targeting FA C02		Freezer 1A	KCG-09-003	8/17/09	R. Stuart	ChemX Inc.	This batch came v
5	SN-8435	COc1cc(O)cc	Ethanol	Targeting FA D02		Freezer 1A	KCG-09-004	8/17/09	R. Stuart	ChemX Inc.	This batch came v
6	SN-6981	Nc1ccc(OC)cc	Ethanol	Targeting FA E02		Freezer 1A	KCG-09-005	8/17/09	R. Stuart	ChemX Inc.	This batch came v
7	CDD-54581	C1c1cc(OC)c(C	Ethanol	Targeting FA F02		Freezer 1A	KCG-09-006	8/17/09	R. Stuart	ChemX Inc.	This batch came v
8	SN-14610	Oc1cc(C)cc(C	Ethanol	Targeting FA G02		Freezer 1A	KCG-09-007	8/17/09	R. Stuart	ChemX Inc.	This batch came v
9	SN-472	Oc1c(C)cc(C	Ethanol	Targeting FA H02		Freezer 1A	KCG-09-008	8/17/09	R. Stuart	ChemX Inc.	This batch came v
10	SN-6412	Oc1c(C)cc(C	Ethanol	Targeting FA A03		Freezer 1A	KCG-09-009	8/17/09	R. Stuart	ChemX Inc.	This batch came v
11	SN-13959	OCc1cc(C)cc	Ethanol	Targeting FA B03		Freezer 1A	KCG-09-010	8/17/09	R. Stuart	ChemX Inc.	This batch came v
12	SN-14418	Oc1ccc(C)cc	Ethanol	Targeting FA C03		Freezer 1A	KCG-09-011	8/17/09	R. Stuart	ChemX Inc.	This batch came v
13	SN-9924	Oc1c(C)cc(C	Ethanol	Targeting FA D03		Freezer 1A	KCG-09-012	8/17/09	R. Stuart	ChemX Inc.	This batch came v
14	2-[4-(2,4-dicl	CC(O)C1CC(O	Ethanol	Targeting FA E03		Freezer 1A	KCG-09-013	8/17/09	R. Stuart	ChemX Inc.	This batch came v
15	4-(4-Chlorop	Nc1ccc(OC)cc	Ethanol	Targeting FA F03		Freezer 1A	KCG-09-014	8/17/09	R. Stuart	ChemX Inc.	This batch came v
16	2-(2,4-dichio	NCCOC1CC(C	Ethanol	Targeting FA G03		Freezer 1A	KCG-09-015	8/17/09	R. Stuart	ChemX Inc.	This batch came v
17	SN-14417	Oc1c(C)ccc(C	Ethanol	Targeting FA H03		Freezer 1A	KCG-09-016	8/17/09	R. Stuart	ChemX Inc.	This batch came v
18	SN-12060	Nc1cc(C)cc(O	Ethanol	Targeting FA A04		Freezer 1A	KCG-09-017	8/17/09	R. Stuart	ChemX Inc.	This batch came v
19	SN-11479	COCCOCCOC	Ethanol	Targeting FA B04		Freezer 1A	KCG-09-018	8/17/09	R. Stuart	ChemX Inc.	This batch came v
20	SN-115166	CCC(C)OC1cc	Ethanol	Targeting FA C04		Freezer 1A	KCG-09-019	8/17/09	R. Stuart	ChemX Inc.	This batch came v
21	SN-15167	CC(C)COC1cc	Ethanol	Targeting FA D04		Freezer 1A	KCG-09-020	8/17/09	R. Stuart	ChemX Inc.	This batch came v
22	SN-15177	COCCOCCOC	Ethanol	Targeting FA E04		Freezer 1A	KCG-09-021	8/17/09	R. Stuart	ChemX Inc.	This batch came v

- Each type of data must be entered into a separate column.

For example, if chemist name and vendor name are in the same column then separate them out.

- The very first row of the file must contain column headers/labels. There may not be multiple header rows. CDD will use the headers to link the data in the column to the appropriate fields in the database.

Columns with a blank first row will be ignored.

The headers do not need to match CDD fields exactly, as you will specifically link/map your file to vault fields during import.

Please observe, if your file happens to have two rows of headers, where the second row has sub-headers, these must be combined into a single row.

- Each row of data must include a molecule and batch identifier.

Recall that all data in CDD is pivoted around molecules, and we will not know how to associate the data unless a molecule ID and batch ID are supplied.

For example, if you are importing batch attributes, you need to include both the molecules' name and batch name.

A molecule may be uniquely identified by a combination of its' primary ID and batch name, or synonym and batch name, or plate and well reference, once the plate map is imported. If you have a unique batch name field defined, you can use this instead of the molecule name and batch name combination.

- Molecule structure must be in [MOL](#), [IUPAC](#) or [SMILES](#) format. Name to structure conversion based on common names and CAS numbers is possible but is not guaranteed.

SMILES is usually safest for CSV files, since it is not dependent on formatting such as line breaks (like MOL), and is not prone to errors (like IUPAC).

- For replicate data include each replicate on a separate row.

For example, if you are importing 3 batches of a single compound, include one row for each replicate, repeating the same molecular identifier on each row.

Import Step 1: Choose data file (and project)

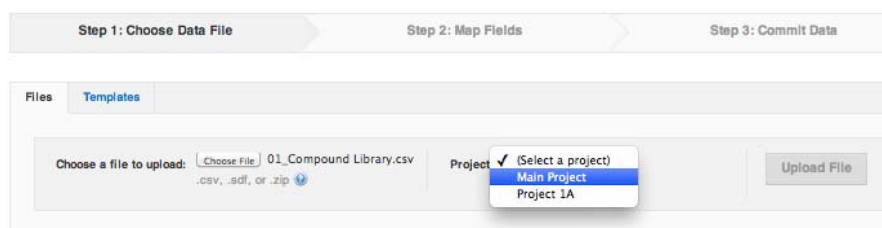
In the this section we will practice importing a file containing a compound library. You can download a

sample CSV file attached to this lesson, or use your own - prepared in Excel according to above rules, or a vendor SDF file. At the very minimum, your file should contain structures and required batch fields. Files are imported, not surprisingly, on the **"Import Data"** tab. The process follows a three step wizard starting a file upload and project selection, followed by a "mapping" step, where we instruct the database how to parse the file. Finally the wizard has us finish with a quality control and commit step.

The first step of the import wizard is to choose the file to upload from your computer, and to pick a project in CDD Vault where this file is going.

All data in a CDD Vault must be assigned to at least one project, so if you only have one project available, it will be automatically selected. If there are multiple projects, you must choose the project in step 1 of the import, otherwise CDD Vault will not let you move to the next step. The chosen file and it's contents will become associated with this project.

We will complete our example with a library file attached to this lesson, called [Lesson2_CMPD_Library.csv](#). Whichever file you are using for practice, download it to your computer first. When a file and a project are selected, the **"Upload file"** button become available - click it to proceed to the next step.



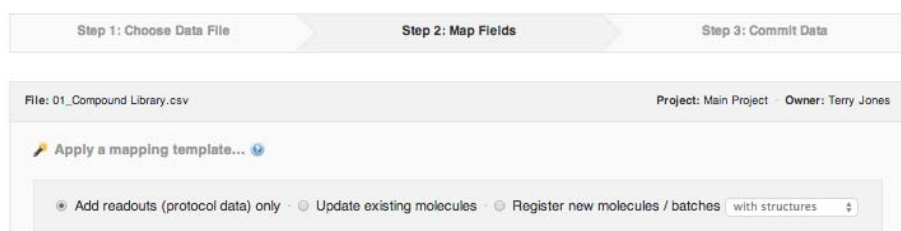
Import Step 2: Map fields

This is the main part of the import process, where you instruct the database how to parse your file, by "mapping" columns (in csv format) or fields (in SDF format) from your file to database fields. Since mapping is done explicitly by you, there is no need to match headers exactly between your file and CDD. Once you have completed a mapping, it may be saved for future use as a template. If you will use the same import file format for many libraries, this template will eliminate the need to go through the mapping one column at a time.

Choose the import setting that applies to your data:

- **Add readouts (protocol data) only** - this is the default setting, and it applies any time you are importing assay (protocol) results. It is assumed that all compounds referenced in the import have been previously registered in your CDD Vault, otherwise you will receive errors. You may register plates in this mode (for previously registered compounds).
- **Update existing molecules** - use this setting for most flexible, since you can both update existing molecules with new attributes (e.g. synonyms, batch fields, plates), and also import assay or protocol data. It is still assumed that all compounds referenced in the import have been previously registered in your CDD Vault, otherwise you will receive errors.
- **Register new molecules/batches (with structure/without structure)** - use this setting when registering new compounds, or when registering new batches of existing compounds. It is assumed that the molecules and batches found in the file do not exist, and thus will be created on import. Should a compound already exist in the vault, a new batch will be created, but not without giving you a chance to review this before committing to the import (see later, [Import Step 3](#)).

Please note, that should you choose this setting for a dose-response data import file, you will inadvertently generate as many new batches, as you have serial dilutions.



Since we're registering a library of compounds, make sure to select **Register new molecules / batches**

with structures without structures setting.

The rest of the mapping interface consists of your file preview, with a horizontal scroll-bar above to move across different columns and fields. The first column is automatically selected and highlighted in blue. Following this blue arrow down the page, you see a panel with CDD field selections. These are the fields in the CDD database available for mapping. The panel is divided into sections: **Molecule fields**, **Batch fields**, **Plate and Well fields**, **Protocol fields**, and lastly a **Do not import** option. CDD will try to help expedite the process by trying to guess the mapping of a given column according to its header. Most often the system will guess correctly, but if no best guess is available, the pre-selected field will be "Do not import". You can always override the best guess by choosing a different field. In case of the sample file, the best guess mapping for the first column is correct: the column called "Molecule Name" should be mapped to the field "Molecule Name or Synonym" in the database. Notice that you have additional options below the "Best Guess": you can choose to add a prefix to the molecule names, and to assign data in this column as a primary molecule names. In this example, we will not need any prefix. When you are satisfied with the mapping, click "Apply".

	A	B	C	D	E	F
1	Synonym	SMILES	Solvent (UDF)	Plate Name	Well	Plate Location
2	trilosan	<chem>Oc1cc(C)...(C)cc1C</chem>	Ethanol	Targeting... Pathway	A02	Freezer 1A
3	3,4,5-tri...xyphenol	<chem>C1c1c(OC)...(C)c1C1</chem>	Ethanol	Targeting... Pathway	B02	Freezer 1A
4	Pentachlo...ybenzene	<chem>C1c1c(OC)...(C)c1C1</chem>	Ethanol	Targeting... Pathway	C02	Freezer 1A

Molecule Name or Synonym

Best guess: SMILES is Structure

You can map structures in the following formats: SMILES, Molfiles, ChemAxon extended SMILES and IUPAC names. The chiral flag will automatically be set to absolute stereochemistry, but enhanced stereochemistry features will be preserved.

The core structure will be neutralized. Salts and hydrates will automatically be stripped and stored in newly registered batches. If you map the Molecule Salt field, all components of the structure will be registered as a mixture. You can view and download a list of all available salts.

Apply

Once the mapping is applied, the blank status in the file preview changes to the green mapped status, and the blue highlighting jumps to the next column. A best guess is done on the second column, correctly surmising that column "SMILES" contains SMILES representation of compound structures, and should be mapped to the **Structure** field under **Molecule fields**. We can just click "Apply", and continue to the next column.

	A	B	C	D	E	F
1	Synonym	SMILES	Solvent (UDF)	Plate Name	Well	Plate Location
2	trilosan	<chem>Oc1cc(C)...(C)cc1C1</chem>	Ethanol	Targeting... Pathway	A02	Freezer 1A
3	3,4,5-tri...xyphenol	<chem>C1c1c(OC)...(C)c1C1</chem>	Ethanol	Targeting... Pathway	B02	Freezer 1A
4	Pentachlo...ybenzene	<chem>C1c1c(OC)...(C)c1C1</chem>	Ethanol	Targeting... Pathway	C02	Freezer 1A

Molecule Name or Synonym **Structure** **Solvent (UDF)** **Plate Name** **Well Location**

Plate Location will not be imported

Next

The next column is "Solvent". We will want to associate solvent information with the molecule record, and will store it in a **User-defined field** under **Molecule fields**. When creating a new field for the first time, choose "Create new field" option from the drop-down, otherwise choose an existing field. Columns "Plate" and "Well" will also be mapped correctly under **Plate and Well fields**, so all we need to do, is click "Apply" each time. Here's what the mapping should look like now: columns A through E are mapped, with the corresponding database field is shown in the bottom row of the preview, column F "Plate Location" is highlighted with the "Do not import" option selected. CDD can not provide a best guess for this column

mapping, but we can map it manually by assigning this column to **Batch Fields**, **Batch Note** (if they are defined, else you can create as above mentioned manually, or simply skip them).

	F	G	H	I	J	K
1	Location	Batch External ID	Batch Date	Batch Person	Batch Vendor	Warning (Batch Note)
2	er 1A	KCG-09-001	8/17/09	R. Stuart	ChemX Inc.	This batc... it last
3	er 1A	KCG-09-002	8/17/09	R. Stuart	ChemX Inc.	This batc... it last
4	er 1A	KCG-09-003	8/17/09	R. Stuart	ChemX Inc.	This batc... it last

Warning (Batch Note) is mapped to Batch Note

Next

Columns G through K contain batch information, and will be guessed by the wizard. Do not forget to click "Apply" for each column. The last column called "Warning (Batch Note)" will be manually mapped to **Batch Fields**, **Batch Note**. The final mapping preview should look something like the screen-shot on the left.

You can use the horizontal scroll-bar at the top to review your mappings. If everything looks correct, you are ready for the next steps. You can Save this mapping as a template..., and "Process File".

Import Step 3: Commit Data

The final step of the import process requires the user to **commit** or **reject** the import after reviewing a validation report. At this stage, no data has been written to the database, and the page displays a report with a summary of import events to be finalized.

The summary is divided into sections of **Noteworthy events**, **Suspicious events** and **Errors**. The report will include a preview of the problematic data and suggest possible causes of the problem in the section descriptions. Each section may be selectively accepted or rejected. You can also download any section of the report by clicking the Download link at the bottom of each section. Data review is an important step, learn more about it in [this article](#).

File: Lesson2_CMPD_Library.csv (Review mapping) Project: CDD Tutorial Owner: Demo User

This data import is ready for review
 80 records will be imported Only records that are not associated with any rejected event will be imported.

Noteworthy Events - Usually fine. Associated records will be imported except if you choose otherwise.

- 79 New Molecules [ACCEPT] [REJECT]
- 1 New Batch for New Molecule [ACCEPT] [REJECT]
- 1 New Plate [ACCEPT] [REJECT]

[Commit Data Import] [Reject Data Import]

Importing our example file, your report should look like this, if everything went well: 79 new molecules, one new batch for a new molecule (this file has the same structure in two rows, so two batches are created). In this case, it's the expected behavior, we are expecting two batches for one compound, and we will accept this section. Lastly, one new plate is being registered. Click on the sections to see expanded explanations and previews. When ready, click **Commit Data Import**. Compounds are now being registered in your vault. You will see a progress notification on top of the report. For large import, you can click on the "notify by email" link; this allows you to do other things in vault (or otherwise). Once the import is finished you will

receive an email notification!

Where are my compounds?

A data import is completed when the "commit" step is finished, and the validation report is updated. You can access your newly registered compounds directly from the validation report by clicking on the magnifying glass "Explore Data" link. You can also download a file with new compound IDs. The primary view of the data, however, is from the main Explore Data tab. Here you can browse all the compounds in your projects, and perform queries on any molecule-related information, including structures. Learn more about reporting and searching data in [this section](#) of the knowledgebase.

Navigate to the **Molecules** sub-tab on **Explore Data**. If you don't see any structures, make sure you have selected your project in the side-bar on the left of the screen. Once the project check box is clicked, you should see 79 structure records (you will end up on this page directly if you clicked the "Explore data" link mentioned above). Compounds are always displayed in reverse chronological order, so you see the most recently imported compounds first. Click on the first compound - it should be compound 80. The form should already be familiar to you - we have seen it when registering niacin.

The screenshot shows the CDD Support: RegistrationTutorial interface. At the top, there's a navigation bar with 'Dashboard', 'Explore Data', 'Import Data', and 'Manage Vault'. The 'Explore Data' tab is active. Below the navigation bar, there's a search bar and a 'Create a new molecule' button. The main content area displays a list of 80 molecules, with the first few visible. Each molecule entry includes a chemical structure, its CDD ID, and synonyms. For example, the first molecule is ASCDDPR-0000080 with CDD-187834 and synonyms: SMR000137200. The second molecule is ASCDDPR-0000078 with CDD-343992 and synonyms: 4-Hydroxy-3,5-dimethyl-5-prop-2-ynyl-2H-thiophan-2-one. The third molecule is ASCDDPR-0000078 with CDD-341020 and synonyms: 4-Hydroxy-5-... The fourth molecule is ASCDDPR-0000077 with CDD-341020 and synonyms: 5-(E)-But-2-...

Go back to view all molecules and search for "354132". You will find that this is a synonym for compound #68. You will also see that this compound has another ID associated to it: CDD-341044. Any compound with a CDD ID, is registered in CDD's public domain, in addition to being a private compound in your vault. Obviously no one but you knows of the existence of this molecule in your vault. You are able to see public data registered by someone else, but not vice versa - anyone viewing public compounds won't see who else might have them or not. You can find out exactly what other data is available for this compound when you open the detailed record, and follow the links under the CDD number.

Summary

In this lesson we learned how to register molecules in a project. We manually registered Niacin using MarvinSketch structure editor, and then imported an entire plate of compounds from file. We learned about all the required fields during compound registration, how to use the import wizard, and looked at the import validation report.

By looking at the compound registration example, we have learned about data import in general, since all the basic steps will be the same regardless of the imported data type. We will now move on to setting up the vault for biological data. Structure-activity relationships cannot be established until the newly imported chemical structures are annotated with biological assay results.

Next Lesson- [Setting up the first screening assay](#)

[Lesson 2_ Importing the first compound library _ Support Portal.pdf](#) (900 KB)

[Lesson2_CMPD_Library.csv](#) (50 KB)

Was this article helpful?



1 out of 1 found this helpful



Have more questions? [Submit a request](#)

0 Comments

Please [sign in](#) to leave a comment.