# CSP Data Guidance for CSD-Theory

## Q2 2022 On-Site WebCSD

**Conditions of Use**

The Cambridge Structural Database Portfolio (CSD Portfolio) including, but not limited to, the following: ConQuest, CSD-Editor, Decifer, Mercury, Mogul, IsoStar, CSD Conformer Generator, Hermes, GOLD, SuperStar, the CSD Python API, web accessible CSD tools and services, WebCSD, CSD sketchers, CSD data files, CSD data updates, the CSD database, sub-files derived from the foregoing data files, documentation and command procedures, test versions of any existing or new program, code, tool, data files, sub-files, documentation or command procedures which may be available from time to time (each individually a Component) encompasses database and copyright works belonging to the Cambridge Crystallographic Data Centre (CCDC) and its licensors and all rights are protected.

Any use of a Component of the CSD Portfolio, is permitted solely in accordance with a valid Licence of Access Agreement or Products Licence and Support Agreement and all Components included are proprietary. When a Component is supplied independently of the CSD Portfolio its use is subject to the conditions of the separate licence. All persons accessing the CSD Portfolio or its Components should make themselves aware of the conditions contained in the Licence of Access Agreement or Products Licence and Support Agreement or the relevant licence.

In particular:

- The CSD Portfolio and its Components are licensed subject to a time limit for use by a specified organisation at a specified location.
- The CSD Portfolio and its Components are to be treated as confidential and may NOT be disclosed or re-distributed in any form, in whole or in part, to any third party.
- Software or data derived from or developed using the CSD Portfolio may not be distributed without prior written approval of the CCDC. Such prior approval is also needed for joint projects between academic and for-profit organisations involving use of the CSD Portfolio.
- The CSD Portfolio and its Components may be used for scientific research, including the design of novel compounds. Results may be published in the scientific literature, but each such publication must include an appropriate citation as indicated in the Schedule to the Licence of Access Agreement or Products Licence and Support Agreement and on the CCDC website.
- No representations, warranties, or liabilities are expressed or implied in the supply of the CSD Portfolio or its Components by CCDC, its servants or agents, except where such exclusion or limitation is prohibited, void or unenforceable under governing law.

Licences may be obtained from:

Cambridge Crystallographic Data Centre
12 Union Road
Cambridge CB2 1EZ, United Kingdom

Web: http://www.ccdc.cam.ac.uk
Telephone: +44-1223-336408
Email: admin@ccdc.cam.ac.uk

# Contents

# 1 Overview

This document provides guidance on how CSD-Theory reads, and translates into software, any input CSP CIF data on the basis of the evolving CSP CIF Dictionary. The CSP CIF Dictionary itself is an on-going effort supported by the CCDC in order to bring much greater consistency, and clarity, in how CSP data and metadata should be handled within the CIF format.

As of May 2021, the CSP CIF Dictionary reached version 0.8, and this version has been publicly shared alongside the seventh CSP Blind Test (2020 to 2022) to guide submissions in CIF format. Finalising a working version 1.0 of the CSP CIF dictionary will likely require further coordination with software vendors, users, and the IUCr Committee for the Maintenance of the CIF Standard (COMCIFS). For reference, information on CIF syntax in general is available from the CCDC website: A short guide to Crystallographic Information Files.

# 2 Importing Data into CSD-Theory

Unlike proprietary experimental databases, please note that you do not need to use the CSD-Editor software to curate predicted structures into a database for use with CSD-Theory. There is instead a dedicated CSP landscape importing Python script available alongside the CSD-Theory API. This should be visible in the utilities area of the CSD-Theory API, alongside scripts for deleting a CSP landscape from the CSD-Theory platform, plus utility scripts to cross-reference your predicted data with your experimental data, amongst other things.

# 3 CSP CIF Data Versions and CSD-Theory

As the CSP CIF Dictionary has been evolving in line with the CSP Data Standards conversation worldwide, we have been updating how our software reads CSP CIFs. Here we provide a simple overview of which versions of our software support different versions of the CSP CIF Dictionary at a high level. Please note that the latest CSD-Theory software does not provide support for older versions of the CSP CIF Dictionary, but once we reach v1.0 of the CSP CIF Dictionary we will ensure that we can maintain backwards compatibility with the CSP CIF format.

| CSP CIF Dictionary version | CSD-Theory software compatibility |
|---|---|
| v0.1 | CSD-Theory API: versions 2021.1, 2021.2 & 2022.1 |
| v0.8 | CSD-Theory API: version 2022.2<br>CSD Landscape Generator: versions 2021.3, 2022.1 |

The information provided below details which specific data fields have been historically used by CSD-Theory, are in active use now, and how these translate into the software.

- Prior to July 2022 (i.e., CSD-Theory API versions 2021.1 up to 2022.1), CSD-Theory used an early version (v0.1) of the CSP CIF dictionary for CSP metadata field mapping. The key fields are given in the table below in the first column.
- From July 2022 onwards (i.e., CSD-Theory API version 2022.2), CSD-Theory will use the v0.8 CSP CIF dictionary fields in order to read and display CSP metadata. Please note that at the point of launch (2021.3 CSD Release), the CSD Landscape Generator component in Mercury and the CSD Python API has also generated CIFs in the v0.8 CSP CIF dictionary format. The key fields interpreted by CSD-Theory are given in the table below in the second column.
- The Type column indicates what format is expected for the contents of that field, either “numb” for a number, or “char” for a string of characters.
- A description of the meaning of each field is provided in the Description column.
- The CSD-Theory field column indicates which specific field in the CSD-Theory software (shown under Computational Details in the interface) is populated from the input CSP CIF data.

CSD-Theory field: Temperature (K)

- Description: The temperature at which the structure is calculated.
- Type: numb
- v0.1 CSP CIF field mapping: _ccdc_csp_simulation_temperature_value
- v0.8 CSP CIF field mapping: _ccdc_csp_simulation_temperature

CSD-Theory field: Density (CCDC)

- Description: Density value calculated from the predicted crystal cell and contents.
- Type: numb
- v0.1 CSP CIF field mapping: _ccdc_csp_density_calc

- v0.8 CSP CIF field mapping: _ccdc_csp_density_calc (no change)

CSD-Theory field: Relative energy (kJ/mol)

- Description: The relative lattice energy of the structure with respect to the global minimum on the lattice or absolute energy landscape.
- Type: char
- v0.1 CSP CIF field mapping: _ccdc_csp_classification_energy_relative
- v0.8 CSP CIF field mapping: _ccdc_csp_classification_energy_lattice_relative

CSD-Theory field: Energy model

- Description: Method used for structure optimisation. Ideally this should be one of the examples given but other values are permitted.
- Type: char
- v0.1 CSP CIF field mapping: _ccdc_csp_optimisation_energy_model
- v0.8 CSP CIF field mapping: _ccdc_csp_structure_optimisation_method

CSD-Theory field: Energy model details

- Description: Additional high-level description of the energy model used for structure optimisation.
- Type: char
- v0.1 CSP CIF field mapping: _ccdc_csp_optimisation_energy_model_definition
- v0.8 CSP CIF field mapping: Any of the following fields:
    - _ccdc_csp_structure_optimisation_force_field_description
    - _ccdc_csp_structure_optimisation_semi_empirical_method
    - _ccdc_csp_structure_optimisation_DFT_approximation
    - _ccdc_csp_structure_optimisation_wavefunction_electronic_method

CSD-Theory field: Relative free energy (kJ/mol)

- Description: The relative free energy, at a given temperature, of the structure with respect to the corresponding global minimum on the same free energy landscape.
- Type: numb
- v0.1 CSP CIF field mapping: _ccdc_csp_free_energy_relative
- v0.8 CSP CIF field mapping: _ccdc_csp_free_energy_relative (no change)

CSD-Theory field: Free energy ranking method

- Description: Method used for free energy correction. Ideally this should be one of the examples given but other values are permitted.
- Type: char
- v0.1 CSP CIF field mapping: _ccdc_csp_free_energy_method
- v0.8 CSP CIF field mapping: _ccdc_csp_free_energy_correction_method

These data fields listed below can also be read by CSD-Theory, though these would typically be calculated by the CSD-Theory software at the point of ingestion, so do not need to be supplied at the point of ingestion. To calculate the similarities at the point of ingestion, use the '--similarity' argument when running the ingestion process.

CSD-Theory field: CSDx match similarity

- Description: Ratio of the number of molecules that match in an overlay cluster.
- Type: numb
- v0.1 CSP CIF field mapping: _ccdc_csp_exptl_matching_similarity
- v0.8 CSP CIF field mapping: _ccdc_csp_exptl_matching_similarity (no change)

CSD-Theory field: CSDx match similarity RMSD

- Description: RMSD of atomic positions within an overlay cluster.
- Type: numb
- v0.1 CSP CIF field mapping: _ccdc_csp_exptl_matching_similiarity_rmsd
- v0.8 CSP CIF field mapping: _ccdc_csp_exptl_matching_similiarity_rmsd (no change)

CSD-Theory field: CSDx match

- Description: The CSD database code of an experimental structure that matches this form of a CSP structure. If the structure matches an entry in an internal database then the ID or internal refcode of the entry should be used.
- Type: char
- v0.1 CSP CIF field mapping: _ccdc_csp_exptl_database_code_csd
- v0.8 CSP CIF field mapping: _ccdc_csp_exptl_database_code_csd (no change)

# 4 Example CIF Data

Included below is a selection of v0.8 CSP CIF data examples, and what will be shown in the CSD-Theory software as a result.

## 4.1 Example 1

v0.8 CSP CIF data loop snippet, highlighting the energy model and energy model details sections:

```
loop_
_ccdc_csp_structure_optimisation_method
_ccdc_csp_structure_optimisation_force_field_name
_ccdc_csp_structure_optimisation_force_field_description
_ccdc_csp_structure_optimisation_DFT_approximation
_ccdc_csp_structure_optimisation_DFT_functional
_ccdc_csp_structure_optimisation_DFT_basis_set
_ccdc_csp_structure_optimisation_software
_ccdc_csp_structure_optimisation_software_version
_ccdc_csp_structure_optimisation_stage
'Force field' FIT 'exp-6 atom-atom potential, CHELPG atomic
charges' . . . CrystalOptimizer 1.6 P
DFT . . GGA PBE 6-31G(d,p) CrystalOptimizer 1.6 P
DFT . . Hybrid PBE0 6-31G(d,p) CrystalOptimizer 1.6 F
```

What will be shown in the CSD-Theory interface, note that the details shown will relate to the final step in the loop, as indicated by the “F” value of the _ccdc_csp_structure_optimisation_stage field:

- Energy Model: “DFT”
- Energy Model Details: “PBE0”

If there are multiple optimisation method fields populated for the final step, then these will be all shown together in a comma delimited string.

## 4.2 Example 2

v0.8 CSP CIF data snippet, specifically output from the CSD Landscape Generator component:

```
_ccdc_csp_density_calc           1.44282
_ccdc_csp_simulation_type        static
_ccdc_csp_simulation_temperature 0.0
_ccdc_csp_structure_optimisation_method 'Force Field'
_ccdc_csp_structure_optimisation_software CSD-Theory
_ccdc_csp_structure_optimisation_software_version 2022.1.0
_ccdc_csp_structure_optimisation_stage F
_ccdc_csp_structure_clustering_method None
_ccdc_csp_classification_energy_lattice_relative 0
_ccdc_csp_classification_energy_lattice_absolute -220.332
```

What will be shown in the CSD-Theory interface, note that in this specific case the CSD-Theory software will recognise the term “CSD-Theory” in the _ccdc_csp_structure_optimisation_software field and report the energy model details as shown despite the lack of _ccdc_csp_structure_optimisation_force_field_description field in the CIF:

- Temperature (K): 0.0
- Density (CCDC): 1.448
- Relative energy (kJ mol-1): 0
- Energy Model: “Force Field”
- Energy Model Details: “Unimol force field customised for CSD structures”

## 4.3 Example 3

v0.8 CSP CIF data highlighting free energy data, 1st CIF:

```
_cell_length_a                   11.5647
_cell_length_b                   16.9375
_cell_length_c                   7.10534
_cell_angle_alpha                90
_cell_angle_beta                 90
_cell_angle_gamma                90
_cell_volume                     1391.77
_ccdc_csp_simulation_temperature 0.0
```

v0.8 CSP CIF data highlighting free energy data, 2nd CIF:

```
_cell_length_a                    11.5647
_cell_length_b                    16.9375
_cell_length_c                    7.10534
_cell_angle_alpha                 90
_cell_angle_beta                  90
_cell_angle_gamma                 90
_cell_volume                      1391.77
_ccdc_csp_simulation_temperature 100.0
```

What will be shown in the interface:

- Each structure will be independently searchable and viewable in CSD-Theory Web.
- In the main CSP Landscape visualisation area, the structures will be viewed by temperature separately in different tabs, with the 0 K data being the default view.
- At the bottom of the CSD-Theory Web interface, the free energy plot will be shown if there are structures at different temperatures present, with data related to each unique structure linked.

Further notes:

- When free energy calculations have been performed, there should be a separate CIF file for the predicted structure at each unique temperature, including the base 0 K structure.
- These structures are linked together in the CSD-Theory software using an internal identifier connecting all finite temperature calculations to the 0 K structure identifier as the “base”.
- The linking between structures calculated at finite temperatures is performed by looking for structures with the same unit cell parameters, but different non-zero temperatures.
- This linking only works when a whole landscape of predicted structures is ingested at once.

# 5 General Notes & Troubleshooting

- CIF syntax states that CIF data names are case-insensitive