CSP Core Dictionary

Introduction

Summary

This is the CSP dictionary for describing predicted crystal structures and the methods, parameters and workflows used to calculate these.

Table of contents

- **0. Datablocks identifiers and file cross-referencing** describing the links between methods, workflows and output structures.
- **1. Input Chemical System** describing the input atoms or molecules for CSP.
- **2. Structure Generation Methods** describing the workflow used to generate theoretical crystal structures.
- **3. Structure Ranking Methods** describing the energy evaluation models used to generate and rank the structures.
- **4. Output Structure Properties** describing the properties of each output structure, such as their energy or density.
- **5. Conventions** specifying guidelines to avoid multiple labels for the same term.
- **6. Future Developments** describing what is missing from the current dictionary.

0. Data blocks identifiers and file cross-referencing

Category `_csp.data_block_[]` : This section specifies the class type of the data block and assigns a unique identifier to it.

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Data Block	`class`	char	Class type of the data block	- "Input" - "Generation Method" - "Ranking Method" - "Workflow" - "Theoretical Structure"		"Input"
CSP	Data Block	`id`	char	Unique identifier of the data block. This will be used to link the different datablocks in a workflow or output structure.	Recommended unique identifiers generation protocol use such as UUID		`dd55207f-9649-435b-9708-c8154c33fc03`
CSP	Data Block	`description`	char	Text identifier of a datablock for human readability.	Free text		"Molecule 1"
CSP	Data Block	`additional_files`	list	If datablocks are specified in different files, add the position of these files.	`List[str]`		`[ "generation_methods.cif" "ranking_methods.cif" "workflows.cif" ]`

Single inputs systems, generation methods and ranking methods must be described in separate datablocks and a unique identifier should be assigned to them. We recommend the use of Universally Unique IDentifiers (UUIDs), described in here and naturally implemented in most programming languages. The Workflow datablock is meant to connect different generation and ranking methods in multistep approaches. Finally, the output theoretical structures will have a link to the specific stage in the workflow and the previous structure. Examples of each one of these data blocks are available below.

Datablocks can be stored in different files or in multiple files depending on the user. In the former case, the `_csp.data_block_additional_files` field should be used. In practice, one can have a set of default settings for the generation, ranking methods and workflows, with the different landscapes that will have only the input and output structures data blocks that differ.

1. Input Chemical System

Category `_csp.input_[]` : This section specifies the atoms used in inorganic CSP or the input molecular entities for molecular crystal generation.

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Input	`name`	char	See name_common and name_systematic from Core CIF dictionary.	Free Text		urea hydrate

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Input	`composition_calculation`	char	"fixed" or "variable" composition calculation.	- Fixed - Variable		Fixed
CSP	Input	`composition_coefficients`	list	List of possible compositions for fixed-composition calculations or extremes for variable-composition simulations.	`List[PositiveInt]`		`[1 1]` `[2 1]`
CSP	Input	`maximum_number_of_components`	numb	The maximum number of components (atoms or molecules) in the unit cell.	1:		4
CSP	Input	`minimum_number_of_components`	numb	The minimum number of components (atoms or molecules) in the unit cell.	1:		2
CSP	Input Atoms	`types`	list	List of atomic species defining the composition.	`List[str]` or `List[PositiveInt]`		`[Mg O]` (atomic symbols), `[12 8]` (Atomic number)
CSP	Input Molecule	`number`	char	Molecule component index.	1:		1
CSP	Input Molecule	`identifier`	char	Label used to identify the molecule.	Free Text		urea, water
CSP	Input Molecule	`smiles`	char	SMILES of the component.	-		C(=O)(N)N
CSP	Input Molecule	`molecule_number`	char	Molecule component index for each atom.	1:		1
CSP	Input Molecule	`molecule_identifier`	char	Label used to identify the molecule for each atom.	Free Text		urea, water
CSP	Input Molecule	`atom_label`	char	Label of atom in the component.	Free Text		C1

Additional details on atoms in molecule ad their connectivity can be specified through the CIF Chemical dictionary, available at:
https://www.iucr.org/__data/iucr/cifdic_html/1/cif_core.dic/index.html

Examples

Inorganic CSP input with fixed stoichiometry:

```
data_ferrosilite
# Datablock Details
_csp.data_block_class               "Input"
_csp.data_block_id                  1ac303d1-ea55-439a-8f3d-d9bd462b3b25
_csp.data_block_description         Ferrosilite

# Input Details
_csp.input_name                     Ferrosilite
_csp.input_atoms_types              [ Fe Si O ]
_csp.input_composition_calculation  fixed
_csp.input_composition_coefficients [ 1 1 3 ]
```

Inorganic CSP input with variable stoichiometry:

```
data_input2
# Datablock Details
_csp.data_block_class                   "Input"
_csp.data_block_id                      d5c894e9-561c-490a-89b0-877bbc516b14
_csp.data_block_description             Hypersthene

# Input Details
_csp.input_name                         Hypersthene
_csp.input_atoms_types                  [ Fe Mg Si O ]
_csp.input_composition_calculation      variable
_csp.input_composition_coefficients     [[1 0 1 3] [0 1 1 3]]
_csp.input_minimum_number_of_components 2
```

```
_csp.input_maximum_number_of_components 10
```

This implies that resulting structures will have formula *x*(FeSiO3)+*y*(MgSiO3) with *2<x+y<10*. Worth noticing that the string following `data_` is not used in the linking of datablocks.

Multi-component molecular crystal CSP with fixed stoichiometry:

```
data_molecule
# Datablock Details
_csp.data_block_class       "Input"
_csp.data_block_id          2a2611e3-2021-4b03-a7c6-0ef71239008f
_csp.data_block_description input1

_csp.input.name             Urea_Hydrate

# Molecules
loop_
    _csp.input_molecule_number
    _csp.input_molecule_identifier
    _csp.input_molecule_smiles
    _chemical.name_common
    1 WAT O      water
    2 URE OCN(N) urea

# Atoms in molecules
loop_
    _csp.input_molecule_molecule_number
    _csp.input_molecule_molecule_identifier
    _csp.input_molecule_atom_label
    _chemical.conn_atom_number
    _chemical.conn_atom_type_symbol
    _chemical.conn_atom_charge
    1 WAT O1 1  O -0.800000
    1 WAT H1 2  H  0.400000
    1 WAT H1 3  H  0.400000
    2 URE O1 4  O -0.613359
    2 URE C1 5  C  0.880229
    2 URE N1 6  N -0.923545
    2 URE N2 7  N -0.923545
    2 URE H1 8  H  0.395055
    2 URE H2 9  H  0.395055
    2 URE H3 10 H  0.395055
    2 URE H4 11 H  0.395055

# Bonds
loop_
    _chemical.conn_bond_atom_1
    _chemical.conn_bond_atom_2
    _chemical.conn_bond_type
    1 2  sing
    1 3  sing
    4 5  doub
    5 6  sing
    5 7  sing
    6 8  sing
    6 9  sing
    7 10 sing
    8 11 sing

_csp.input.composition_calculation  "fixed"
_csp.input.composition_coefficients [ 2 1 ] # Indexes from molecule section (2 water molecules and one urea)
```

`composition_coefficients` here refers to the molecule number. Worthy of note the use of the `Chemical` dictionary in defining the molecules.

Variable stoichiometry search can be specified in the same way as for inorganic systems:

```
...
_csp.input_composition_calculation      "variable"
_csp.input_composition_coefficients     [[1 0] [0 1]]
_csp.input_maximum_number_of_components 4
_csp.input_minimum_number_of_components 2
```

For metal-organic systems, the `input_molecule` and `Chemical` dictionaries can be used specifying metallic atoms as "individual molecules":

```
data_mo
# Datablock Details
_csp.data_block_class       "Input"
_csp.data_block_id          fbbe2b09-da53-4505-ba9c-d4952a096dbb
_csp.data_block_description input1

_csp.input.name "(mi-tricyanomethanide)-silver"
```

```
# Molecules
loop_
    _csp.input_molecule_number
    _csp.input_molecule_identifier
    _chemical.name_common
    1 Metal Silver
    2 c4n3  tricyanomethanide

# Atoms in molecules
loop_
    _csp.input_molecule_molecule_number
    _csp.input_molecule_molecule_identifier
    _csp.input_molecule_atom_label
    _chemical.conn_atom_number
    _chemical.conn_atom_type_symbol
    1 Metal Ag1 1  Ag
    2 c4n3  C1  2  C
    2 c4n3  C2  3  C
    2 c4n3  C3  4  C
    2 c4n3  C4  5  C
    2 c4n3  N1  6  N
    2 c4n3  N2  7  N
    2 c4n3  N3  8  N

# Bonds
loop_
    _chemical.conn_bond_atom_1
    _chemical.conn_bond_atom_2
    _chemical.conn_bond_type
    1 6  sing
    1 7  sing
    1 8  sing
    2 3  doub
    2 4  sing
    2 5  sing
    3 6  doub
    4 7  trip
    5 8  trip


_csp.input_composition_calculation   "fixed"
_csp.input_composition_coefficients  [1 1]
```

2. Structure Generation Methods

This section helps delineate the space search range and specify the parameters used for different methods.

2.1 General Fields

Category `_csp.structure_generation_[]` : Category for structure generation methods.

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Structure Generation	`space_group_number_list`	char/numb/list	Space group selection could be â€œallâ€ or a subset (list) specifying which spacegroups were used.	Either "all" or list of sg numbers		`[14 2 15 61 19 4 33 29 5 1]`
CSP	Structure Generation	`method`	char/list	Structure generation method or list of methods.	- Evolutionary Algorithm (Sec. 2.2) - Particle Swarm Optimisation (Sec. 2.3) - Simulated Annealing (Sec. 2.4) - Monte Carlo Parallel tempering		Simulated Annealing

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
					(Sec. 2.5) - Random Sampling (Sec. 2.6) - Analogue Templates - Other		
CSP	Structure Generation	`stage`	numb	To be used in the "Workflow" datablock, it specify the stage number of the generation method.			
CSP	Structure Generation	`preceding_stage`	numb	To be used in the "Workflow" datablock, it specify the stage number of the previous generation method.			
CSP	Structure Generation	`data_block_id`	char	To be used in the "Workflow" datablock, it specify the identifier of the datablock in which the structure generation method is described.			
CSP	Structure Generation	`data_block_description`	char	To be used in the "Workflow" datablock, it is a short, human-readable description of the generation method.			
CSP	Structure Generation	`software`	char	Name of the software used for structure generation.	Free text		
CSP	Structure Generation	`software_citation`	char	Details of the software used for structure generation. Either URL to webpage or DOI of the related publication.	Free text		
CSP	Structure Generation	`software_version`	char	Version of software used for structure generation.	Free text		

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Structure Generation	`density_lower_limit`	numb	Minimum Cell Density.	>0	kg m^{-3}	800
CSP	Structure Generation	`density_upper_limit`	numb	Maximum Cell Density.	>0	kg m^{-3}	1400
CSP	Structure Generation	`reference_temperature`	numb	Reference temperature for finite-temperature simulations.	>0	K	0
CSP	Structure Generation	`reference_pressure`	numb	Reference pressure for variable cell calculations.	>0	Pa	100000
CSP	Structure Generation	`stopping_criteria`	char/list	List of rules for stopping the generation of new structures.	Free text		"Max Structures", "Low-Energy Structures Unchanged"
CSP	Structure Generation	`stopping_criteria_max_structures_evaluated`	numb	The maximum total number of unique crystal structures that will be generated and evaluated during the search.	>0		10000
CSP	Structure Generation	`stopping_criteria_iterations_without_improvement`	numb	The maximum number of consecutive iterations (generations, MC steps, etc.) where the global minimum (or the lowest few structures) does not change.	>0		50
CSP	Structure Generation	`stopping_criteria_energy_range`	numb	An energy threshold for the selection of low-energy structures to be considered in the convergence criteria.	>0	kJ mol^{-1}	5
CSP	Structure Generation	`stopping_criteria_structures_range`	numb	The number of low-energy structures to be considered in the convergence criteria.	>0		1000

2.2 Evolutionary Algorithms

Category `_csp.evolutionary_algorithm_[]` : Subgroup for CSP Structure Generation methods that use Evolutionary Algorithms. For these fields to be used, the `_csp.structure_generation_method` must include "Evolutionary Algorithm".

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Evolutionary Algorithms	`population_size`	numb	The number of candidate structures in each generation.	>0		100
CSP	Evolutionary Algorithms	`initial_population_size`	numb	The number of candidate structures in the first generation.	>0		50
CSP	Evolutionary Algorithms	`number_of_generations`	numb	The maximum number of evolutionary cycles the algorithm will run before termination (unless other stopping criteria are met).	>0		50
CSP	Evolutionary Algorithms	`nextgen_structure_selection`	numb	The number of individuals that survives in the next generation.	>1		5
CSP	Evolutionary Algorithms	`parents_structure_fraction`	numb	The fraction of individuals in the current population that is used to generate structures in the next cycle.	0-1		0.75
CSP	Evolutionary Algorithms	`mutation_fraction`	numb	The fraction of individuals in the population that will undergo mutation in each generation.	0-1		0.2
CSP	Evolutionary Algorithms	`heredity_fraction`	numb	The fraction of individuals in the population that will be generated through heredity (crossover/recombination) operations between two or more parents.	0-1		0.6
CSP	Evolutionary Algorithms	`permutation_fraction`	numb	The fraction of individuals in the population that will undergo a permutation operation (e.g., swapping atom positions within a structure) in each generation.	0-1		0.1

2.3 Particle Swarm Optimisation Algorithms

Category `_csp.particle_swarm_optimisation_[]` : Subgroup for CSP Structure Generation methods that use Particle Swarm Optimisation. For these fields to be used, the `_csp.structure_generation_method` must include "Particle Swarm Optimisation".

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Particle Swarm Optimisation	`population_size`	numb	The number of candidate crystal structures (particles) in the swarm.	>0		50
CSP	Particle Swarm Optimisation	`number_of_generations`	numb	The maximum number of optimization cycles (generations or iterations) the PSO algorithm will run.	>0		100
CSP	Particle Swarm Optimisation	`inertia_weight`	numb	A parameter controlling the contribution of the previous velocity of the particle to its current velocity.	0-1		0.7
CSP	Particle Swarm Optimisation	`max_inertia_weight`	numb	If the inertia weight changes with each iteration, this parameter specify the maximum value it can have.	0-1		0.9
CSP	Particle Swarm Optimisation	`min_inertia_weight`	numb	If the inertia weight changes with each iteration, this parameter specify the minimum value it can have.	0-1		0.4
CSP	Particle Swarm Optimisation	`cognitive_coefficient`	numb	A parameter (also called self-confidence factor) controlling the influence of the particle's own best position found so far on its movement.	>=0		2
CSP	Particle Swarm Optimisation	`social_coefficient`	numb	A parameter (also called swarm confidence factor) controlling the influence of the swarm's best position found so far on the particle's movement.	>=0		2
CSP	Particle Swarm Optimisation	`velocity_clamp_max`	numb	The maximum allowed velocity for each dimension if velocity clamping is enabled.	>0		0.2

2.4 Simulated Annealing

Category `_csp.simulated_annealing_[]` : Subgroup for CSP Structure Generation methods that use Simulated Annealing. For these fields to be used, the `_csp.structure_generation_method` must include "Simulated Annealing".

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Simulated Annealing	`initial_temperature`	numb	The starting temperature of the simulated annealing process.	>0	K	500
CSP	Simulated Annealing	`cooling_rate`	numb	The parameter that determine how the temperature is decreased over the course of the simulation.	0-1		0.95
CSP	Simulated Annealing	`number_of_steps`	numb	The number of attempted structure generation and acceptance steps performed at each temperature.	>0		10

2.5 Monte Carlo Parallel Tempering

Category `_csp.monte_carlo_parallel_tempering_[]` : Subgroup for CSP Structure Generation methods that use Monte Carlo Parallel tempering. For these fields to be used, the `_csp.structure_generation_method` must be set to "Monte Carlo Parallel Tempering".

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Monte Carlo Parallel Tempering	`number_of_replicas`	numb	The number of independent Monte Carlo simulations (replicas) running in parallel at different temperatures.	1		3
CSP	Monte Carlo Parallel Tempering	`temperatures_list`	list	The list of temperatures at which the replicas are run.	`[T >= 0]`	K	`[0, 300, 600]`
CSP	Monte Carlo Parallel Tempering	`number_of_steps`	numb	The number of Monte Carlo steps performed by each replica at its assigned temperature in each parallel tempering cycle.	>0		100

2.6 Random Search

Category `_csp.random_[]` : Subgroup for CSP Structure Generation methods that use Random, Quasi-random algorithms. For these fields to be used, the `_csp.structure_generation_method` should be set to "Random Sampling".

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Random Search	`random_numbers_algorithm`	char	Specifies the type of random algorithm used.	"Pseudorandom", "Quasirandom", "Other"		"Pseudorandom"
CSP	Random Search	`number_of_samples`	numb	The total number of unique crystal structures to be generated and evaluated during the random search.	>0		5000

Examples

Search in all space groups until 10000 structures are generated with an evolutionary algorithm:

```
data_ea
# Datablock Details
_csp.data_block_class                                          "Generation Method"
_csp.data_block_id                                             af534fed-8153-4af2-bd9f-29b0fef8d805
_csp.data_block_description                                    "ea"

# Method Details
_csp.structure_generation_space_group_number_list              "all"
_csp.structure_generation_method                               "Evolutionary Algorithm"
_csp.structure_generation_density_lower_limit                  750
_csp.structure_generation_density_upper_limit                  1500
_csp.structure_generation_stopping_criteria                    "Max Structures"
_csp.structure_generation_stopping_criteria_max_structures_evaluated 10000
```

Combination of different structure generation methods and search limited on most popular space groups for organic crystals:

```
# Random Search
data_rs
    _csp.data_block_class                                          "Generation Method"
    _csp.data_block_description                                    rs
    _csp.data_block_id                                             6e0147be-0454-44a1-a3bb-de7b326dde1b

    _csp.structure_generation_stopping_criteria                    "Max Structures"
    _csp.structure_generation_stopping_criteria_max_structures_evaluated 1000
    _csp.random_random_numbers_algorithm                           "Quasi-random"

# Simulated Annealing
data_sa
    _csp.data_block_class                                          "Generation Method"
    _csp.data_block_description                                    sa
    _csp.data_block_id                                             95f28b3c-d029-4840-a69a-3ced34219c28

    _csp.structure_generation_stopping_criteria                    "Max Structures"
    _csp.structure_generation_stopping_criteria_max_structures_evaluated 2000000
    _csp.simulated_annealing_initial_temperature                   400
    _csp.simulated_annealing_cooling_rate                          0.95
    _csp.simulated_annealing_number_of_steps                       100

# Relations between methods
data_workflow
    _csp.data_block_class                                          Workflow
    _csp.data_block_description                                    wf
    _csp.data_block_id                                             29ba6f2f-0a56-47be-bc90-9c8adc7760e9

    # General Settings
    _csp.structure_generation_space_group_number_list              [14 2 15 61 19 4 33 29 5 1]
    _csp.structure_generation_density_lower_limit                  750

    # Structure Generetion Methods
    loop_
        _csp.structure_generation_stage
        _csp.structure_generation_preceding_stage
        _csp.structure_generation_data_block_description
        _csp.structure_generation_method
        _csp.structure_generation_data_block_id
        0 . "rs" "Random Sampling"     6e0147be-0454-44a1-a3bb-de7b326dde1b
        1 0 "sa" "Simulated Annealing" 95f28b3c-d029-4840-a69a-3ced34219c28

    # Structure Ranking Methods
    ...
```

For the last section, the mandatory data fields to identify and link the different datablocks are `_csp.structure_generation_stage`, `_csp.structure_generation_preceding_stage` and `_csp.structure_generation_data_block_id`. Other data fields are included to make the file more human-readable.

3. Structure Ranking Methods (High-level)

Within this section, you can define the workflow used to rank the different crystals and give high-level details of the methods used. To allow compatibility with other dictionaries and possible future works on computational chemistry calculations, single methods datafields don't have the `_csp` prefix.

3.1 General Fields

Categories:

- `_csp.structure_ranking_[]` : Category to define structure ranking stages in multistep approaches.
- `_compchem.[]` : Details on calculation types and software citation.
- `_compchem.geometry_optimistaion_[]` : Geometry optimisation datils.

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Structure Ranking	`stage`	numb	To be used in the "Workflow" datablock, it specify the stage number of the ranking method.	>=0		0
CSP	Structure Ranking	`preceding_stage`	numb	To be used in the "Workflow" datablock, it specify the stage number of the previous ranking method.			
CSP	Structure Ranking	`data_block_description`	char	To be used in the "Workflow" datablock, it specify the identifier of the datablock in which the structure ranking method is described.	Free Text		FF, PBE, PBE0

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
CSP	Structure Ranking	`data_block_id`	char	To be used in the "Workflow" datablock, it is a short, human-readable description of the ranking method.			
CompChem	-	`method`	char	The energy or scoring model used to rank structures.	- pDFT (Sec. 3.2) - Forcefield (Sec. 3.3) - Semi-Empirical (Sec. 3.4) - Wavefunction (Sec. 3.5) - ML Potentials (Sec. 3.6) - Other		Forcefield
CompChem	-	`calculation_type`	char	Indicates how atomic positions are changed.	- "Optimisation" - "Dynamic Ensemble" - "Single point"		"Single point"
CompChem	-	`software`	char	Name of the software used for structure ranking.	Free Text		
CompChem	-	`software_citation`	char	Details of the software used for structure ranking.	Free Text		
CompChem	-	`software_version`	char	Version of software used for structure ranking.	Free Text		
CompChem	Geometry Optimisation	`algorithm`	char	Geometry optimisation algorithm.	- BFGS - L-BFGS - Quasi-Newton - FIRE - Steepest Descent - Conjugate Gradient - Other		
CompChem	Geometry Optimisation	`cell`	char	It can be "fixed" for no cell optimisation, "isotropic" or "anisotropic" for cell relaxation calculations.	- fixed - isotropic - anisotropic		
CompChem	Geometry Optimisation	`atoms`	char	It can be "fixed" for no atoms' position optimisation, "all" for all-atoms geometry optimization, "hydrogens" for optimisation of only H atoms, "non-hydrogens" for non-H atoms or a list of atoms for custom relaxation.	- fixed - all - hydrogens - non-hydrogens - `[List of _chemical.conn_atom.number]`		`[1 2 3 4]`
CompChem	Geometry Optimisation	`relax_force_convergence`	numb	Convergence criteria for stopping the geometry optimisation. Present in TCOD as `_dft_atom_relax_force_conv`.	>0	kJ mol^{-1} nm^{-1}	0.1
CompChem	Geometry Optimisation	`max_steps`	numb	Maximum number of steps in the geometry optimisation.	>0		

3.2 Periodic Density Functional Theory

Category `_dft_[]` : Subgroup for CSP Structure Ranking methods that use pDFT methods (the *p* of *pDFT* is removed in `_dft` for consistency with the TCOD Dictionary). For these fields to be used, the `_compchem.method` should be set to "pDFT".

Top Category	Data Field	Type	Definition	Constraints	Units	Example
pDFT	`exchange_correlation_functional_type`	char	Specifies the type of exchange-correlation functional used.	- LDA - GGA - meta-GGA - Hybrid - Other		GGA

Top Category	Data Field	Type	Definition	Constraints	Units	Example
pDFT	`exchange_correlation_functional_name`	char	Specifies the name of exchange-correlation functional used.	- PBE - PBE0 - SCAN - ...		PBE
pDFT	`pseudopotential_type`	char	Defines the type of pseudopotentials used.	- Plane-waves - PAW - Norm-conserving - Ultrasoft		PAW
pDFT	`dispersion_correction`	char	The Van der Waals correction used.	- Grimme-D2 - Grimme-D3 - Tkatchenko-Scheffler - Many-body dipersion - XDM - Other		XDM

3.3 Forcefields

Category `_forcefield_[]` : Subgroup for CSP Structure Ranking methods that use forcefield or mixed inter/intra molecular methods. For these fields to be used, the `_compchem.method` should be set to "Forcefield".

Top Category	Data Field	Type	Definition	Constraints	Units	Example
Forcefield	`name`	char	Name of the force field.	Free Text		OPLS, GAFF
Forcefield	`intramolecular_term`	char	The energy evaluation method for intramolecular interactions.	- "Bonded Parameters" - "Isolated Molecule Energy" - None - Other		
Forcefield	`electrostatic_term`	char	Functional form of electrostatic interactions.	- "Point-Charges" - Multipoles - Other		
Forcefield	`vdw_term`	char	Functional form of van der Waals interactions.	- LJ(C6 -C12) - LJ(epsilon -sigma) - Buckingham - ReaxFF Morse-Potential - 14-7 function - Other		
Forcefield	`parameterization_method_intra`	char	Briefly describes the primary method used to derive the intramolecular force field parameters.	Free Text		"Fitting to gas-phase QM data", "Transferable parameters based on atom types"
Forcefield	`parameterization_method_coulomb`	char	Briefly describes the primary method used to derive the electrostatic force field parameters.	Free Text		"Fitting to gas-phase QM data", "Transferable parameters based on atom types"
Forcefield	`parameterization_method_vdw`	char	Briefly describes the primary method used to derive the dispersion/repulsion force field parameters.	Free Text		"Fitting to gas-phase QM data", "Transferable parameters based on atom types"
Forcefield	`qm_parameterization_functional`	char	The exchange-correlation functional used in the gas-phase quantum mechanical	- "MP2" - "CCSD(T)"		MP2

Top Category	Data Field	Type	Definition	Constraints	Units	Example
			calculations when fitting force field parameters.	- "B3LYP" - ...		
Forcefield	`qm_parameterization_basis_set`	char	The basis set used in the gas-phase quantum mechanical calculations when fitting force field parameters.	- "aug-cc-pVTZ" - "6-31G(d,p)" - ...		6-31G

3.4 Semi-Empirical

Category `_semiempirical_[]` : Subgroup for CSP Structure Ranking methods that use Semi-Empirical methods. For these fields to be used, the `_compchem.method` should be set to "Semi-Empirical".

Top Category	Data Field	Type	Definition	Constraints	Units	Example
Semi-Empirical	`type`	char	Type of semi-empirical method.	- Tight-Binding - "Self-Consistent Tight-Bonding" - Other		
Semi-Empirical	`method`	char	Specifies the name of the Semi-Empirical method used.	- AM1 - PM3 - PM6 - xTB - ...		PM6
Semi-Empirical	`electronic_parameters`	char	The Slater-Koster tables or equivalent defining the atomic orbitals and pairwise element-element interactions.	- mio - 3ob - ...		mio
Semi-Empirical	`repulsive_potential`	char	When not included in the SK files, the repulsive potential term used.			
Semi-Empirical	`dispersion_correction`	char	Dispersion corrections for semi-empirical methods.	- D3 - TS - MBD - ...		D3
Semi-Empirical	`hydrogen_bond_correction`	char	H-bond corrections for semi-empirical methods.	- H+ - H4 - ...		H+
Semi-Empirical	`halogen_bond_correction`	char	Correction term for halogen bonds interactions.	- X		

3.5 Wavefunction

Category `_wavefunction_[]` : Subgroup for CSP Structure Ranking methods that use wavefunction methods. For these fields to be used, the `_compchem.method` should be set to "Wavefunction".

Top Category	Data Field	Type	Definition	Constraints	Units	Example
Wavefunction	`exchange_correlation_functional`	char	Specifies the name of functional used.	- HF - MP2 - CC - ...		MP2
Wavefunction	`basis_set_type`	char	Defines the type of basis used.	- GTH - NAO - ...		NAO

3.6 ML Potentials

Category `_ml_potential_[]` : Subgroup for CSP Structure Ranking methods that use machine learning potentials methods. For these fields to be used, the `_compchem.method` should be set to "ML Potentials".

Top Category	Data Field	Type	Definition	Constraints	Units	Example
ML Potential	`method`	char	Specifies the name of the ML Potential used. In case of ML parameterisation of classical forcefields, refer to the Forcefields dictionaries.	- ANI - MACE - ...		
ML Potential	`model`	char	The specific model used to rank structures.	- 2x - OFF24 - ...		
ML Potential	`precision`	char	Float precision in calculations.	- `float32` - `float64`		

3.7 Free Energy

Category `_free_energy_[]` : Subgroup for CSP Structure Ranking methods that use free energy methods.

Top Category	Data Field	Type	Definition	Constraints	Units	Example
Free Energy	`method`	char	Specifies the name of the approach used to calculate free energies.	- HA - QHA - PSCP - EC - ...		QHA
Free Energy	`reference_temperature`	numb/list	The temperature or list of temperatures at which free energies are calculated.	>0 or `List[PositiveFloat]`	K	`[ 100.0, 200.0, 300.0 ]`
Free Energy	`reference_pressure`	numb/list	The pressure or list of pressures at which free energies are calculated.	>0 or `List[PositiveFloat]`	Pa	`[ 100000.0, 200000.0 ]`

Examples

pDFT with hybrid XC functional and additional data fields taken from the TCOD DFT dictionary:

```
# Datablock Details
_csp.data_block_class       "Ranking Method"
_csp.data_block_id          fe97f09e-5c9c-41b2-930e-bd14f3a418a9
_csp.data_block_description pbe_xdm

# pDFT settings
_compchem.calculation_type               "Single-Point"

_dft.exchange_correlation_functional_type GGA
_dft.exchange_correlation_functional_name PBE
_dft.pseudopotential_type                 PAW
_dft.dispersion_correction                XDM

_dft.kinetic_energy_cutoff_wavefunctions 600
_dft.atom_relax_force_conv               0.002
_dft.BZ_integration.method               "Monkhorst-Pack"
_dft.BZ_integration.grid_dens_X          0.5
_dft.BZ_integration.grid_dens_Y          0.5
_dft.BZ_integration.grid_dens_Z          0.5
```

Multiple energy evaluation steps and different workflows can be described in a cif file. After specifying parameters for each ranking method, two possible workflow examples are described:

- **wf1**: Different methods of increasing computational cost are used. Two single-point hybrid functionals are then used on the PBE optimised structures.
- **wf2**: Different methods of increasing computational cost are used. The last step is a free energy calculation.

```
#
# Structure Ranking Methods
#

# General Purpose FF
data_gaff
    # Data blocks details
    _csp.data_block_class       "Ranking Method"
    _csp.data_block_description gaff
    _csp.data_block_id          83f824d3-6d17-4e42-9952-31ed161ef811

    # Forcefield details
    _compchem.calculation_type         "Optimisation"
```

```
    _ff.name                          "GAFF"
    _ff.intramolecular_term           "Bonded-Parameters"
    _ff.electrostatic_term            "Point-Charges"
    _ff.vdw_term                      "LJ(epsilon,sigma)"
    _ff.parameterization_method       "BCC"
    _ff.qm_parameterization_functional "AM1"

    # Geometry Optimisation
    _compchem.geometry_optimisation_algorithm               steep
    _compchem.geometry_optimisation_cell                    fixed
    _compchem.geometry_optimisation_atoms                   all
    _compchem.geometry_optimisation_relax_force_convergence 0.01
    _compchem.geometry_optimisation_max_steps               10000


# Multipoles-based Approach
data_psi_mol
    # Data blocks details
    _csp.data_block_class       "Ranking Method"
    _csp.data_block_description psi_mol
    _csp.data_block_id          d6f196c5-88d9-4ecd-b388-bcd92fd93a05

    # Forcefield details
    _compchem.calculation_type         "Optimisation"
    _ff.name                           "Psi_mol"
    _ff.intramolecular_term            "Isolated Molecule Energy"
    _ff.electrostatic_term             "Multipoles"
    _ff.vdw_term                       "Buckingham"
    _ff.parameterization_method        "GDMA"
    _ff.qm_parameterization_functional "PBE0"
    _ff.qm_parameterization_basis_set  "6-31G(d,p)"

    # Geometry Optimisation
    _compchem.geometry_optimisation_algorithm               CG
    _compchem.geometry_optimisation_cell                    anisotropic
    _compchem.geometry_optimisation_atoms                   all
    _compchem.geometry_optimisation_relax_force_convergence 0.01
    _compchem.geometry_optimisation_max_steps               200


# GGA DFT
data_pbe
    _csp.data_block_class       "Ranking Method"
    _csp.data_block_description pbe
    _csp.data_block_id          17ad684a-2337-4a96-9808-b8b8d3013dc3

    # DFT details
    _compchem.calculation_type                "Optimisation"
    _dft.exchange_correlation_functional_type "GGA"
    _dft.exchange_correlation_functional_name "PBE"

    # Geometry Optimisation
    _compchem.geometry_optimisation_algorithm               FIRE
    _compchem.geometry_optimisation_cell                    anisotropic
    _compchem.geometry_optimisation_atoms                   all
    _compchem.geometry_optimisation_relax_force_convergence 0.01
    _compchem.geometry_optimisation_max_steps               50

# Hybrid DFT
data_pbe0
    # Data blocks details
    _csp.data_block_class       "Ranking Method"
    _csp.data_block_description pbe0
    _csp.data_block_id          a741eea0-d308-436a-916f-31964b86b649

    # DFT details
    _compchem.calculation_type                "Single-Point"
    _dft.exchange_correlation_functional_type "Hybrid"
    _dft.exchange_correlation_functional_name "PBE0"
    _dft.dispersion_correction                "TS"

data_optb88
    # Data blocks details
    _csp.data_block_class       "Ranking Method"
    _csp.data_block_description optb88
    _csp.data_block_id          2ba152e5-4690-4af4-be55-68789b38b166

    # DFT details
    _compchem.calculation_type                "Single-Point"
    _dft.exchange_correlation_functional_type "Hybrid"
    _dft.exchange_correlation_functional_name "optb88"
    _dft.dispersion_correction                "TS"

# Free Energy
data_pbe0_qha
    # Data blocks details
    _csp.data_block_class       "Ranking Method"
    _csp.data_block_description pbe0
    _csp.data_block_id          11d2779e-6396-4c2c-91ff-d62dddaf9cc1
```

```
    # DFT details
    _dft.exchange_correlation_functional_type "Hybrid"
    _dft.exchange_correlation_functional_name "PBE0"
    _dft.dispersion_correction               "MBD"

    # Free energy
    _compchem.calculation_type         "Dynamic Ensemble"
    _free_energy.method                "QHA"
    _free_energy.reference_temperature [100 200 300]
    _free_energy.reference_pressure    100000.0

#
# Workflows
#

data_workflow_1
    _csp.data_block_class       Workflow
    _csp.data_block_description wf1
    _csp.data_block_id          29ba6f2f-0a56-47be-bc90-9c8adc7760e9

    # Structure Generetion Methods
    ...

    # Structure Ranking Methods
    loop_
        _csp.structure_ranking_stage
        _csp.structure_ranking_preceding_stage
        _csp.structure_ranking_data_block_description
        _compchem.calculation_type
        _compchem.geometry_optimisation_atoms
        _compchem.geometry_optimisation_cell
        _compchem.method
        _csp.structure_ranking_data_block_id
        0 . "gaff"    "Optimisation" "all" "anisotropic" "Forcefield" 83f824d3-6d17-4e42-9952-31ed161ef811
        1 0 "psi_mol" "Optimisation" "all" "anisotropic" "Forcefield" d6f196c5-88d9-4ecd-b388-bcd92fd93a05
        2 1 "pbe"     "Optimisation" "all" "anisotropic" "pDFT"       17ad684a-2337-4a96-9808-b8b8d3013dc3
        3 2 "pbe0"    "Single-Point" .     .             "pDFT"       a741eea0-d308-436a-916f-31964b86b649
        4 2 "optb88"  "Single-Point" .     .             "pDFT"       2ba152e5-4690-4af4-be55-68789b38b166

data_workflow_2
    _csp.data_block_class       Workflow
    _csp.data_block_description wf2
    _csp.data_block_id          964fcd15-82e1-4c4d-a7cb-61b0b34c3421

    # Structure Generetion Methods
    ...

    # Structure Ranking Methods
    loop_
        _csp.structure_ranking_stage
        _csp.structure_ranking_preceding_stage
        _csp.structure_ranking_data_block_description
        _compchem.calculation_type
        _compchem.geometry_optimisation_atoms
        _compchem.geometry_optimisation_cell
        _compchem.method
        _csp.structure_ranking_data_block_id
        0 . "gaff"     "Optimisation"     "all" "anisotropic" "Forcefield" 83f824d3-6d17-4e42-9952-31ed161ef811
        1 0 "psi_mol"  "Optimisation"     "all" "anisotropic" "Forcefield" d6f196c5-88d9-4ecd-b388-bcd92fd93a05
        2 1 "pbe0_qha" "Dynamic Ensemble" .     .             "pDFT"       00d2779e-6396-4c2c-91ff-d62dddaf9cc1
```

Also in this example, mandatory data fields in the loop are `_csp.structure_ranking_stage`, `_csp.structure_ranking_preceding_stage` and `_csp.structure_ranking_data_block_id`. Additional data fields are for improving the human-readability of the CIF file.

4. Theoretical Crystal Structure

Describes the structure-specific outputs of CSP methods. Categories:

- `_theoretical_structure.[]`: Properties of the structure.
- `_theoretical_structure.csp_[]`: Stage identifiers in a multistep ranking approach.

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
Theoretical Structure	-	`temperature`	numb	The temperature at which the energy and other properties of the theoretical structure were calculated.	>=0	K	298.15
Theoretical Structure	-	`pressure`	numb	The pressure at which the energy and other properties of the theoretical structure were calculated.	:	Pa	101325.0

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
Theoretical Structure	-	`calculated_density`	numb	The calculated density of the crystal.	>=0	kg m^{-1}	1420.0
Theoretical Structure	-	`total_energy`	numb	The total energy of the theoretical structure, i.e. energy relative to all of the nuclei and electrons seperated to an infinite distance.	:	kJ mol^{-1}	-1500.5
Theoretical Structure	-	`absolute_lattice_energy`	numb	The absolute lattice energy of the crystal, i.e. energy relative to all the molecules seperated to an infinite distance.	:	kJ mol^{-1}	-1600.8
Theoretical Structure	-	`absolute_free_energy`	numb	The absolute free energy of the crystal.	:	kJ mol^{-1}	-1450.2
Theoretical Structure	-	`free_energy_correction`	numb	The correction applied to the lattice energy to obtain the free energy, accounting for vibrational and other thermal effects.	:	kJ mol^{-1}	50.6
Theoretical Structure	-	`relative_lattice_energy`	numb	The lattice energy of the theoretical structure relative to the lowest energy structure found in the CSP.	>=0	kJ mol^{-1}	0.0, 5.2
Theoretical Structure	-	`energy_uncertainty`	numb	An estimate of the uncertainty associated with the calculated energy of the theoretical structure.	>=0	kJ mol^{-1}	0.1
Theoretical Structure	-	`score`	numb	To allow for methods that may rank by criteria other than energies (e.g., based on stability or other desired properties).	:		1, 0.3333, 0.01
Theoretical Structure	-	`rank`	numb	The rank of the structure when ordered by chosen criteria where 1 is considered to be the most favorable or likely structure.	>=1		2, 7, 12
Theoretical Structure	CSP	`reference_temperature`	numb	In case of free-energy calculations, specify the reference temperature.	>=0	K	
Theoretical Structure	CSP	`reference_pressure`	numb	In case of free-energy calculations, specify the reference pressure.	>=0	Pa	
Theoretical Structure	CSP	`input_system_description`	char	Human-readable description of the input system data block.			
Theoretical Structure	CSP	`input_system_id`	char	Identifier of the input system data block.			
Theoretical Structure	CSP	`workflow_description`	char	Human-readable description of the workflow data block.			
Theoretical Structure	CSP	`workflow_id`	char	Identifier of the workflow data block.			
Theoretical Structure	CSP	`generation_stage_description`	char	Human-readable description of the structure generation method data block.			
Theoretical Structure	CSP	`generation_stage_id`	char	Identifier of the structure generation method data block.			
Theoretical Structure	CSP	`ranking_stage_description`	char	Human-readable description of the structure ranking method data block.			
Theoretical Structure	CSP	`ranking_stage_id`	char	Identifier of the structure ranking method data block.			
Theoretical Structure	CSP	`previous_stage_structure_description`	char	Human-readable description of the data block of the structure from			

Top Category	Category	Data Field	Type	Definition	Constraints	Units	Example
				previous stage. This is the structure used as input for the current stage.			
Theoretical Structure	CSP	`previous_stage_structure_id`	char	Identifier of the data block of the structure from previous stage.			

Details on composition, unit cell, symmetry, and atomic coordinates can be specified through the CIF Core dictionary.

Examples

Structure optimised using a multipoles approach:

```
data_structure_1_1
# Structure is thoeretically generated
_exptl.method                                                'theoretical model'

# Datablock Details
_csp.data_block_class                                        "Theoretical Structure"
_csp.data_block_description                                  structure_1_1
_csp.data_block_id                                           00d2779e-6396-4c2c-91ff-d62dddaf9cc1

# Include files specifying input, methods and workflow data blocks
_csp.data_block_additional_files                             [ "csp_input.cif" "workflow.cif" ]

# Stage identifiers
_theoretical_structure.csp_input_system_description          Urea_Hydrate
_theoretical_structure.csp_input_system_id                   2a2611e3-2021-4b03-a7c6-0ef71239008f

_theoretical_structure.csp_workflow_description              wf2
_theoretical_structure.csp_workflow_id                       964fcd15-82e1-4c4d-a7cb-61b0b34c3421

_theoretical_structure.csp_generation_stage_description      ea
_theoretical_structure.csp_generation_stage_id               af534fed-8153-4af2-bd9f-29b0fef8d805

_theoretical_structure.csp_ranking_stage_description         psi_mol
_theoretical_structure.csp_ranking_stage_id                  d6f196c5-88d9-4ecd-b388-bcd92fd93a05

# Properties
_theoretical_structure.temperature             0
_theoretical_structure.relative_lattice_energy 1.5
_theoretical_structure.rank                    5

# Crystal
_symmetry.cell_setting           monoclinic
_symmetry.space_group_name_H-M   'P 21/c'
_symmetry.Int_Tables_number      14
_space_group_name_Hall           '-P 2ybc'
loop_
_symmetry.equiv_pos_site_id
_symmetry.equiv_pos_as_xyz
1 x,y,z
2 -x,1/2+y,1/2-z
3 -x,-y,-z
4 x,1/2-y,1/2+z
_cell.length_a                   16.8168
_cell.length_b                   7.0178
_cell.length_c                   14.5475
_cell.angle_alpha                90
_cell.angle_beta                 115.28
_cell.angle_gamma                90
_cell.volume                     1552.44
loop_
_atom_site.label
_atom_site.type_symbol
_atom_site.fract_x
_atom_site.fract_y
_atom_site.fract_z
C0  C  0.402802  0.483043 0.842739
C1  C  0.387973  0.677363 0.792505
H2  H  0.372266  0.370900 0.785899
C3  C  0.290934  0.723546 0.722019
C4  C  0.299744  0.784092 0.627602
O5  O  0.430454  0.669327 0.717154
N6  N  0.375029  0.749323 0.626530
C7  C  0.433253  0.840791 0.865082
H8  H  0.373059  0.480692 0.896797
H9  H  0.473281  0.453133 0.883342
H10 H  0.406023  0.852375 0.921481
H11 H  0.422667  0.976203 0.823613
H12 H  0.504341  0.815441 0.904886
H13 H  0.247676  0.598250 0.706716
H14 H  0.264985  0.835703 0.754024
S15 S  0.217342  0.894753 0.520420
```

```
O16 O  0.132826  0.836886 0.515882
O17 O  0.235781  0.880775 0.432108
C18 C  0.229140  1.158880 0.556705
F19 F  0.218005  1.168970 0.646067
F20 F  0.314654  1.210640 0.581184
C21 C  0.161031  1.270540 0.472953
C22 C  0.174831  1.346830 0.391627
C23 C  0.075521  1.280400 0.468012
C24 C  0.007484  1.364770 0.384876
F25 F  0.256854  1.339740 0.393375
C26 C  0.107309  1.429290 0.307776
C27 C  0.023141  1.436700 0.304553
H28 H -0.030676  1.496570 0.238225
H29 H -0.058002  1.370540 0.382481
H30 H  0.062589  1.219340 0.529056
H31 H  0.120713  1.485310 0.246093
```

Structures at different temperature conditions generated from the same structure in the previous stage:

```
#
# Structure at 100K
#
data_structure_1_2_0
# Structure is thoeretically generated
_exptl.method                                                'theoretical model'

# Datablock Details
_csp.data_block_class                                        "Theoretical Structure"
_csp.data_block_description                                  structure_1_2_0
_csp.data_block_id                                           56cfac30-680a-4821-b29a-28c5991beba9

# Include files specifying input, methods and workflow data blocks
_csp.data_block_additional_files                             [ "csp_input.cif" "workflow.cif" ]

# Stage identifiers
_theoretical_structure.csp_input_system_description          Urea_Hydrate
_theoretical_structure.csp_input_system_id                   2a2611e3-2021-4b03-a7c6-0ef71239008f

_theoretical_structure.csp_workflow_description              wf2
_theoretical_structure.csp_workflow_id                       964fcd15-82e1-4c4d-a7cb-61b0b34c3421

_theoretical_structure.csp_generation_stage_description      ea
_theoretical_structure.csp_generation_stage_id               af534fed-8153-4af2-bd9f-29b0fef8d805

_theoretical_structure.csp_ranking_stage_description         pbe0_qha
_theoretical_structure.csp_ranking_stage_id                  11d2779e-6396-4c2c-91ff-d62dddaf9cc1

_theoretical_structure.csp_reference_temperature             100.0
_theoretical_structure.csp_reference_pressure                100000.0

_theoretical_structure.csp_previous_stage_structure_description    structure_1_1
_theoretical_structure.csp_previous_stage_structure_id       00d2779e-6396-4c2c-91ff-d62dddaf9cc1

# Crystal properties and details
...

#
# Structure at 200K
#
data_structure_1_2_1
# Structure is thoeretically generated
_exptl.method                                                'theoretical model'

# Datablock Details
_csp.data_block_class                                        "Theoretical Structure"
_csp.data_block_description                                  structure_1_2_1
_csp.data_block_id                                           28775053-e8fb-4e2b-bdb7-0ae255a422ef

# Include files specifying input, methods and workflow data blocks
_csp.data_block_additional_files                             [ "csp_input.cif" "workflow.cif" ]

# Stage identifiers
_theoretical_structure.csp_input_system_description          Urea_Hydrate
_theoretical_structure.csp_input_system_id                   2a2611e3-2021-4b03-a7c6-0ef71239008f

_theoretical_structure.csp_workflow_description              wf2
_theoretical_structure.csp_workflow_id                       964fcd15-82e1-4c4d-a7cb-61b0b34c3421

_theoretical_structure.csp_generation_stage_description      ea
_theoretical_structure.csp_generation_stage_id               af534fed-8153-4af2-bd9f-29b0fef8d805

_theoretical_structure.csp_ranking_stage_description         pbe0_qha
_theoretical_structure.csp_ranking_stage_id                  11d2779e-6396-4c2c-91ff-d62dddaf9cc1

_theoretical_structure.csp_reference_temperature             200.0
_theoretical_structure.csp_reference_pressure                100000.0

_theoretical_structure.csp_previous_stage_structure_description    structure_1_1
```

```
_theoretical_structure.csp_previous_stage_structure_id            00d2779e-6396-4c2c-91ff-d62dddaf9cc1

# Crystal properties and details
...


#
# Structure at 300K
#
data_structure_1_2_2
# Structure is thoeretically generated
_exptl.method 'theoretical model'

# Datablock Details
_csp.data_block_class                                             "Theoretical Structure"
_csp.data_block_description                                       structure_1_2_2
_csp.data_block_id                                                28775053-e8fb-4e2b-bdb7-0ae255a422ef

# Include files specifying input, methods and workflow data blocks
_csp.data_block_additional_files                                  [ "csp_input.cif" "workflow.cif" ]

# Stage identifiers
_theoretical_structure.csp_input_system_description               Urea_Hydrate
_theoretical_structure.csp_input_system_id                        2a2611e3-2021-4b03-a7c6-0ef71239008f

_theoretical_structure.csp_workflow_description                   wf2
_theoretical_structure.csp_workflow_id                            964fcd15-82e1-4c4d-a7cb-61b0b34c3421

_theoretical_structure.csp_generation_stage_description           ea
_theoretical_structure.csp_generation_stage_id                    af534fed-8153-4af2-bd9f-29b0fef8d805

_theoretical_structure.csp_ranking_stage_description              pbe0_qha
_theoretical_structure.csp_ranking_stage_id                       11d2779e-6396-4c2c-91ff-d62dddaf9cc1

_theoretical_structure.csp_reference_temperature                  300.0
_theoretical_structure.csp_reference_pressure                     100000.0

_theoretical_structure.csp_previous_stage_structure_description   structure_1_1
_theoretical_structure.csp_previous_stage_structure_id            00d2779e-6396-4c2c-91ff-d62dddaf9cc1

# Crystal properties and details
...
```

5. Conventions

A few guidelines are adopted in the description of specific data fields as highlighted in the table below. Except for *pDFT*, full names are preferred.

Category	Data Field	Suggested Input Item	Alternatives to avoid
Structure Generation	`method`	Random Sampling	Quasirandom, Pseudorandom (specified in separate datafield)
Structure Generation	`method`	Evolutionary Algorithm	Genetic Algorithm, EA, GA
CompChem	`method`	Forcefield	Force Field, Force-Field, FF
CompChem	`method`	pDFT	DFT, Density Functional Theory, periodic-DFT
CompChem	`method`	Semi-Empirical	Semi Empirical

In addition, the Structure Ranking `method` "ML Potential" refers to methods using *ad hoc* descriptors for neural network training to directly compute energy and forces. On the other hand, ML models used to parameterise models constants should be classified in the related method. For example, forcefield constants parametrised with a ML network should be classified as "Forcefield".

6. Future Developments

A few areas relevant to CSP have not been explored yet and might be included in later updates of the dictionary. In general, new or specific methods can use the "Other" option and specify possible publications describing the workflow.

A list of missing sections is shown below:

- Initial molecule (or list of molecules and conformers) coordinates and properties.
- Explicit search variables description (cell parameters, components' positions and orientation, internal degrees of freedom)
- ML-based Structure Generation methods.
- Clustering algorithms used to remove duplicates.
- While the TCOD dictionary is available for DFT methods and a draft dictionary for forcefield methods is being developed, data fields of other energy evaluation methods are limited to basic identification labels. This includes:
  - Semi-Empirical methods
  - ML Potentials

    - Free energy correction methods
- Output structure properties are limited to the energy or score of the crystal. Other measurable properties (the band gap for example) are not currently included.

7. Computational Chemistry Dictionary

The `_csp` prefix in energy/scoring methods and output structures have been intentionally left out so that the present data fields could be used as a basis for the development of a more general computational chemistry dictionary. While this is currently limited to a selection of data fields relevant to CSP application, below is an example of how these can be used to describe a single geometry optimisation on a known structure:

```
data_optimised_structure
# Structure is theoretically generated
_exptl.method 'theoretical model'

# General and Software details
_compchem.method pDFT
_compchem.calculation_type Optimisation

_compchem.software "Quantum Espresso"
_compchem.software_version 6.0
_compchem.software_citation "https://doi.org/10.1088/1361-648X/aa8f79"

# Energy Evaluation
_dft.exchange_correlation_functional_type GGA
_dft.exchange_correlation_functional_name PBE
_dft.pseudopotential_type PAW
_dft.dispersion_correction XDM
_dft.kinetic_energy_cutoff_wavefunctions 600  # TCOD Dictionary
_dft.BZ_integration.method "Monkhorst-Pack"
_dft.BZ_integration.grid_dens_X 0.5
_dft.BZ_integration.grid_dens_Y 0.5
_dft.BZ_integration.grid_dens_Z 0.5

# Geometry Optimisation
_compchem.geometry_optimisation_algorithm FIRE
_compchem.geometry_optimisation_cell anisotropic
_compchem.geometry_optimisation_atoms all
_compchem.geometry_optimisation_relax_force_convergence 0.01
_compchem.geometry_optimisation_max_steps 200

# Output Structure Details
_theoretical_structure.temperature 0.0
_theoretical_structure.calculated_density 1.314748
_theoretical_structure.total_energy -85493.52397

# Crystal
_symmetry.cell_setting           monoclinic
_symmetry.space_group_name_H-M   'P 21/c'
_symmetry.Int_Tables_number      14
_space_group_name_Hall           '-P 2ybc'
loop_
_symmetry.equiv_pos_site_id
_symmetry.equiv_pos_as_xyz
1 x,y,z
2 -x,1/2+y,1/2-z
3 -x,-y,-z
4 x,1/2-y,1/2+z
_cell.length_a                   16.8168
_cell.length_b                   7.0178
_cell.length_c                   14.5475
_cell.angle_alpha                90
_cell.angle_beta                 115.28
_cell.angle_gamma                90
_cell.volume                     1552.44
loop_
_atom_site.label
_atom_site.type_symbol
_atom_site.fract_x
_atom_site.fract_y
_atom_site.fract_z
C0  C  0.402802  0.483043 0.842739
C1  C  0.387973  0.677363 0.792505
H2  H  0.372266  0.370900 0.785899
C3  C  0.290934  0.723546 0.722019
C4  C  0.299744  0.784092 0.627602
O5  O  0.430454  0.669327 0.717154
N6  N  0.375029  0.749323 0.626530
C7  C  0.433253  0.840791 0.865082
H8  H  0.373059  0.480692 0.896797
H9  H  0.473281  0.453133 0.883342
H10 H  0.406023  0.852375 0.921481
H11 H  0.422667  0.976203 0.823613
H12 H  0.504341  0.815441 0.904886
H13 H  0.247676  0.598250 0.706716
H14 H  0.264985  0.835703 0.754024
S15 S  0.217342  0.894753 0.520420
```

```
O16 O  0.132826  0.836886 0.515882
O17 O  0.235781  0.880775 0.432108
C18 C  0.229140  1.158880 0.556705
F19 F  0.218005  1.168970 0.646067
F20 F  0.314654  1.210640 0.581184
C21 C  0.161031  1.270540 0.472953
C22 C  0.174831  1.346830 0.391627
C23 C  0.075521  1.280400 0.468012
C24 C  0.007484  1.364770 0.384876
F25 F  0.256854  1.339740 0.393375
C26 C  0.107309  1.429290 0.307776
C27 C  0.023141  1.436700 0.304553
H28 H -0.030676  1.496570 0.238225
H29 H -0.058002  1.370540 0.382481
H30 H  0.062589  1.219340 0.529056
H31 H  0.120713  1.485310 0.246093
```