CSP Data Standards

CCDC Update - October 2025

CSP Data Standards

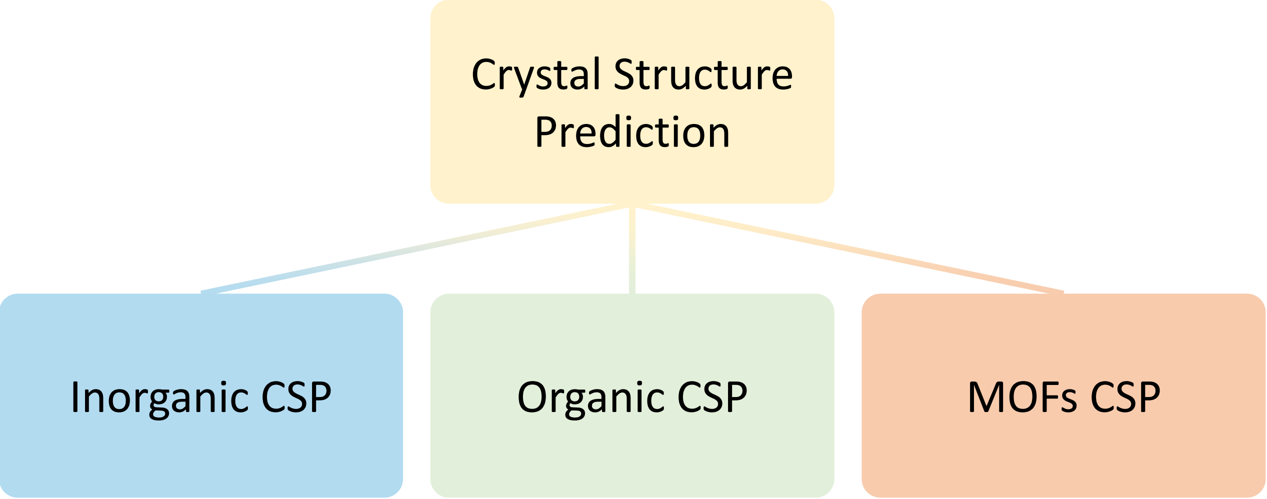
Scope of the project:

- An inclusive initiative that establishes **community standards** for predicted crystal structures
- Enables effective publication of results and easy **search and retrieval** across resources
- Supports levels of **reproducibility** deemed appropriate by different communities
- **Flexible structure** for future expansion as CSP methods are continuing to evolve

IUCr
International Union
of Crystallography

CCDC

TCOD



Timeline

- **Kick-off meetings** October/November 2023
- **Use case discussions** February/March 2024
- **Assimilation of input** June/July 2024
- **Share and seek further community input** December 2024
- **First Dictionary Draft** September 2025

A first draft of a CSP dictionary for molecular crystals was proposed during the **7th CSP Blind Test**. Subsequent meetings were held to address participant feedback and integrate input from researchers in the **inorganic** and **MOF** communities, as well as **industrial** partners.

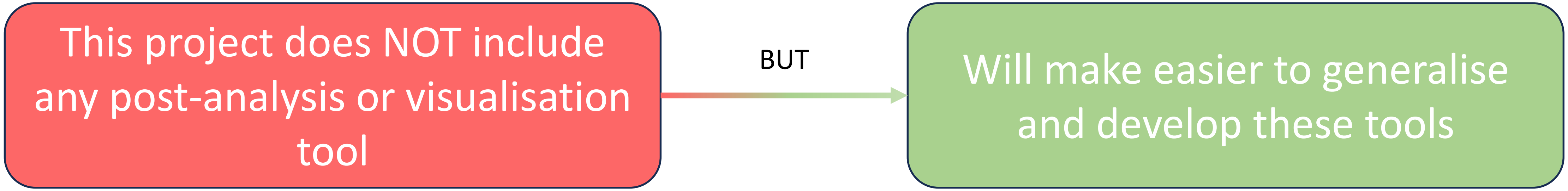
We focused on creating new data fields that use **controlled vocabularies** instead of free text and improving the description of **energy evaluation methods**.

New data fields were developed to enhance the description of **structure generation methods** and to meet the specific needs of specific communities. For instance, new fields were added for inorganic CSP to support **variable composition** searches, and for organic CSP to detail **dispersion corrections**.

A **GitHub Repository** has been made publicly available

Reflections on initial discussions

- Perception that **different communities may have quite different requirements** - different starting points, different outcomes
- Recognition that methods in some areas are **still evolving** so concepts may be hard to pin down
- Different expectations around what **levels of reproducibility** are necessary – also some methods are stochastic
- Need to consider what validation against standards is desirable – **what data items are essential**, desirable, optional etc.

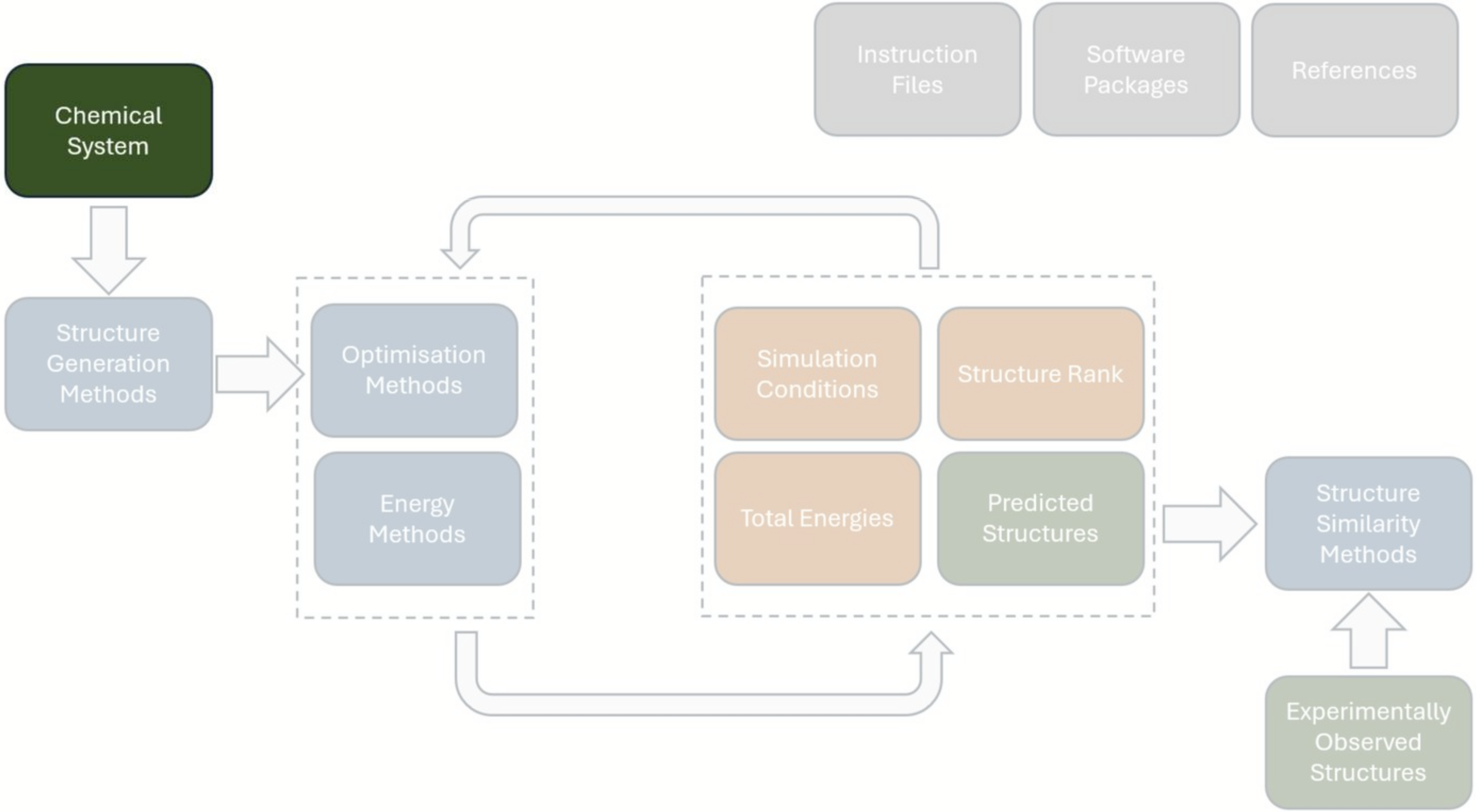


This project does NOT include any post-analysis or visualisation tool

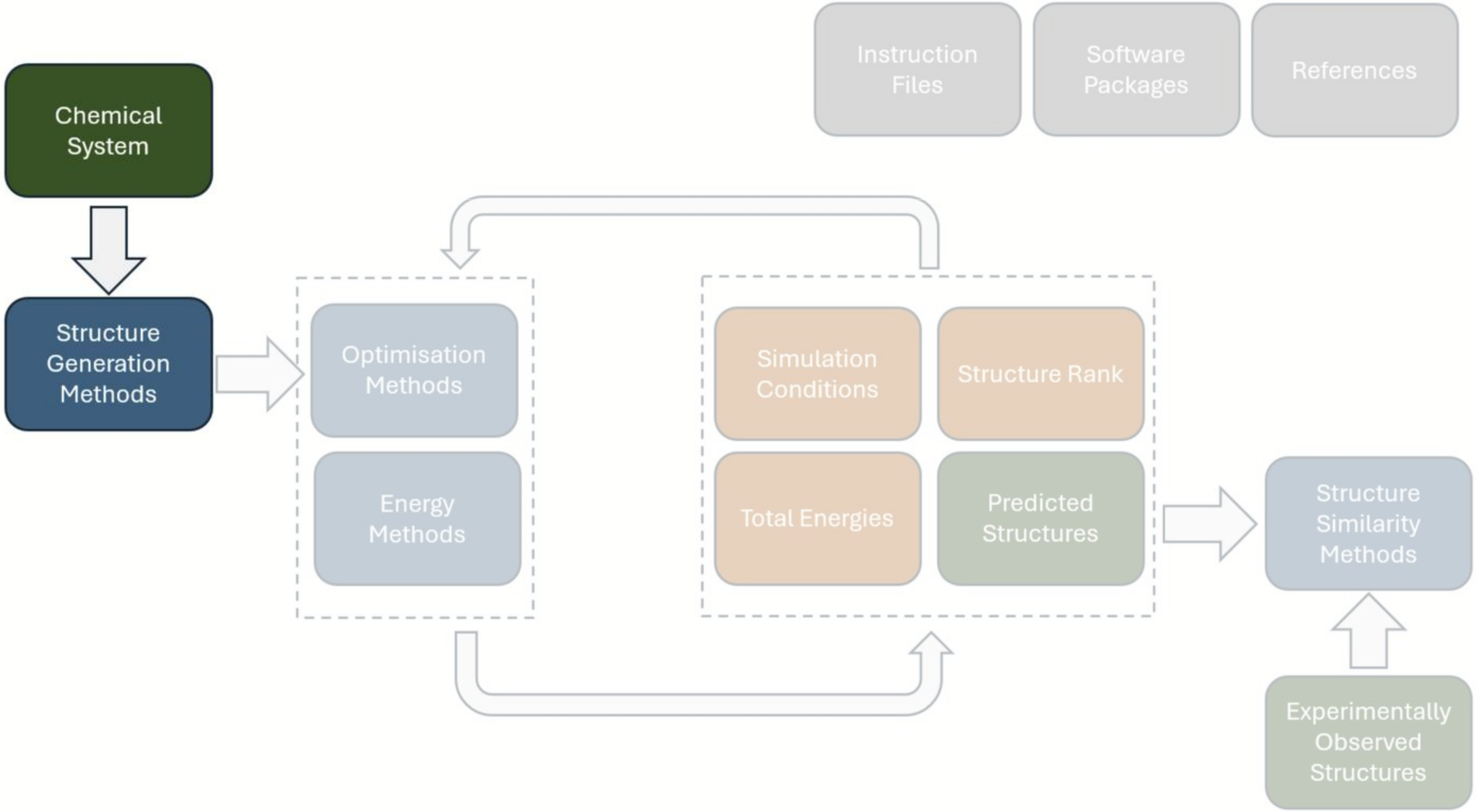
BUT

Will make easier to generalise and develop these tools

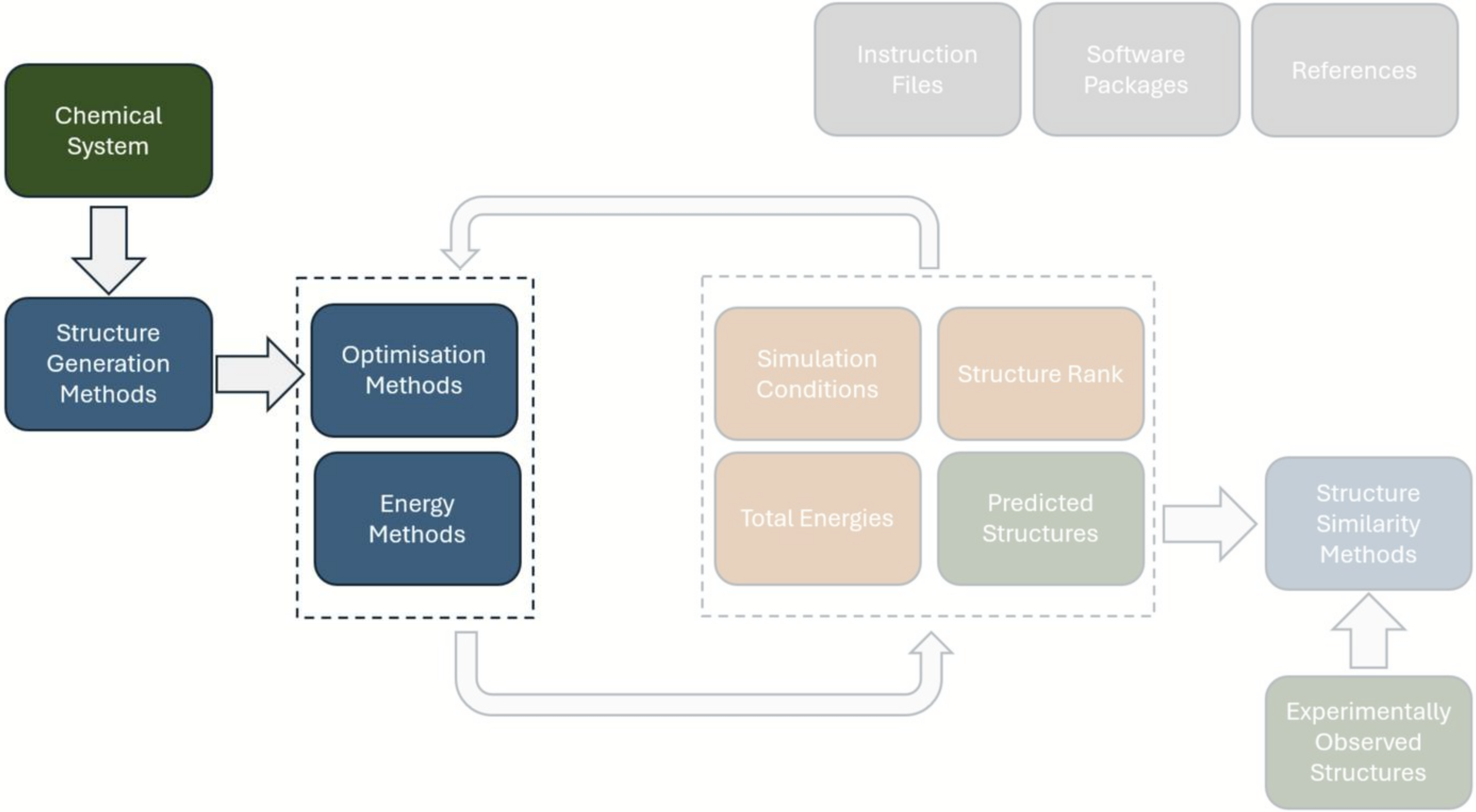
Conceptual Model



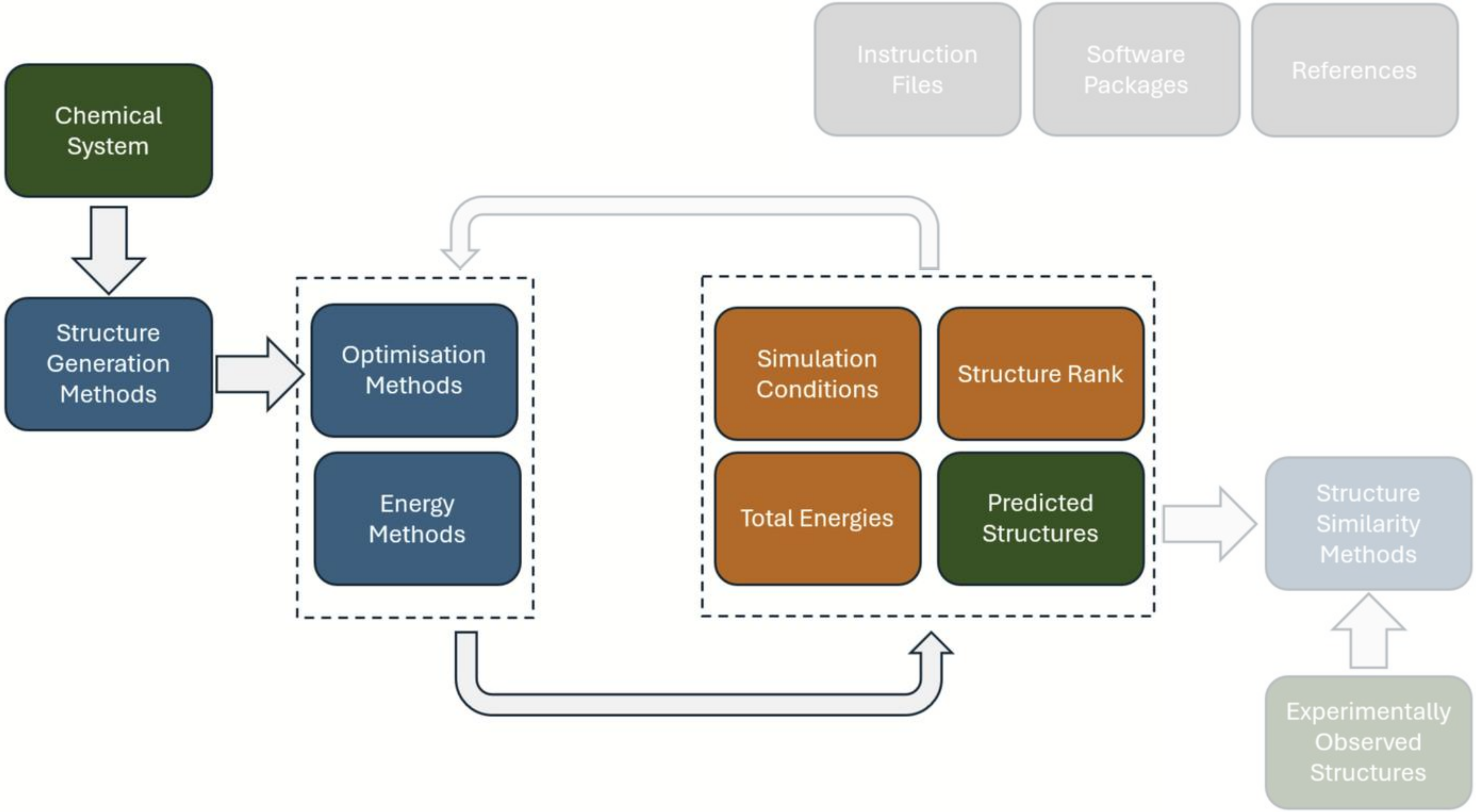
Conceptual Model



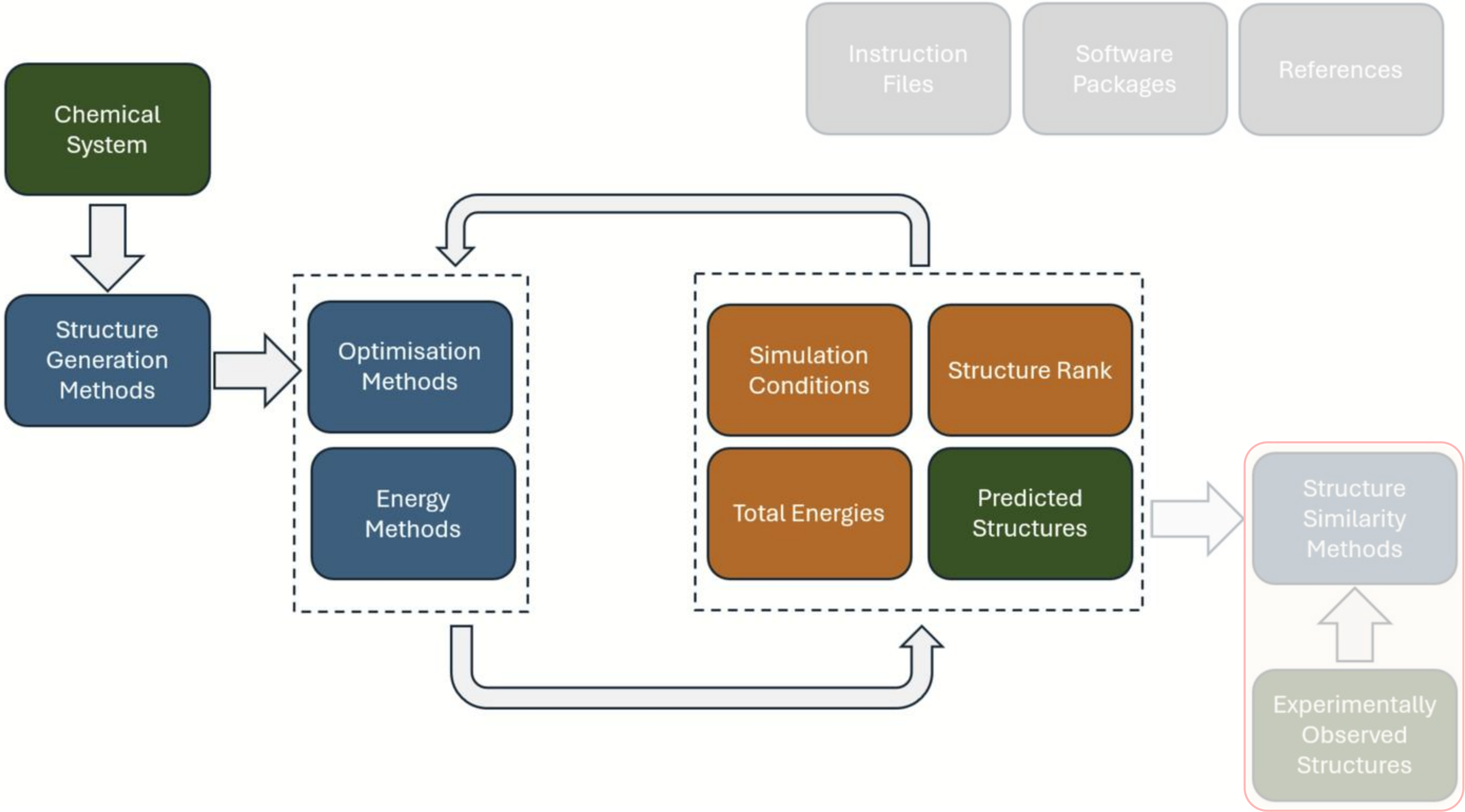
Conceptual Model



Conceptual Model



Conceptual Model

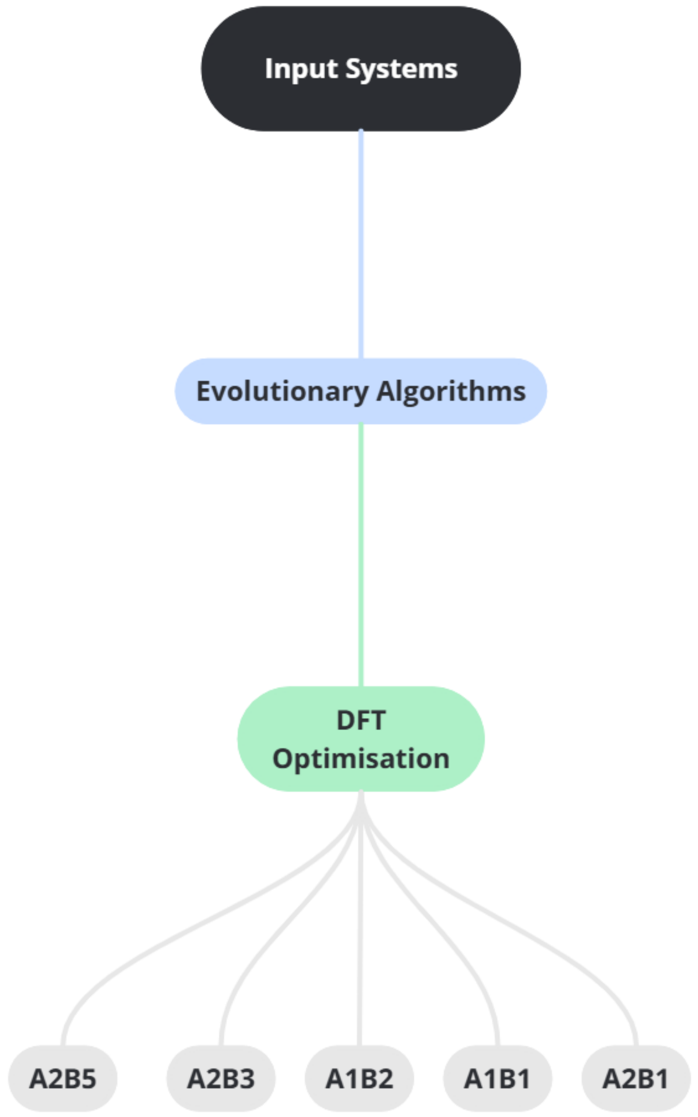


The CSP Dictionary

A good dictionary should be able to accurately describe the complex workflows used in CSP today:

- Variable composition searches

Variable Composition

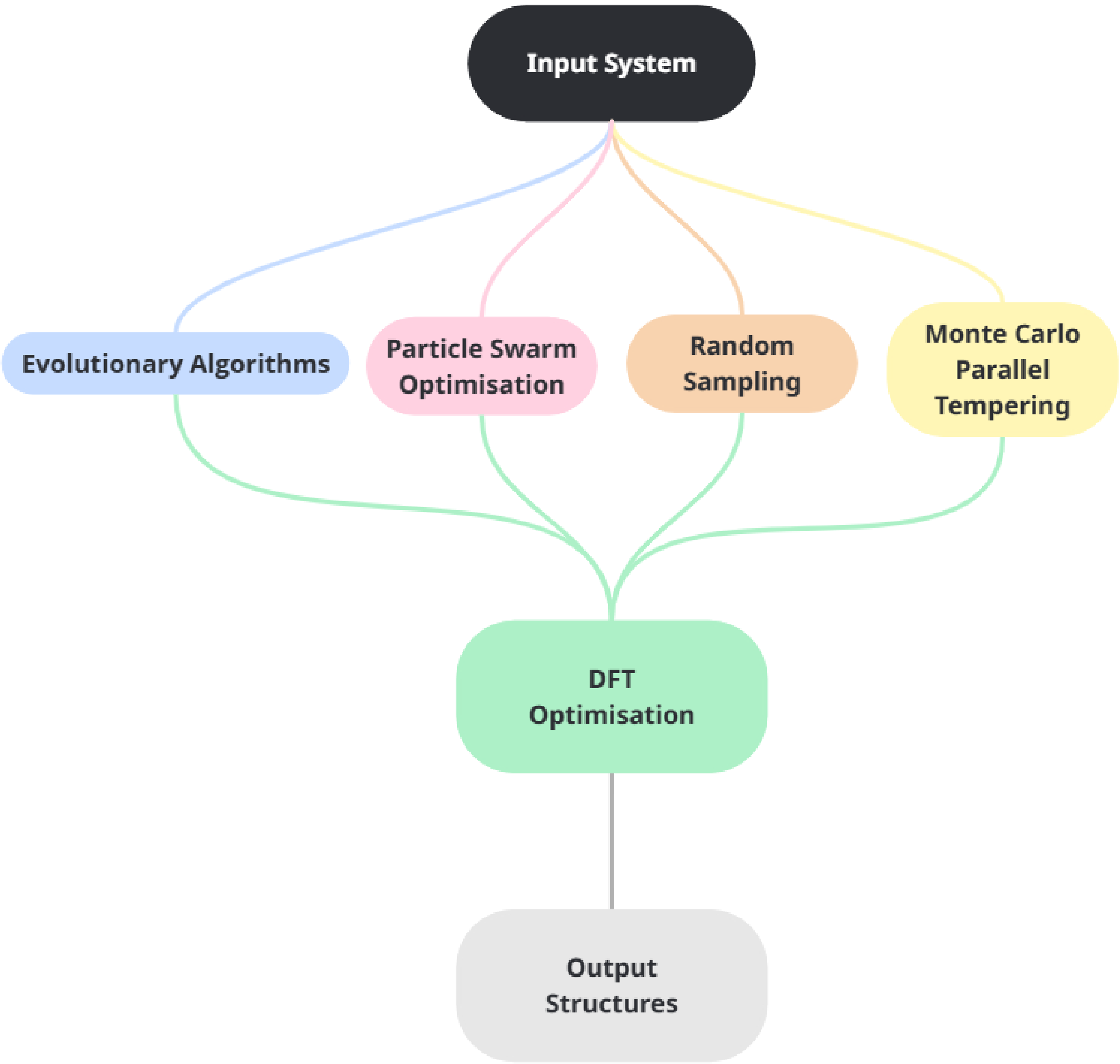


The CSP Dictionary

A good dictionary should be able to accurately describe the complex workflows used in CSP today:

- Variable composition searches
- The use of different generation methods

Multiple Generation Methods

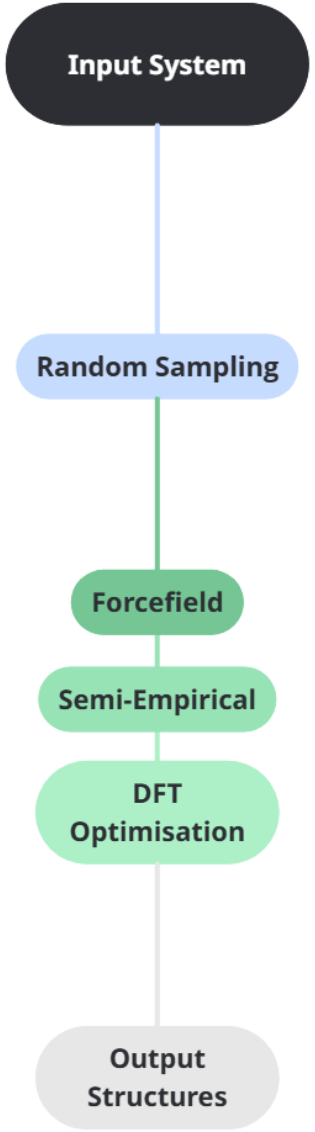


The CSP Dictionary

A good dictionary should be able to accurately describe the complex workflows used in CSP today:

- Variable composition searches
- The use of different generation methods
- The use of multiple geometry optimisation and energy evaluation methods

Multi-Step Optimisation



The Current CSP Dictionary

Input Chemical System

- Atoms
- Molecules
- Fixed or variable stoichiometry

Structure Generation Methods

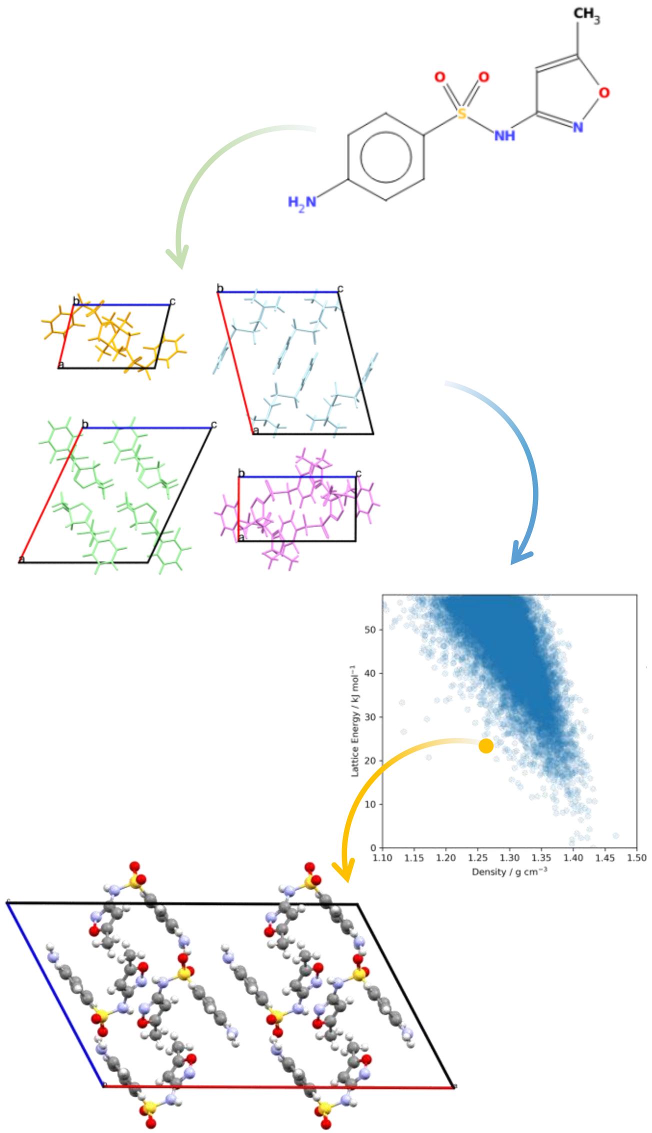
- Evolutionary Algorithms
- Particle Swarm Optimisation
- MC Simulated Annealing
- MC Parallel Tempering
- Random Search

Structure Ranking Methods

- Periodic Density Functional Theory
- Forcefields
- Semi-Empirical
- Wavefunction
- ML Potentials
- Free Energy

Output Structure

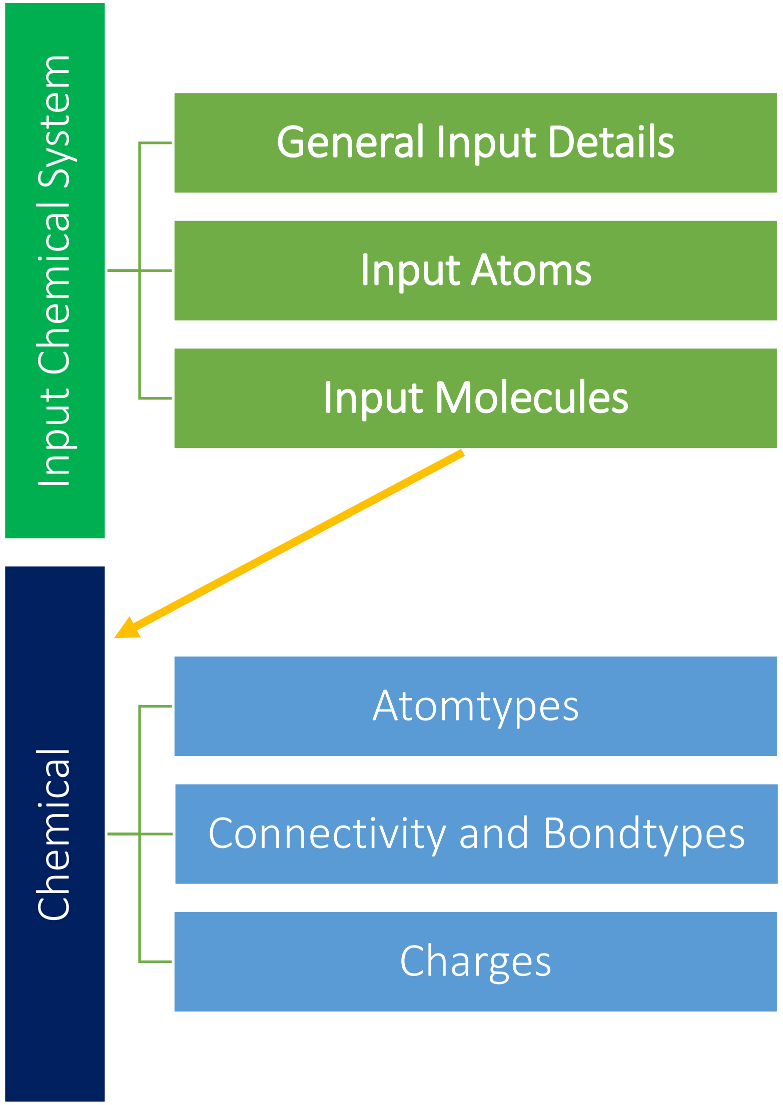
- Energy and rank
- 3D Structure
- Pressure and temperature



Additional Dictionaries

- CIF Core Dictionary
- Chemical Dictionary (Core CIF)
- TCOD DFT Dictionary
- Draft Forcefield Dictionary

Input Chemical System



Input Chemical System

General Input Details

Input Atoms

Input Molecules

Chemical

Atomtypes

Connectivity and Bondtypes

Charges

```
data_CSP_Meth
_csp.input.nam

# Molecules
```

```
# Atoms in molecules
loop_
   _csp.input_molecule.molecule_number
   _csp.input_molecule.molecule_identifier
   _csp.input_molecule.atom_label
   _chemical.conn_atom.number
                               pe_symbol
                               arge
```

```
# 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
```

Variable Composition Search

Given 3 atoms (A, B, C) you can either have a fixed stoichiometry or variable composition search

Fixed stoichiometry

```
_csp.input.name AxByCz
_csp.input_atoms.types [ A B C ]
_csp.input.composition_calculation "fixed"
_csp.input.composition_coefficients [[1 1 1] [1 1 2] [1 2 1] [2
1 1]]
```

Single or multiple stoichiometries

Variable composition

```
_csp.input.name AxByCz
_csp.input_atoms.types [ A B C ]
_csp.input.composition_calculation "variable"
_csp.input.minimum_number_of_components 2
_csp.input.maximum_number_of_components 12
# xA + yB + zC with 2<x+y+z<12
_csp.input.composition_coefficients [[1 0 0] [0 1 0] [0 0 1]]
```

- Atoms (Inorganic CSP)
- Molecules (Molecules)

Variable Composition Search

Fayalite

```
_csp.input.name Ferrosilite
_csp.input_atoms.types Fe Si O
_csp.input.composition_calculation fixed
_csp.input.composition_coefficients [2 1 4]
```

$$Fe_2SiO_4$$

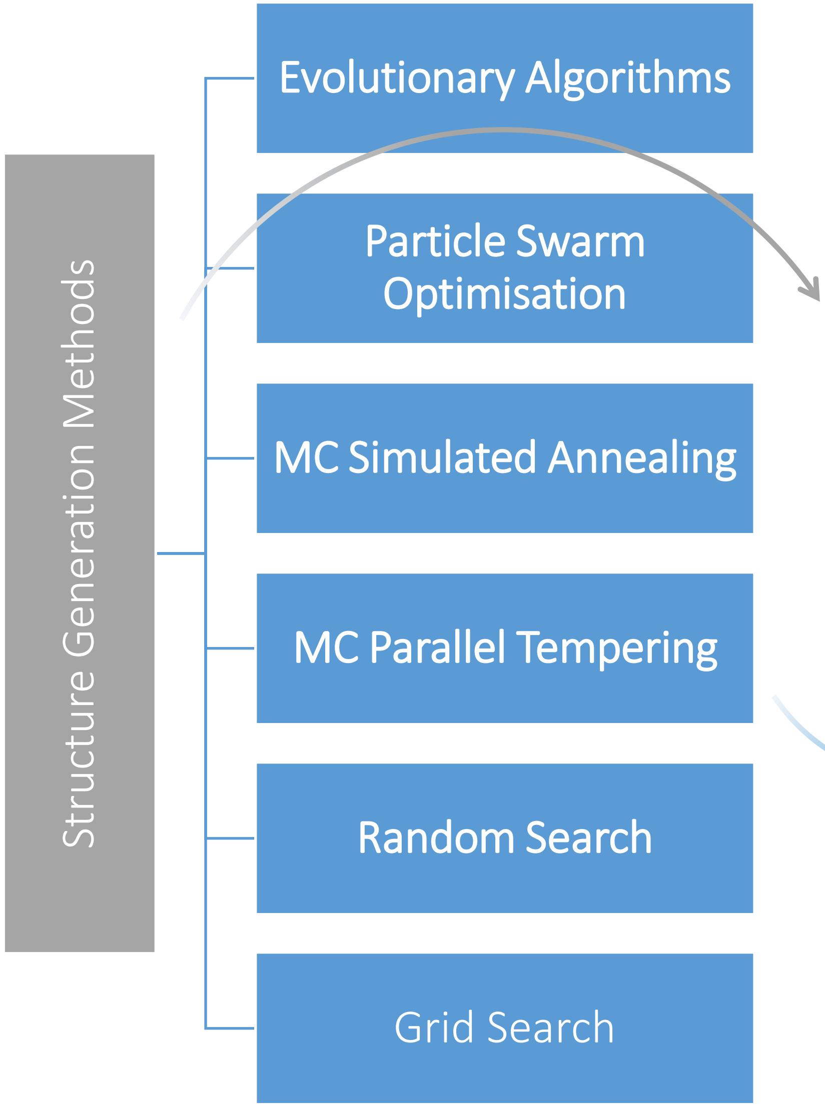
Olivine

```
_csp.input.name Hypersthene
_csp.input_atoms.types Fe Mg Si O
_csp.input.composition_calculation variable
_csp.input.composition_coefficients [[2 0 1 4] [0 2 1 4]]
_csp.input.minimum_number_of_components 2
_csp.input.maximum_number_of_components 10
```

$$x(\mathrm{Fe_2SiO_4}) + y(\mathrm{Mg_2SiO_4})$$

$$(Mg_aFe_{1-a})_2\mathrm{SiO_4}$$

Structure Generation Methods



Structure Generation	method	Monte Carlo Parallel Tempering
	space_group_number_list	all
	stopping_max_structures_evaluated	10000
	density_lower_limit	800
	density_upper_limit	1600
MC Parallel Tempering	number_of_replicas	5
	temperatures_list	50, 150, 250, 300, 350
	number_of_steps	100

Structure Generation Methods

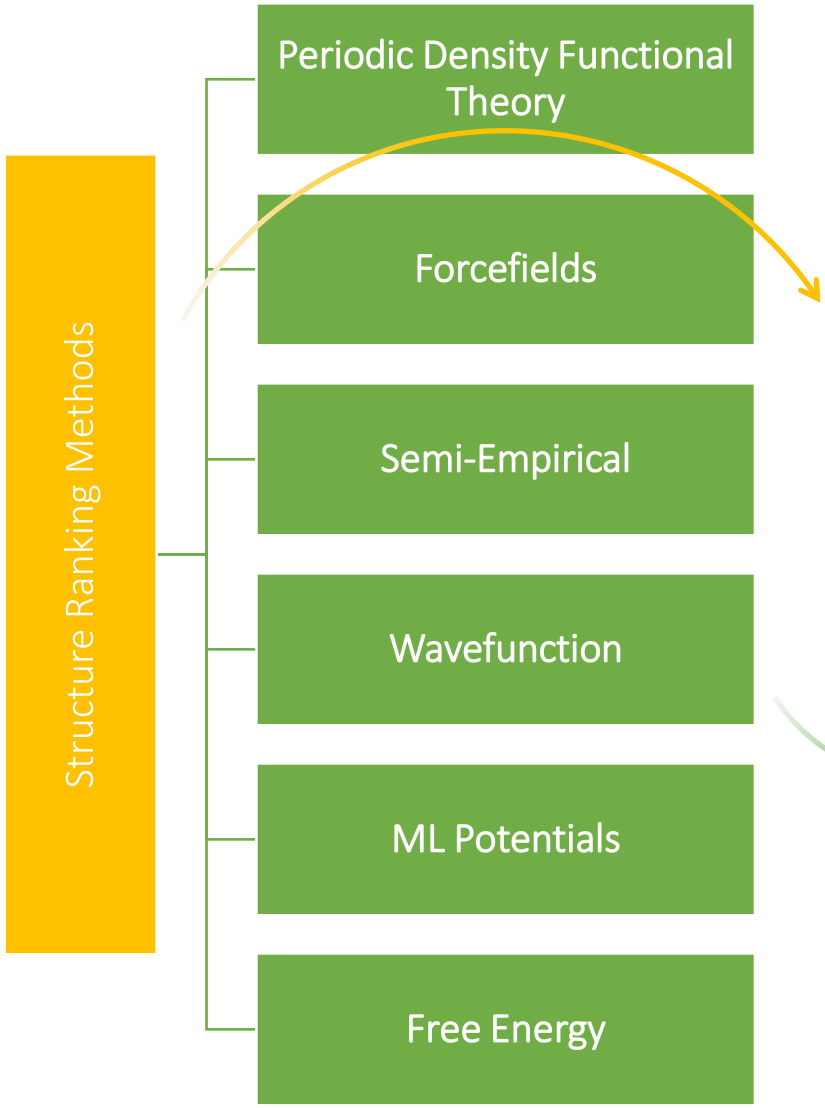
Single Method

```
# General Settings
_csp.structure_generation.space_group_number_list "all“
_csp.structure_generation.density_lower_limit 750
_csp.structure_generation.density_upper_limit 1500

# Evolutionary Algorithm Settings
_csp.structure_generation.method "Evolutionary Algorithm"
_csp.evolutionary_algorithm.population_size 20
_csp.evolutionary_algorithm.number_of_generations 100
_csp.evolutionary_algorithm.parents_structure_fraction 0.7
_csp.evolutionary_algorithm.heredity_fraction 0.5
_csp.evolutionary_algorithm.mutation_fraction 0.2
_csp.evolutionary_algorithm.permutation_fraction 0.1
```

Multiple methods

```
# 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
_csp.structure_generation.density_upper_limit 1500
_csp.structure_generation.method ["Random Sampling" "Simulated
Annealing"]

# Random Search
_csp.random.random_numbers_algorithm "Quasi-random"
_csp.random.number_of_samples 50

# Simulated Annealing
_csp.simulated_annealing.initial_temperature 400
_csp.simulated_annealing.cooling_rate 0.95
_csp.simulated_annealing.number_of_steps 100
```

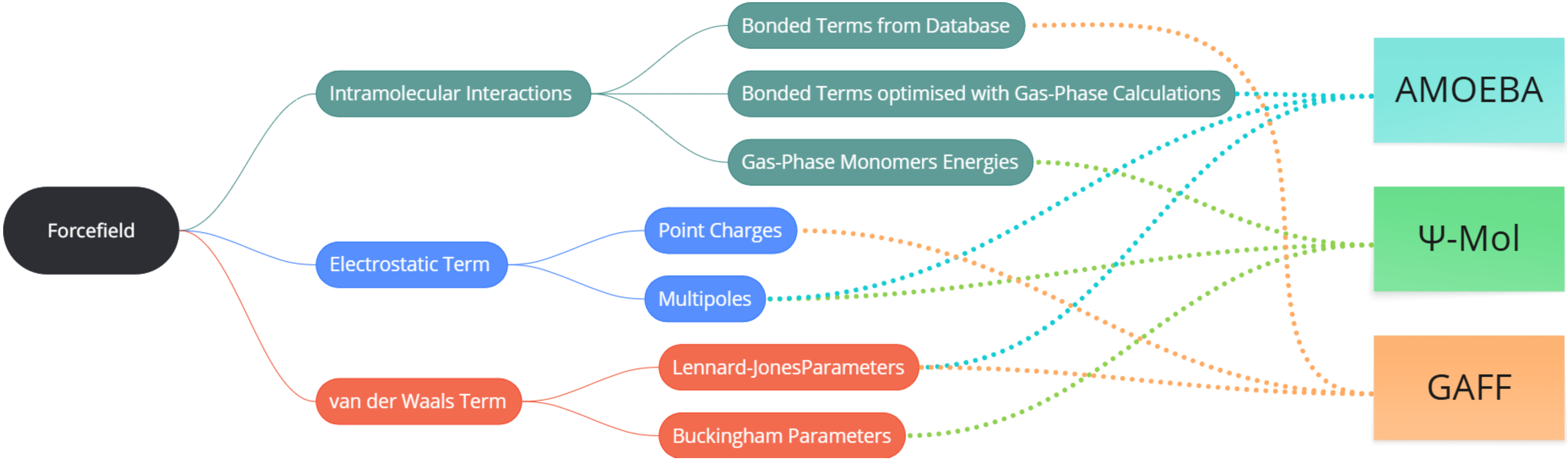
Structure Ranking Methods



Structure Ranking	method	pDFT
	stage_id	final
DFT	exchange_correlation_functional_type	GGA
	exchange_correlation_functional_name	PBE
	basis_set_type	PAW
	dispersion_correction	XDM

The Forcefield Dictionary

The forcefield dictionary aims at describing the different ways the potential terms can be combined


Forcefield
Intramolecular Interactions
Bonded Terms from Database
Bonded Terms optimised with Gas-Phase Calculations
Gas-Phase Monomers Energies
Electrostatic Term
Point Charges
Multipoles
van der Waals Term
Lennard-JonesParameters
Buckingham Parameters
AMOEBA
Ψ-Mol
GAFF

Multi-step Ranking Methods

Example of a 4 steps approach:

Stage	Stage ID	Calculation Type	Method (subdictionary)	Field	Value
0	GAFF	Optimisation	Forcefield	name	GAFF
				intramolecular_term	Bonded-Parameters
				electrostatic_term	Point-Charges
				vdw_term	"LJ(epsilon,sigma)"
				parameterization_method	BCC
				qm_parameterization_functional	AM1
		Optimisation	Forcefield	name	Psi_mol
				intramolecular_term	Isolated Molecule Energy
				electrostatic_term	Multipoles
				vdw_term	Buckingham
				parameterization_method	GDMA
				qm_parameterization_functional	PBE0
				qm_parameterization_basis_set	6-31G(d,p)
		Optimisation	DFT	exchange_correlation_functional_type	GGA
				exchange_correlation_functional_name	PBE
		Single-Point	DFT	exchange_correlation_functional_type	Hybrid
				exchange_correlation_functional_name	PBE0

Geometry Optimisation Datafields

Category	Data Field	Value
Geometry Optimisation	algorithm	FIRE
Geometry Optimisation	cell	anisotropic
Geometry Optimisation	atoms	all
Geometry Optimisation	relax_force_convergence	0.01

Example: Ranking Methods Datablocks

Main Dictionary	Subsection	Field	Value
Chemical	-	name	Urea
CSPCore	Structure Generation	space_group_number_list	all
		method	"Particle Swarm Optimisation"

Stage	Stage ID	Calculation Type	Method	Field	Value
0	GAFF	Optimisation	Forcefield	name	GAFF
				intramolecular_term	Bonded-Parameters
				electrostatic_term	Point-Charges
				vdw_term	"LJ(epsilon,sigma)"
				parameterization_method	BCC
				qm_parameterization_functional	AM1
1	Ψmol	Optimisation	Forcefield	name	Psi_mol
				intramolecular_term	Isolated Molecule Energy
				electrostatic_term	Multipoles
				vdw_term	Buckingham
				parameterization_method	GDMA
				qm_parameterization_functional	PBE0
				qm_parameterization_basis_set	6-31G(d,p)
2	PBE	Optimisation	DFT	exchange_correlation_functional_type	GGA
				exchange_correlation_functional_name	PBE
3	PBE0	Single-Point	DFT	exchange_correlation_functional_type	Hybrid
				exchange_correlation_functional_name	PBE0

```
########################################################################
#
#    Example 1: Multistep Ranking Approach
#
########################################################################
data_method
_chemical.name urea
_csp.structure_generation.space_group_number_list all
_csp.structure_generation.method "Particle Swarm Optimisation"

loop_
   _csp.structure_ranking.stage
   _csp.structure_ranking.stage_id
   _csp.structure_ranking.calculation_type
   _csp.structure_ranking.method
   0 "gaff" "Optimisation" "Forcefield"
   1 "psi_mol" "Optimisation" "Forcefield"
   2 "pbe" "Optimisation" "DFT"
   3 "pbe0" "Single-Point" "DFT"

# END

########################################################################
#
#    Additional Parameters for each method
#
########################################################################
data_gaff
_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"
# END

data_psi_mol
_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)"
# END

data_pbe
_dft.exchange_correlation_functional_type "GGA"
_dft.exchange_correlation_functional_name "PBE"
# END

data_pbe0
_dft.exchange_correlation_functional_type "Hybrid"
_dft.exchange_correlation_functional_name "PBE0"
_dft.dispersion_correction "TS"
# END
```

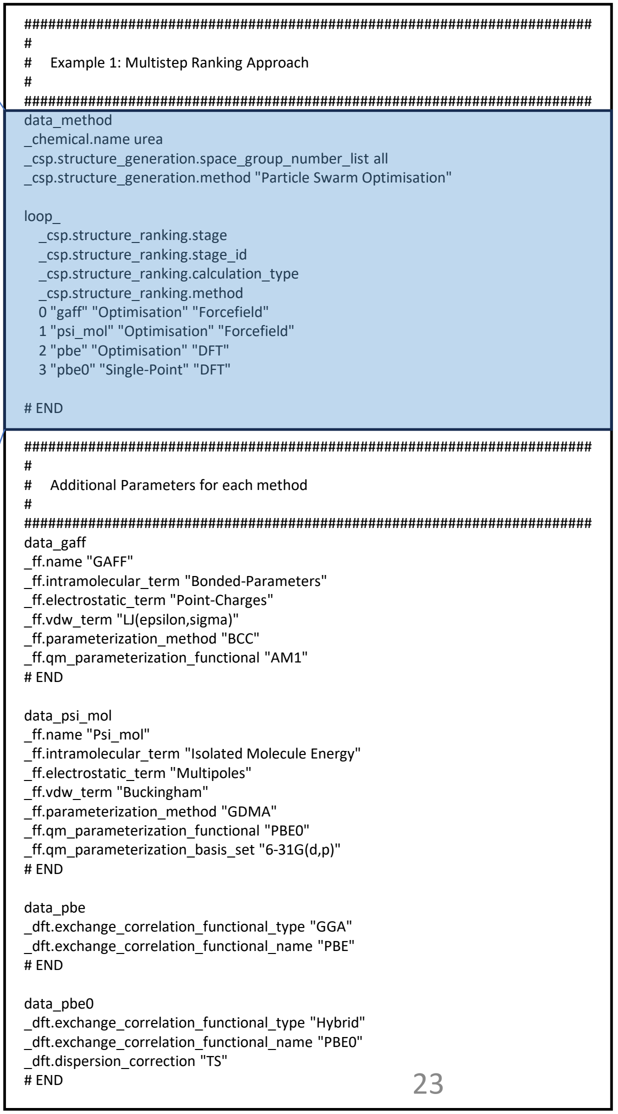
Example: Ranking Methods Datablocks

Main Dictionary	Subsection	Field	Value
Chemical	-	name	Urea
CSPCore	Structure Generation	space_group_number_list	all
		method	"Particle Swarm Optimisation"

Stage	Stage ID	Calculation Type	Method	Field	Value
0	GAFF	Optimisation	Forcefield	name	GAFF
				intramolecular_term	Bonded-Parameters
				electrostatic_term	Point-Charges
				vdw_term	"LJ(epsilon,sigma)"
				parameterization_method	BCC
				qm_parameterization_functional	AM1
1	Ψmol	Optimisation	Forcefield	name	Psi_mol
				intramolecular_term	Isolated Molecule Energy
				electrostatic_term	Multipoles
				vdw_term	Buckingham
				parameterization_method	GDMA
				qm_parameterization_functional	PBE0
				qm_parameterization_basis_set	6-31G(d,p)
2	PBE	Optimisation	DFT	exchange_correlation_functional_type	GGA
				exchange_correlation_functional_name	PBE
3	PBE0	Single-Point	DFT	exchange_correlation_functional_type	Hybrid
				exchange_correlation_functional_name	PBE0

Block 1

```
data_method
_chemical.name urea
_csp.structure_generation.space_group_number_list all
_csp.structure_generation.method "Particle Swarm Optimisation"

loop_
  _csp.structure_ranking.stage
  _csp.structure_ranking.stage_id
  _csp.structure_ranking.calculation_type
  _csp.structure_ranking.method
  0 "gaff" "Optimisation" "Forcefield"
  1 "psi_mol" "Optimisation" "Forcefield"
  2 "pbe" "Optimisation" "DFT"
  3 "pbe0" "Single-Point" "DFT"
```



```
##########################################################################
#
#    Example 1: Multistep Ranking Approach
#
##########################################################################
data_method
_chemical.name urea
_csp.structure_generation.space_group_number_list all
_csp.structure_generation.method "Particle Swarm Optimisation"

loop_
  _csp.structure_ranking.stage
  _csp.structure_ranking.stage_id
  _csp.structure_ranking.calculation_type
  _csp.structure_ranking.method
  0 "gaff" "Optimisation" "Forcefield"
  1 "psi_mol" "Optimisation" "Forcefield"
  2 "pbe" "Optimisation" "DFT"
  3 "pbe0" "Single-Point" "DFT"

# END
##########################################################################
#
#    Additional Parameters for each method
#
##########################################################################
data_gaff
_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"
# END

data_psi_mol
_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)"
# END

data_pbe
_dft.exchange_correlation_functional_type "GGA"
_dft.exchange_correlation_functional_name "PBE"
# END

data_pbe0
_dft.exchange_correlation_functional_type "Hybrid"
_dft.exchange_correlation_functional_name "PBE0"
_dft.dispersion_correction "TS"
# END
```

Example: Ranking Methods Datablocks

Main Dictionary	Subsection	Field	Value
Chemical	-	name	Urea
CSPCore	Structure Generation	space_group_number_list	all
		method	"Particle Swarm Optimisation"

Stage	Stage ID	Calculation Type	Method	Field	Value
0	GAFF	Optimisation	Forcefield	name	GAFF
				intramolecular_term	Bonded-Parameters
				electrostatic_term	Point-Charges
				vdw_term	"LJ(epsilon,sigma)"
				parameterization_method	BCC
				qm_parameterization_functional	AM1
1	Ψmol	Optimisation	Forcefield	name	Psi_mol
				intramolecular_term	Isolated Molecule Energy
				electrostatic_term	Multipoles
				vdw_term	Buckingham
				parameterization_method	GDMA
				qm_parameterization_functional	PBE0
				qm_parameterization_basis_set	6-31G(d,p)
2	PBE	Optimisation	DFT	exchange_correlation_functional_type	GGA
				exchange_correlation_functional_name	PBE
3	PBE0	Single-Point	DFT	exchange_correlation_functional_type	Hybrid
				exchange_correlation_functional_name	PBE0

Block 1

Block 2

```
data_gaff
_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"
```

```
################################################################################
#
#    Example 1: Multistep Ranking Approach
#
################################################################################
data_method
_chemical.name urea
_csp.structure_generation.space_group_number_list all
_csp.structure_generation.method "Particle Swarm Optimisation"

loop_
   _csp.structure_ranking.stage
   _csp.structure_ranking.stage_id
   _csp.structure_ranking.calculation_type
   _csp.structure_ranking.method
   0 "gaff" "Optimisation" "Forcefield"
   1 "psi_mol" "Optimisation" "Forcefield"
   2 "pbe" "Optimisation" "DFT"
   3 "pbe0" "Single-Point" "DFT"

# END

################################################################################
#
#    Additional Parameters for each method
#
################################################################################
data_gaff
_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"
# END

data_psi_mol
_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)"
# END

data_pbe
_dft.exchange_correlation_functional_type "GGA"
_dft.exchange_correlation_functional_name "PBE"
# END

data_pbe0
_dft.exchange_correlation_functional_type "Hybrid"
_dft.exchange_correlation_functional_name "PBE0"
_dft.dispersion_correction "TS"
# END
```

Example: Ranking Methods Datablocks

Main Dictionary	Subsection	Field	Value
Chemical	-	name	Urea
CSPCore	Structure Generation	space_group_number_list	all
		method	"Particle Swarm Optimisation"

Stage	Stage ID	Calculation Type	Method	Field	Value
0	GAFF	Optimisation	Forcefield	name	GAFF
				intramolecular_term	Bonded-Parameters
				electrostatic_term	Point-Charges
				vdw_term	"LJ(epsilon,sigma)"
				parameterization_method	BCC
				qm_parameterization_functional	AM1
1	Ψmol	Optimisation	Forcefield	name	Psi_mol
				intramolecular_term	Isolated Molecule Energy
				electrostatic_term	Multipoles
				vdw_term	Buckingham
				parameterization_method	GDMA
				qm_parameterization_functional	PBE0
				qm_parameterization_basis_set	6-31G(d,p)
2	PBE	Optimisation	DFT	exchange_correlation_functional_type	GGA
				exchange_correlation_functional_name	PBE
3	PBE0	Single-Point	DFT	exchange_correlation_functional_type	Hybrid
				exchange_correlation_functional_name	PBE0

Block 1

Block 2

Block 3

```
data_psi_mol
_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)"
```

```
###############################################################################
#
#    Example 1: Multistep Ranking Approach
#
###############################################################################
data_method
_chemical.name urea
_csp.structure_generation.space_group_number_list all
_csp.structure_generation.method "Particle Swarm Optimisation"

loop_
   _csp.structure_ranking.stage
   _csp.structure_ranking.stage_id
   _csp.structure_ranking.calculation_type
   _csp.structure_ranking.method
   0 "gaff" "Optimisation" "Forcefield"
   1 "psi_mol" "Optimisation" "Forcefield"
   2 "pbe" "Optimisation" "DFT"
   3 "pbe0" "Single-Point" "DFT"

# END
###############################################################################
#
#    Additional Parameters for each method
#
###############################################################################
data_gaff
_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"
# END

data_psi_mol
_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)"
# END

data_pbe
_dft.exchange_correlation_functional_type "GGA"
_dft.exchange_correlation_functional_name "PBE"
# END

data_pbe0
_dft.exchange_correlation_functional_type "Hybrid"
_dft.exchange_correlation_functional_name "PBE0"
_dft.dispersion_correction "TS"
# END
```

Example: Ranking Methods Datablocks

Main Dictionary	Subsection	Field	Value
Chemical	-	name	Urea
CSPCore	Structure Generation	space_group_number_list	all
		method	"Particle Swarm Optimisation"

Stage	Stage ID	Calculation Type	Method	Field	Value
0	GAFF	Optimisation	Forcefield	name	GAFF
				intramolecular_term	Bonded-Parameters
				electrostatic_term	Point-Charges
				vdw_term	"LJ(epsilon,sigma)"
				parameterization_method	BCC
				qm_parameterization_functional	AM1
1	Ψmol	Optimisation	Forcefield	name	Psi_mol
				intramolecular_term	Isolated Molecule Energy
				electrostatic_term	Multipoles
				vdw_term	Buckingham
				parameterization_method	GDMA
				qm_parameterization_functional	PBE0
				qm_parameterization_basis_set	6-31G(d,p)
2	PBE	Optimisation	DFT	exchange_correlation_functional_type	GGA
				exchange_correlation_functional_name	PBE
3	PBE0	Single-Point	DFT	exchange_correlation_functional_type	Hybrid
				exchange_correlation_functional_name	PBE0

Block 1

Block 2

Block 3

Block 4

```
data_pbe
_dft.exchange_correlation_functional_type "GGA"
_dft.exchange_correlation_functional_name "PBE"
```

```
############################################################################
#
#    Example 1: Multistep Ranking Approach
#
############################################################################
data_method
_chemical.name urea
_csp.structure_generation.space_group_number_list all
_csp.structure_generation.method "Particle Swarm Optimisation"

loop_
   _csp.structure_ranking.stage
   _csp.structure_ranking.stage_id
   _csp.structure_ranking.calculation_type
   _csp.structure_ranking.method
   0 "gaff" "Optimisation" "Forcefield"
   1 "psi_mol" "Optimisation" "Forcefield"
   2 "pbe" "Optimisation" "DFT"
   3 "pbe0" "Single-Point" "DFT"

# END

############################################################################
#
#    Additional Parameters for each method
#
############################################################################
data_gaff
_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"
# END

data_psi_mol
_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)"
# END

data_pbe
_dft.exchange_correlation_functional_type "GGA"
_dft.exchange_correlation_functional_name "PBE"
# END

data_pbe0
_dft.exchange_correlation_functional_type "Hybrid"
_dft.exchange_correlation_functional_name "PBE0"
_dft.dispersion_correction "TS"
# END
```

Example: Ranking Methods Datablocks

Main Dictionary	Subsection	Field	Value
Chemical	-	name	Urea
CSPCore	Structure Generation	space_group_number_list	all
		method	"Particle Swarm Optimisation"

Stage	Stage ID	Calculation Type	Method	Field	Value
0	GAFF	Optimisation	Forcefield	name	GAFF
				intramolecular_term	Bonded-Parameters
				electrostatic_term	Point-Charges
				vdw_term	"LJ(epsilon,sigma)"
				parameterization_method	BCC
				qm_parameterization_functional	AM1
1	Ψmol	Optimisation	Forcefield	name	Psi_mol
				intramolecular_term	Isolated Molecule Energy
				electrostatic_term	Multipoles
				vdw_term	Buckingham
				parameterization_method	GDMA
				qm_parameterization_functional	PBE0
				qm_parameterization_basis_set	6-31G(d,p)
2	PBE	Optimisation	DFT	exchange_correlation_functional_type	GGA
				exchange_correlation_functional_name	PBE
3	PBE0	Single-Point	DFT	exchange_correlation_functional_type	Hybrid
				exchange_correlation_functional_name	PBE0

Block 1

Block 2

Block 3

Block 4

Block 5

```
data_pbe0
_dft.exchange_correlation_functional_type "Hybrid"
_dft.exchange_correlation_functional_name "PBE0"
_dft.dispersion_correction "TS"
```

```
########################################################################
#
#    Example 1: Multistep Ranking Approach
#
########################################################################
data_method
_chemical.name urea
_csp.structure_generation.space_group_number_list all
_csp.structure_generation.method "Particle Swarm Optimisation"

loop_
   _csp.structure_ranking.stage
   _csp.structure_ranking.stage_id
   _csp.structure_ranking.calculation_type
   _csp.structure_ranking.method
   0 "gaff" "Optimisation" "Forcefield"
   1 "psi_mol" "Optimisation" "Forcefield"
   2 "pbe" "Optimisation" "DFT"
   3 "pbe0" "Single-Point" "DFT"

# END

########################################################################
#
#    Additional Parameters for each method
#
########################################################################
data_gaff
_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"
# END

data_psi_mol
_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)"
# END

data_pbe
_dft.exchange_correlation_functional_type "GGA"
_dft.exchange_correlation_functional_name "PBE"
# END

data_pbe0
_dft.exchange_correlation_functional_type "Hybrid"
_dft.exchange_correlation_functional_name "PBE0"
_dft.dispersion_correction "TS"
# END
```

DDLm Data Items

A DDLm-style was adopted in CIF2 file format and we are planning to use it as standard for the CSP Dictionary:

- Explicit data group-subgroup relations:

```
_csp.structure_generation.space_group_number_list
```

- It can include lists and matrixes:

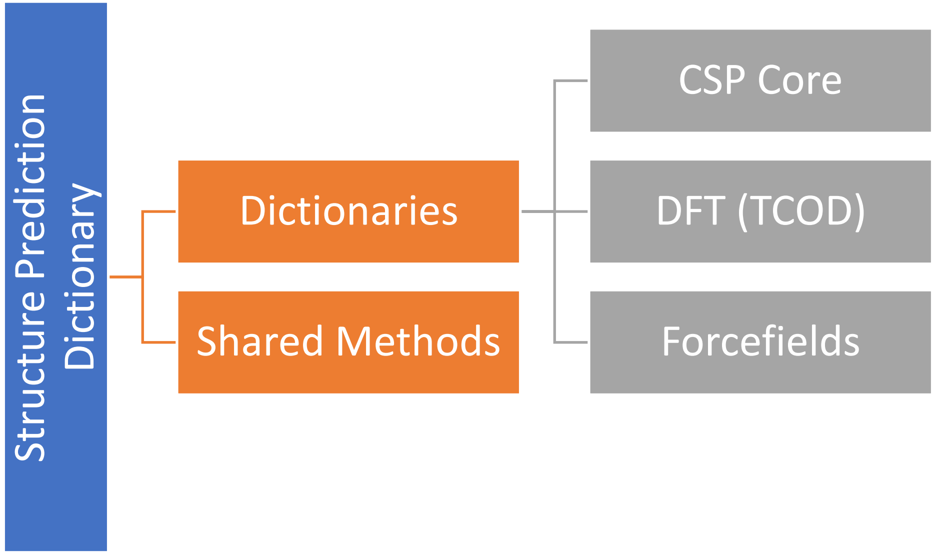
```
_csp.structure_generation.space_group_number_list [14 2 15 61 19 4 33 29 5 1]
```

- It can be back compatible with CIF1-style through aliases:

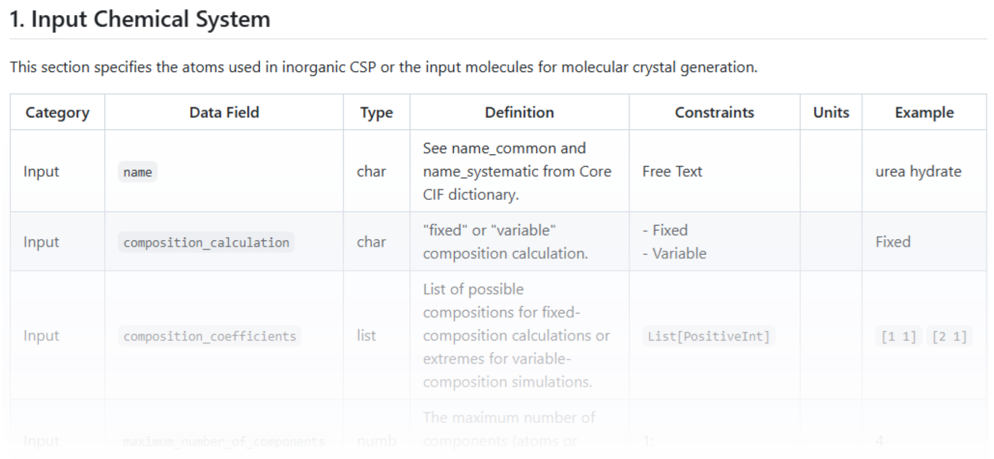
```
_csp_structure_generation_space_group_number_list "14 2 15 61 19 4 33 29 5 1"
```

The GitHub Repository: Dictionaries

Available at: github.com/COMCIFS/Structure_Prediction_Dictionary



- All folders include a text file with existing datafields and examples on how to use them:



1. Input Chemical System

This section specifies the atoms used in inorganic CSP or the input molecules for molecular crystal generation.

Category	Data Field	Type	Definition	Constraints	Units	Example
Input	`name`	char	See name_common and name_systematic from Core CIF dictionary.	Free Text		urea hydrate
Input	`composition_calculation`	char	"fixed" or "variable" composition calculation.	- Fixed - Variable		Fixed
Input	`composition_coefficients`	list	List of possible compositions for fixed-composition calculations or extremes for variable-composition simulations.	`List[PositiveInt]`		`[1 1]` `[2 1]`
Input	`maximum_number_of_components`	numb	The maximum number of components (atoms or	1:		4

- The DFT folder contains a link to the TCOD DFT dictionary and some additional datafields
- A PDF version is also available

The GitHub Repository: How to contribute

If missing fields are present, you can:

- Create a New Issue
- Create a Pull Request
- Add comments to the PDF file and/or email to the CSP Data Standards Team

To ensure consistency and usability of the dictionary, we added a SharedMethods folder, in which you can upload a paper or a file with your CSP workflow and we will use our standardized CSP dictionary to describe your methodology.

Discussion: Missing Data Fields

Feedbacks from different CSP Communities:

Inorganic CSP

Organic CSP

MOFs CSP

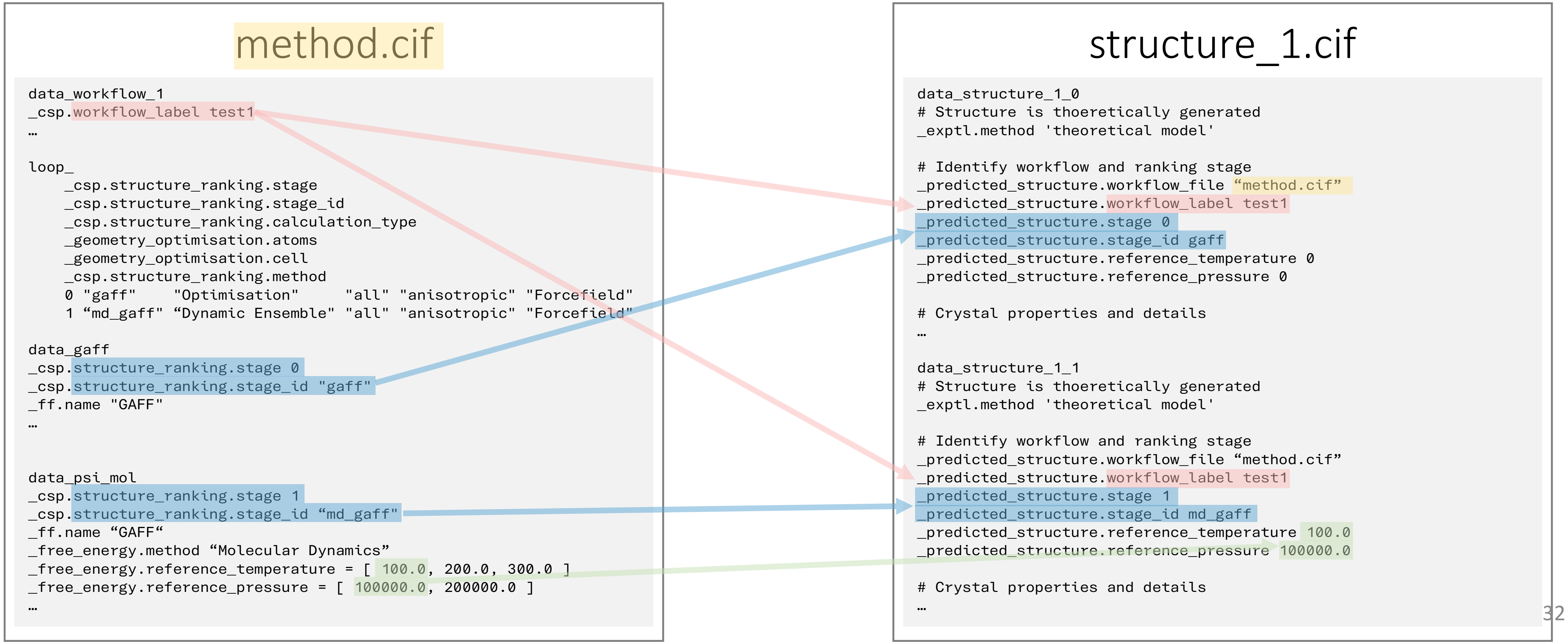
Polymers CSP

Input definition, structure generation and ranking methods not described in detail:

- ML Potentials
- Free Energy methods
- Conformers generation
- Output structure properties
- Search Constraints

Discussion: Files and data blocks

How to link the workflow, ranking stage and external conditions form the methods data blocks to the output structure:



method.cif

```
data_workflow_1
_csp.workflow_label test1
…

loop_
    _csp.structure_ranking.stage
    _csp.structure_ranking.stage_id
    _csp.structure_ranking.calculation_type
    _geometry_optimisation.atoms
    _geometry_optimisation.cell
    _csp.structure_ranking.method
    0 "gaff"    "Optimisation"     "all" "anisotropic" "Forcefield"
    1 “md_gaff" “Dynamic Ensemble" "all" "anisotropic" "Forcefield"

data_gaff
_csp.structure_ranking.stage 0
_csp.structure_ranking.stage_id "gaff"
_ff.name "GAFF"
…


data_psi_mol
_csp.structure_ranking.stage 1
_csp.structure_ranking.stage_id “md_gaff"
_ff.name “GAFF“
_free_energy.method “Molecular Dynamics”
_free_energy.reference_temperature = [ 100.0, 200.0, 300.0 ]
_free_energy.reference_pressure = [ 100000.0, 200000.0 ]
…
```

structure_1.cif

```
data_structure_1_0
# Structure is thoeretically generated
_exptl.method 'theoretical model'

# Identify workflow and ranking stage
_predicted_structure.workflow_file “method.cif”
_predicted_structure.workflow_label test1
_predicted_structure.stage 0
_predicted_structure.stage_id gaff
_predicted_structure.reference_temperature 0
_predicted_structure.reference_pressure 0

# Crystal properties and details
…

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

# Identify workflow and ranking stage
_predicted_structure.workflow_file “method.cif”
_predicted_structure.workflow_label test1
_predicted_structure.stage 1
_predicted_structure.stage_id md_gaff
_predicted_structure.reference_temperature 100.0
_predicted_structure.reference_pressure 100000.0

# Crystal properties and details
…
```

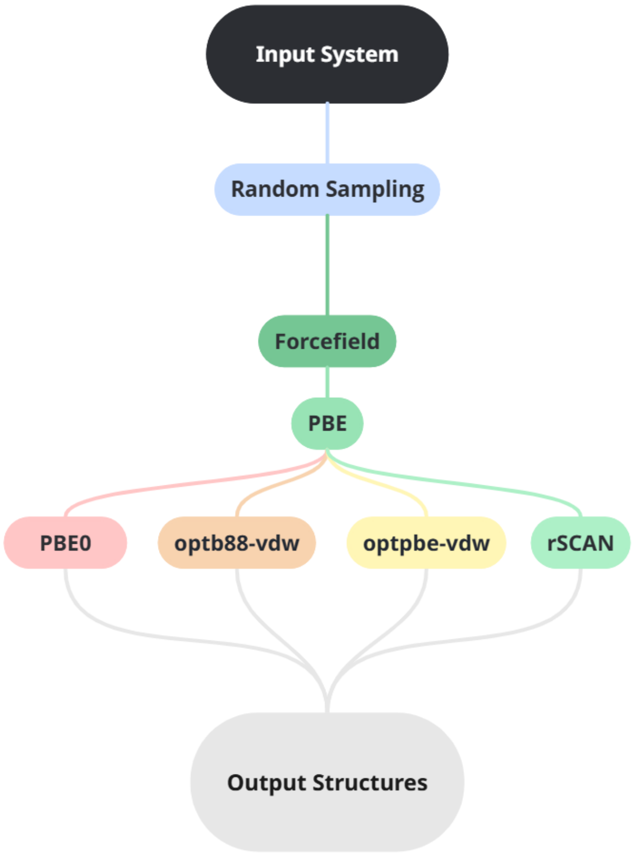
Discussion: Re-ranking with multiple methods

Description of workflows with multiple ranking methods are used to the same set of structures:

method.cif

```
data_workflow_1
# Input and structure generation details
_csp.workflow_label test1
…

# Ranking workflow
loop_
    _csp.structure_ranking.stage
    _csp.structure_ranking.stage_id
    _csp.structure_ranking.calculation_type
    _geometry_optimisation.atoms
    _geometry_optimisation.cell
    _csp.structure_ranking.method
    0  "gaff"     "Optimisation"  "all" "anisotropic" "Forcefield“
    1  “pbe”      “Optimisation”  "all" "anisotropic“ “pDFT”
    2a “pbe0”     “Single-Point”  .     .             “pDFT”
    2b “optb88”   “Single-Point”  .     .             “pDFT”
    2c “optpbe”   “Single-Point”  .     .             “pDFT”
    2d “rscan”    “Single-Point”  .     .             “pDFT”
…
```

Complex Multi-Step Optimisation

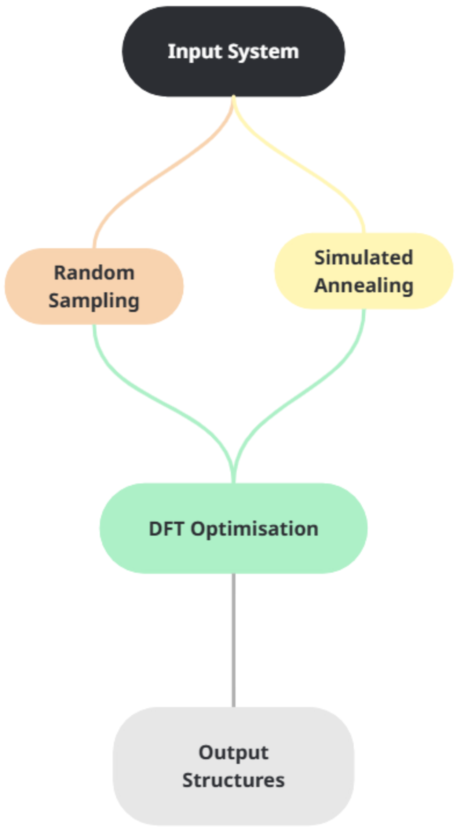


Discussion: Generation methods

Should generation methods be treated in the same way as ranking methods?

```
data_workflow_1

# 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
_csp.structure_generation.method ["Random Sampling" "Simulated Annealing"]

# Random Search
_csp.random.random_numbers_algorithm "Quasi-random"
_csp.random.number_of_samples 1000

# Simulated Annealing
_csp.simulated_annealing.initial_temperature 400
_csp.simulated_annealing.cooling_rate 0.95
_csp.simulated_annealing.number_of_steps 100
```

Multiple Generation Methods

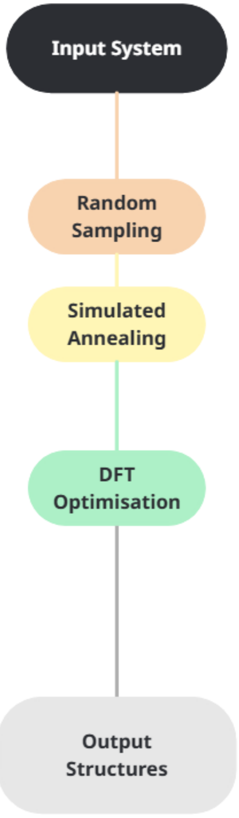


Discussion: Generation methods

Should generation methods be treated in the same way as ranking methods?

```
data_workflow_1

# 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

loop_
    _csp.structure_generation.stage
    _csp.structure_generation.stage_id
    _csp.structure_generation.method
    0 “rs” "Random Sampling"
    1 “sa” "Simulated Annealing“
…

# Random Search
data_rs
_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.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
```

Multi-stage Generation Methods

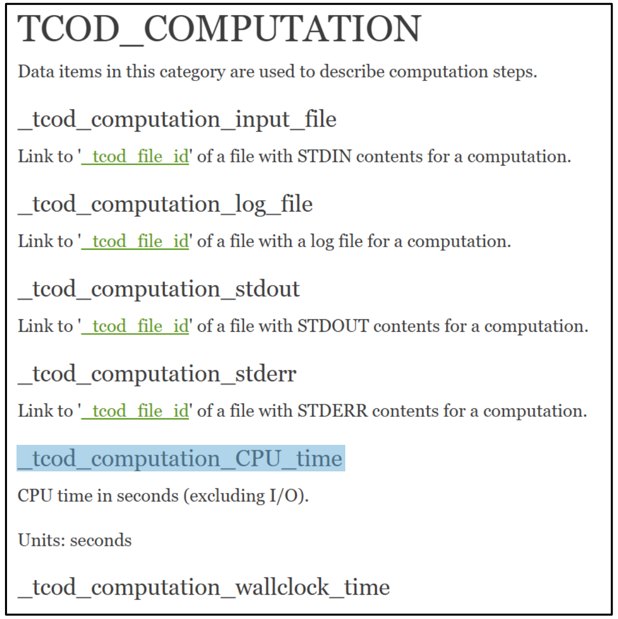


Discussion: Additional *General* Fields

Improve reproducibility by including input files:

```
_csp.input_file_name pw.in
_csp.input_file_content
;
 &CONTROL
    prefix='benzene’
 /

 &SYSTEM
    ibrav = 6
    A = 11.0
    C = 7.0
    ecutwfc =  20.0,
    ecutrho =  200.0,
    nat =  12,
    ntyp =  2,
    nbnd = 16
 /

ATOMIC_SPECIES
   C   1.0   C.pbe-rrkjus.UPF
   H   1.0   H.pbe-rrkjus.UPF

ATOMIC_POSITIONS angstrom
   H   5.5000000   7.98563953   3.5
…

K_POINTS gamma
;
```

Include details on computational costs:



TCOD_COMPUTATION

Data items in this category are used to describe computation steps.

_tcod_computation_input_file

Link to '_tcod_file_id' of a file with STDIN contents for a computation.

_tcod_computation_log_file

Link to '_tcod_file_id' of a file with a log file for a computation.

_tcod_computation_stdout

Link to '_tcod_file_id' of a file with STDOUT contents for a computation.

_tcod_computation_stderr

Link to '_tcod_file_id' of a file with STDERR contents for a computation.

_tcod_computation_CPU_time

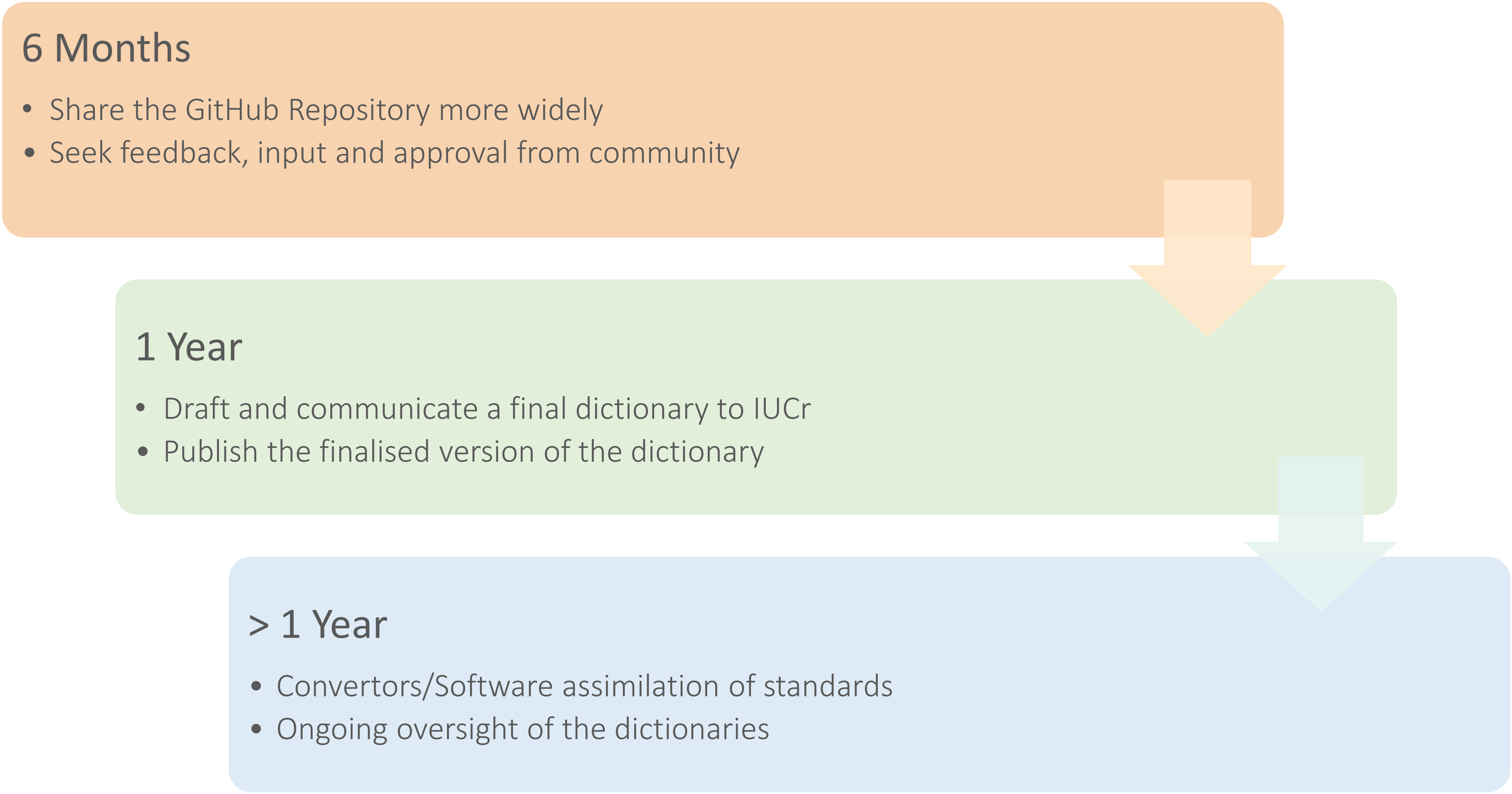
CPU time in seconds (excluding I/O).

Units: seconds

_tcod_computation_wallclock_time

https://wiki.crystallography.net/cif/dictionaries/cif_tcod

Next steps



6 Months

- Share the GitHub Repository more widely
- Seek feedback, input and approval from community

1 Year

- Draft and communicate a final dictionary to IUCr
- Publish the finalised version of the dictionary

> 1 Year

- Convertors/Software assimilation of standards
- Ongoing oversight of the dictionaries

Acknowledgments

CCDC

IUCr
International Union
of Crystallography

TCOD

The CSP Data Standards Team

Ian Bruno

Isaac Sugden

Lily Hunnisett

Jonas Nyman

Nicholas Francia

CCDC staff

Previous Contributors to discussions

Saulius Grazulis
Antanas Vaitkus
Hari Muddana
Dejan Zagorac
Jacco van der Streek
Kamil Dzuibek
Mike Bellucci
Shubham Sharma
Simon Westrip

Stefano Racioppi
Stephan Ruhl
Zhuocen Yang
Chris Pickard
Claire Adjiman
Andrius Merkys
Graeme Day
James Hester
Luca Ghiringhelli

Sally Price
Simon Coles
Simon Parsons
Zahra Momenzadeh
Erin Johnson
Arman Boromond
Susan Reutzel-Edens