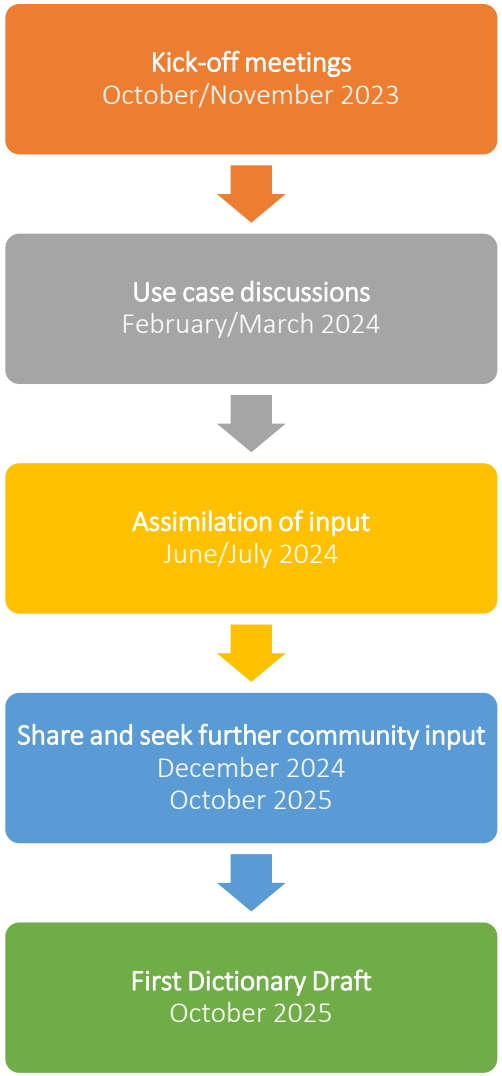
CSP Data Standards

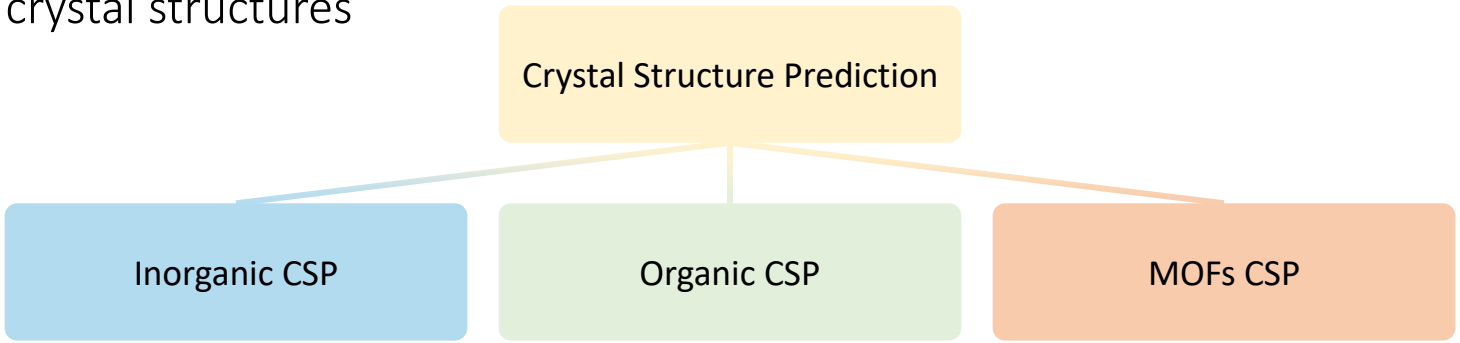
CCDC Update – February 2026

CSP Data Standards



Outcomes:

- An inclusive initiative that establishes **community standards** for predicted crystal structures



- Supports levels of **details** deemed appropriate by different communities
- **Flexible structure** for future expansion as CSP methods are continuing to evolve
- The dictionary format is designed to be **machine-readable**
- The focus of the project is on the development of a CSP dictionary describing structure generation methods and structure ranking methods

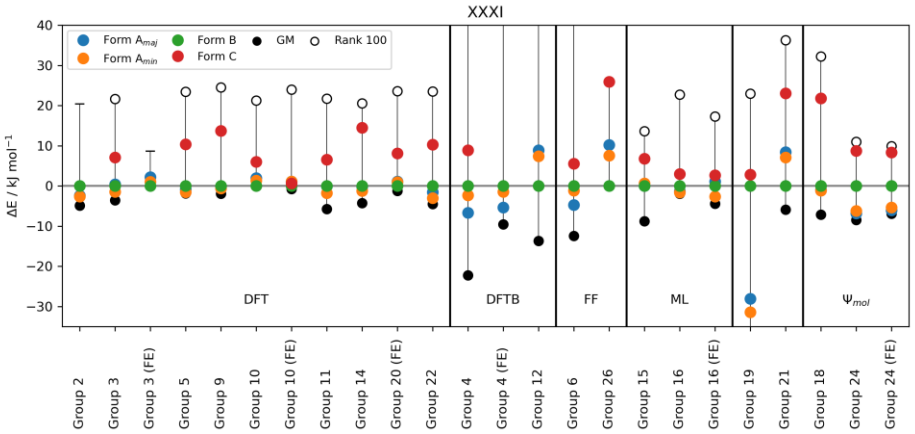
Applications of a CSP Standardised Dictionary

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

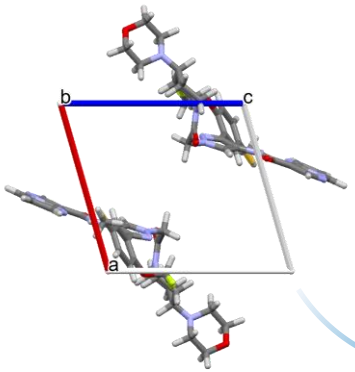
BUT

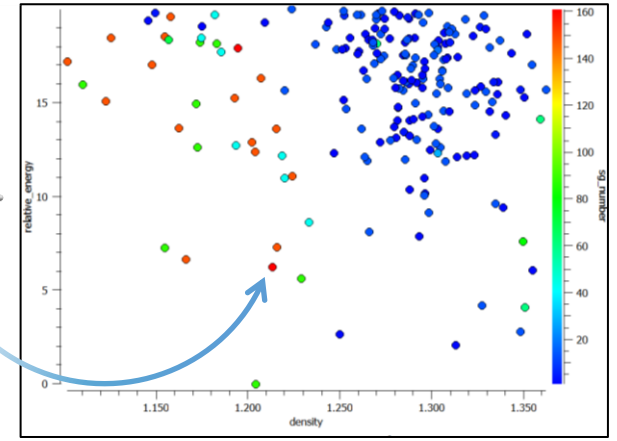
Provides the foundations to generalise and develop these tools

Creation of databases and benchmarking sets

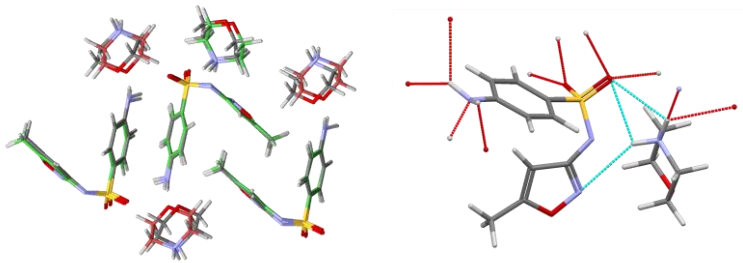


Visualisation of landscapes and structures





Post-analysis of structures



The CSP Dictionary

Input Chemical System

- Atoms
- Molecules
- Fixed or variable stoichiometry

Structure Generation Methods

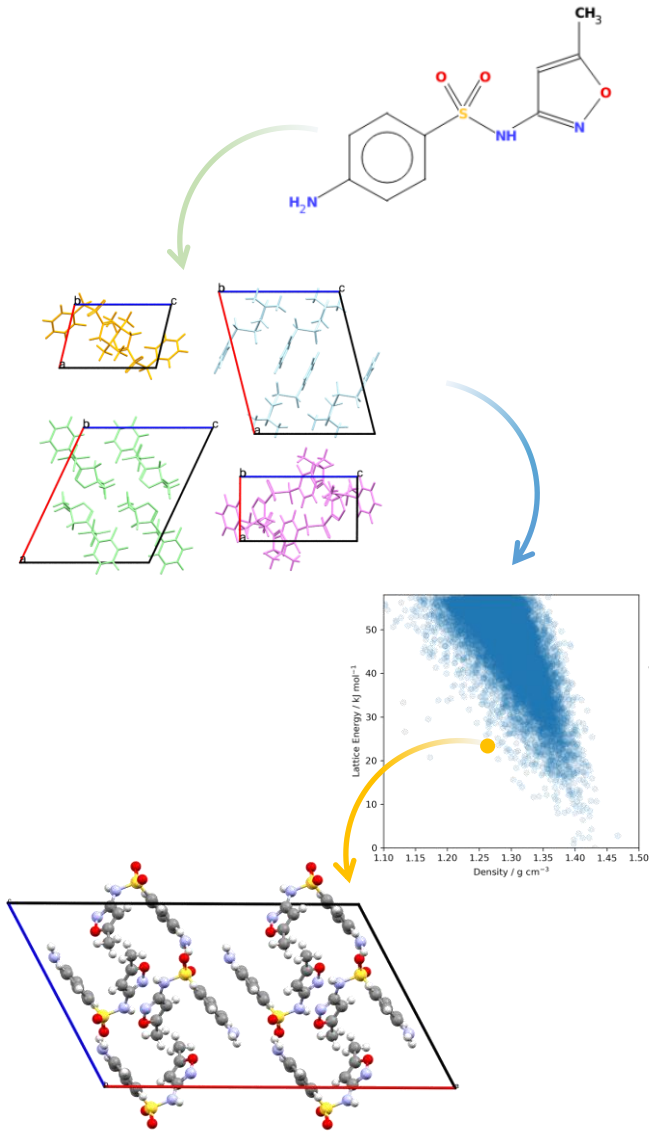
- Evolutionary Algorithms
- Particle Swarm Optimisation
- MC Simulated Annealing
- MC Parallel Tempering
- Random Search

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



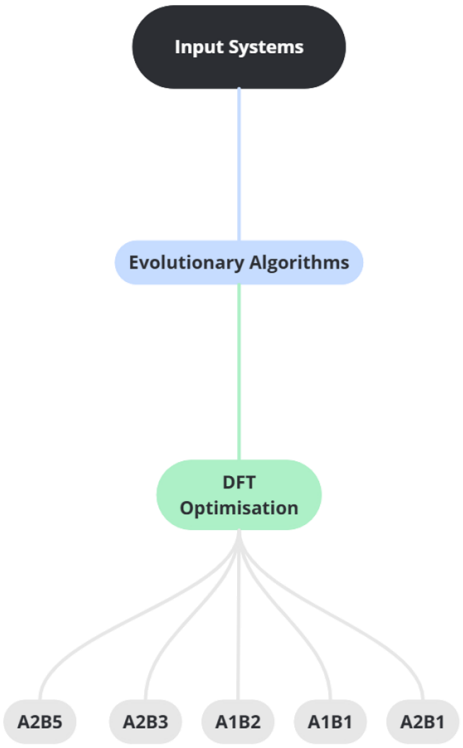
Additional Dictionaries

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

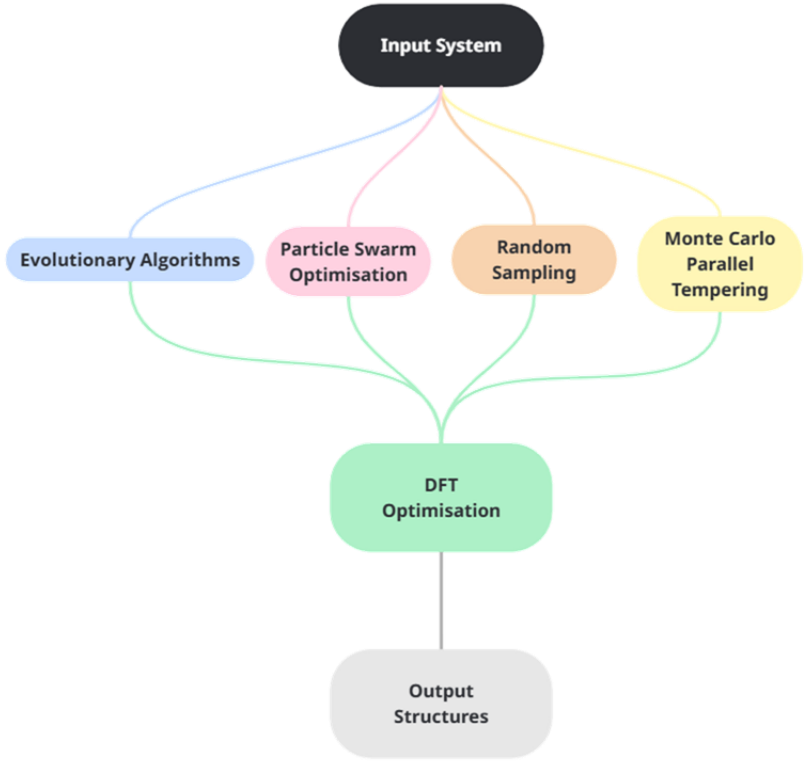
The CSP Dictionary

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

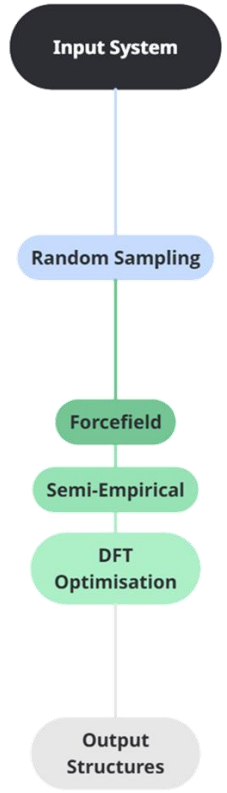
Variable Composition



Multiple Generation Methods



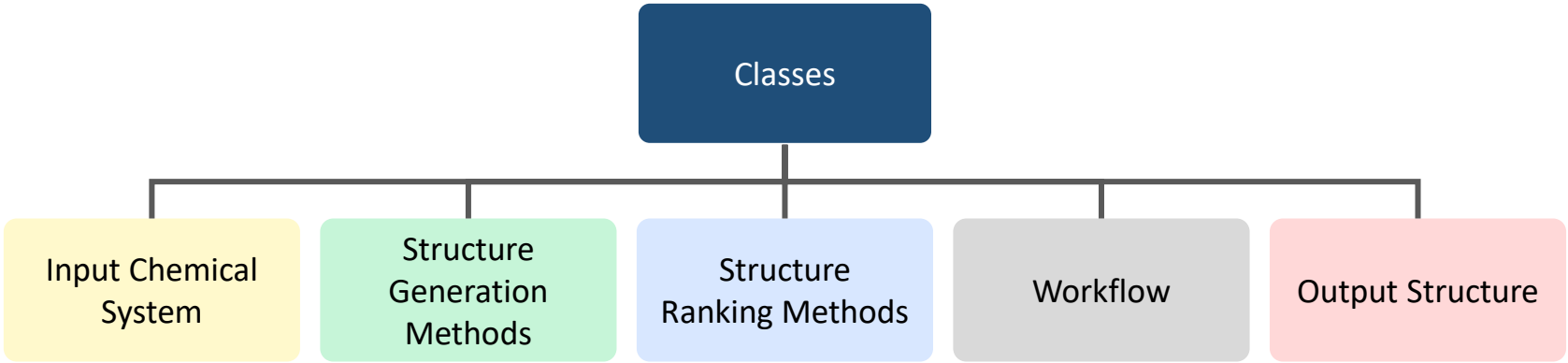
Multi-Step Optimisation

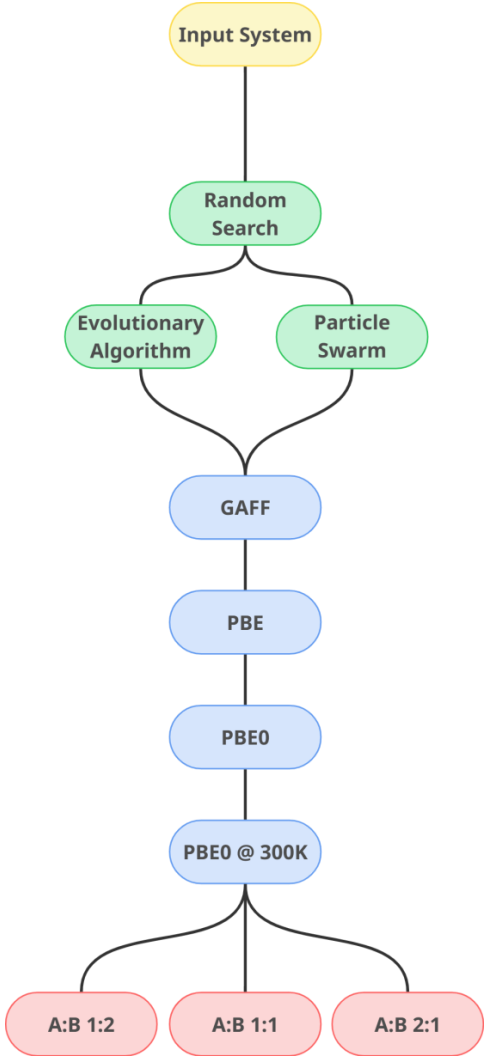


Organisation into Data Blocks

Each data block groups together all the data fields describing a single computational structure-prediction step or output structure, keeping parameters, metadata, and resulting properties organised in a clear, machine-readable way.

The associated classes specify the type of information contained in each data block, making it easy to understand what aspect of the CSP workflow the block represents.





Organisation into Data Blocks

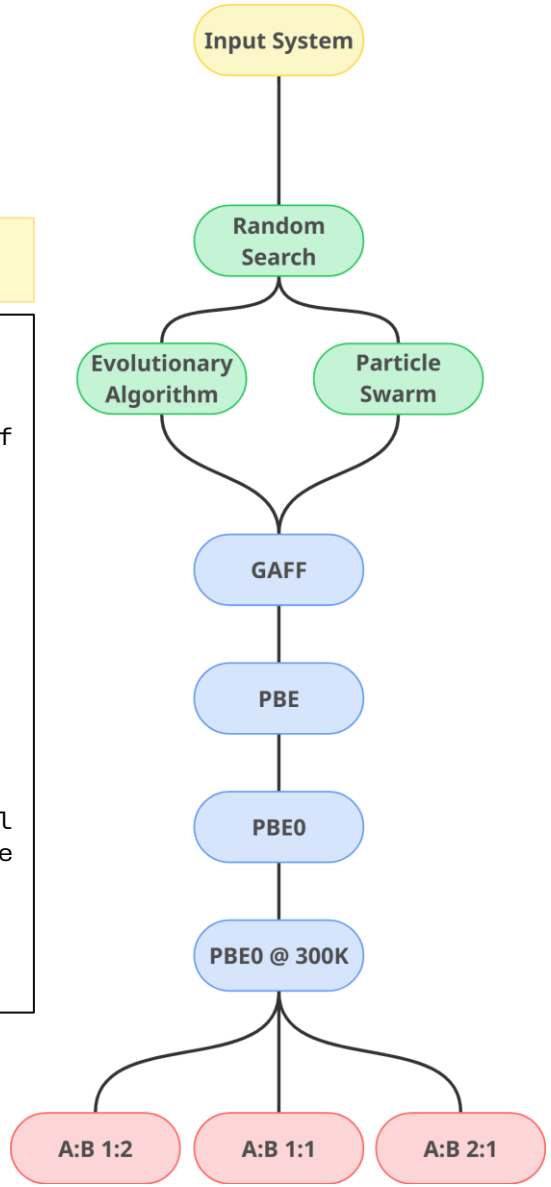
Each part of a CSP calculation is stored in its own data block.

Two essential fields in every data block:

- **Class**: Type of data block
  - Input
  - Generation Method
  - Ranking Method
  - Workflow
  - Theoretical Structure
- **ID**: Unique identifier (typically a UUID) used for linking across files.
- (Optional) A human-readable **description** that provides a concise label to make files understandable to users.

Input System

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

_csp.input_name              cannabinol_tetramethylpyrazine
_csp.input_identifier        BT-XXX

# Molecules
loop_
    _csp.input_molecular_entity_number
    _csp.input_molecular_entity_identifier
    _csp.input_molecular_entity_smiles
    _chemical.name_common
    1 CBN CCCCCc1cc(O)c2c(OC(C)(C)c3ccc([CH6])cc23)c1 cannabinol
    2 TMP Cc1nc(C)c(C)nc1C                          tetramethylpyrazine

# Type of Composition Search
_csp.input_composition_calculation  "fixed"
_csp.input_composition_coefficients [[2 1] [1 1] [1 2]]
```



Data Blocks: Input System

Label of the system:

- Common name
- Identifier (for internal database)

Identity of the system:

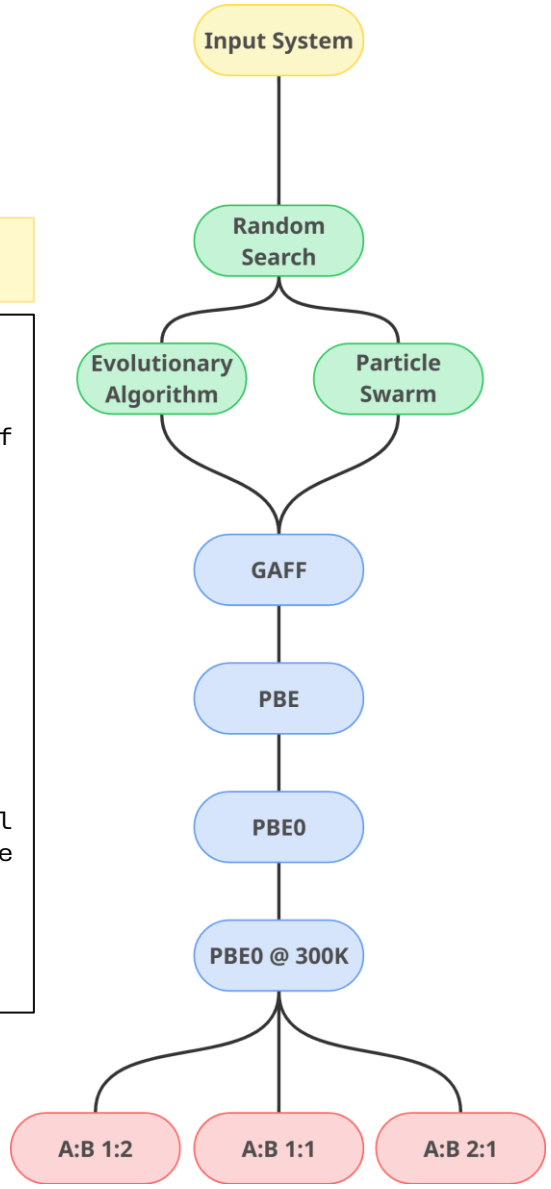
- **Inorganic CSP**: List of atomic species
- **Organic/MOF CSP**: List of Molecular Entities:
  - Identifier and SMILES string
  - Atom-level information and bonds definitions via the CIF Chemical Dictionary

Composition:

- Variable or fixed
- List of stoichiometries or allowed ranges

Input System

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

_csp.input_name                cannabinol_tetramethylpyrazine
_csp.input_identifier          BT-XXX

# Molecules
loop_
    _csp.input_molecular_entity_number
    _csp.input_molecular_entity_identifier
    _csp.input_molecular_entity_smiles
    _chemical.name_common
    1 CBN CCCCCc1cc(O)c2c(OC(C)(C)c3ccc([CH6])cc23)c1 cannabinol
    2 TMP Cc1nc(C)c(C)nc1C                          tetramethylpyrazine

# Type of Composition Search
_csp.input_composition_calculation  "fixed"
_csp.input_composition_coefficients [[2 1] [1 1] [1 2]]
```



Data Blocks: Generation Methods

Methods:

- Method class:
  - Monte Carlo Parallel Tempering
  - Evolutionary Algorithm
  - Random Sampling
  - …
- Single Method parameters

Search Space:

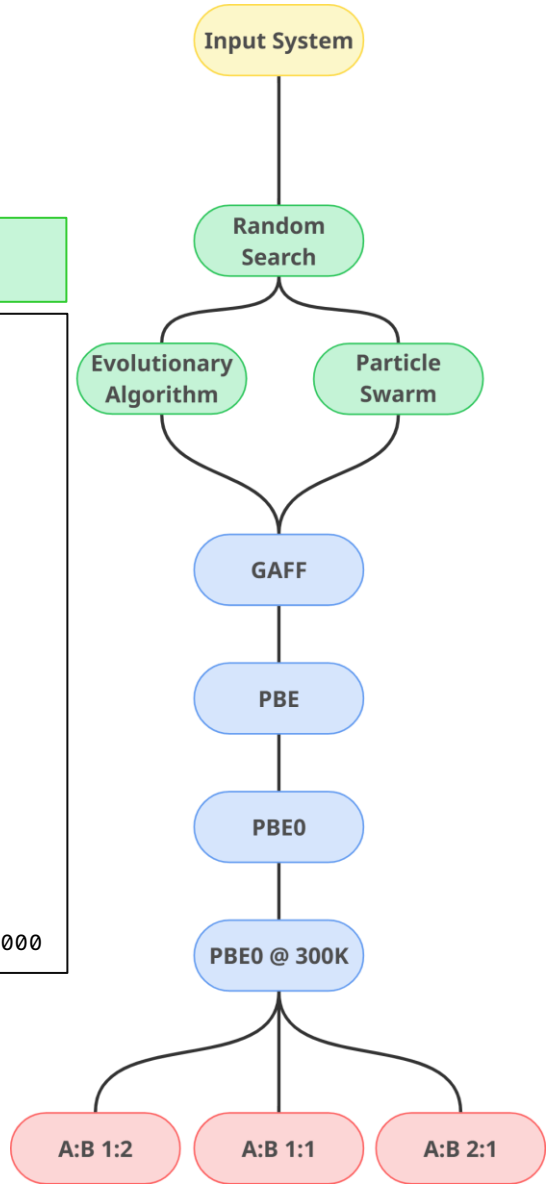
- Space groups
- Density limits

Stopping Criteria:

- Maximum structures evaluated.
- No improvement in low-energy structures
- Criteria can be per space group or global.

Particle Swarm Optimisation

```
data_pso
    # Datablock Details
    _csp.data_block_class       "Generation Method"
    _csp.data_block_id          242618a7-89b6-40d1-a6c4-cc51fd107a13
    _csp.data_block_description "pso"

    # Method Details
    _csp.structure_generation_method      "Particle Swarm Optimisation"
    _csp.particle_swarm_optimisation_population_size       50
    _csp.particle_swarm_optimisation_inertia_weight        0.7
    _csp.particle_swarm_optimisation_max_inertia_weight    0.9
    _csp.particle_swarm_optimisation_min_inertia_weight    0.4
    _csp.particle_swarm_optimisation_cognitive_coefficient 2
    _csp.particle_swarm_optimisation_social_coefficient    2

    # Search Space & Stopping Criteria
    _csp.structure_generation_density_lower_limit     1000
    _csp.structure_generation_density_upper_limit     1750
    _csp.structure_generation_space_group_number_list [14 2 15 61 19]

    _csp.structure_generation_stopping_criteria_description “Max Structures"
    _csp.structure_generation_stopping_criteria_max_structures_evaluated 100000
```



Data Blocks: Ranking Methods

Methods:

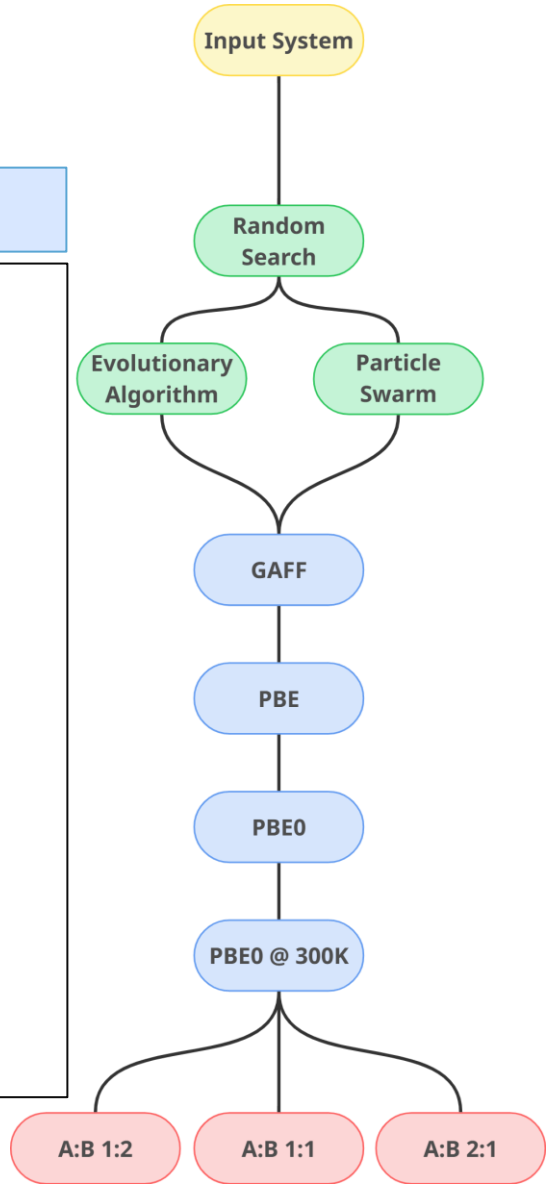
- Energy Model:
  - Forcefield
  - pDFT
  - Semi-Empirical
  - …
- Single method parameters

General Computational Chemistry Fields:

- Calculation Type:
  - Single-Point
  - Optimisation
  - Dynamic Ensemble
- Geometry Optimisation Settings:
  - Optimisation algorithm
  - Force convergence

PBE

```
# DFT
data_pbe
    # Datablock Details
    _csp.data_block_class                 "Ranking Method"
    _csp.data_block_description           "pbe"
    _csp.data_block_id                    97088479-72d3-4953-b858-503639748771

    # Energy Method Details
    _compchem.method                                          "pDFT"
    _dft.exchange_correlation_functional_type                 "GGA"
    _dft.exchange_correlation_functional_name                 "PBE"
    _dft.pseudopotential_type                                 "PAW"
    _dft.dispersion_correction                                "MBD"

    _dft.kinetic_energy_cutoff_wavefunctions                  600
    _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.calculation_type                                "Optimisation"
    _compchem.geometry_optimisation_algorithm                 "FIRE"
    _compchem.geometry_optimisation_cell                      "anisotropic"
    _compchem.geometry_optimisation_atoms                     "all"
    _compchem.geometry_optimisation_relax_force_convergence   0.001
    _compchem.geometry_optimisation_max_steps                 50
```



Data Blocks: Workflow

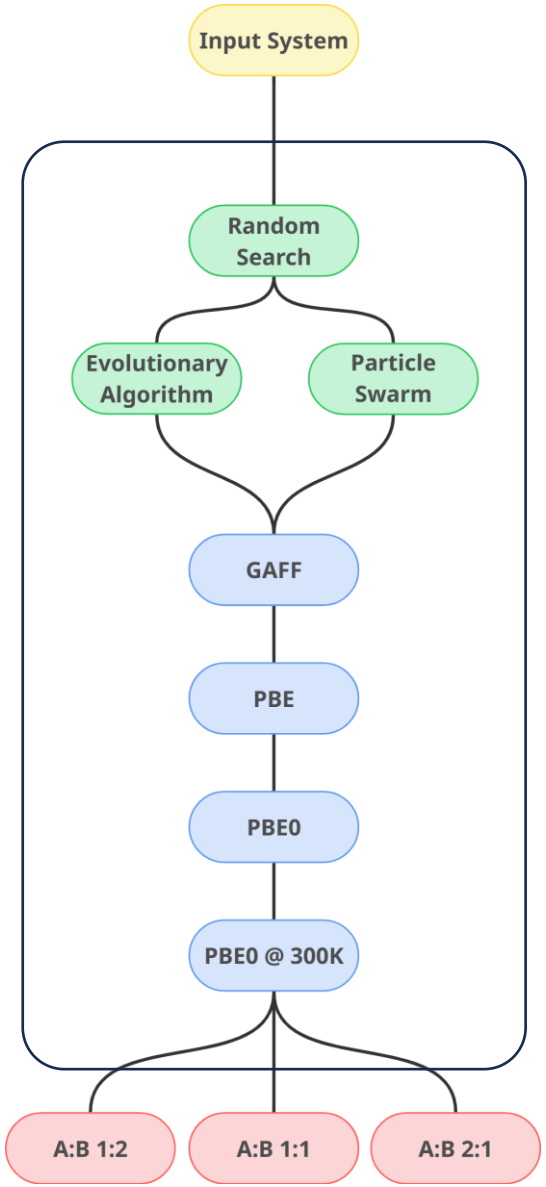
The **Workflow** datablock defines how different structure generation and ranking methods are connected into a full CSP workflow.

- Specifies the order of generation and ranking steps.
- Connects each step to the specific method data block via its unique ID.
- Tracks how structures flow from one ranking step to the next.

Additional data fields include descriptions and method names used for human-readability.

Workflow

```
data_workflow
    # Data blocks details
    _csp.data_block_class             Workflow
    _csp.data_block_description       csp_workflow_long
    _csp.data_block_id                425a77e3-a2a4-40fb-acdb-1e25b2d00bbd
    _csp.data_block_additional_files ["structure_generation_methods.cif"
                                      "structure_ranking_methods.cif"]

    # Structure Generation 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"                1d5b2b53-d70d-44…
        1 0 "ea"    "Evolutionary Algorithm"         af534fed-8153-4a…
        2 0 "pso"   "Particle Swarm Optimisation"    242618a7-89b6-40…

    # Structure Ranking Methods
    loop_
        _csp.structure_ranking_stage
        _csp.structure_ranking_preceding_stage
        _csp.structure_ranking_data_block_description
        _compchem.calculation_type
        _compchem.method
        _csp.structure_ranking_relative_energy_cutoff
        _csp.structure_ranking_max_structures_retained
        _csp.structure_ranking_data_block_id
        0 . "gaff" "Optimisation"    "Forcefield"  30.0 20000 6f36e2d3-a1…
        1 0 "pbe"  "Optimisation"          "pDFT"  10.0  2000 97088479-72…
        2 1 "pbe0" "Optimisation"          "pDFT"  10.0    50 97773a57-f2…
        3 2 "pbe0_qha" "Dynamic Ensemble" "pDFT"     .     . 2b9deed1-11…
```



Data Blocks: Theoretical Structure

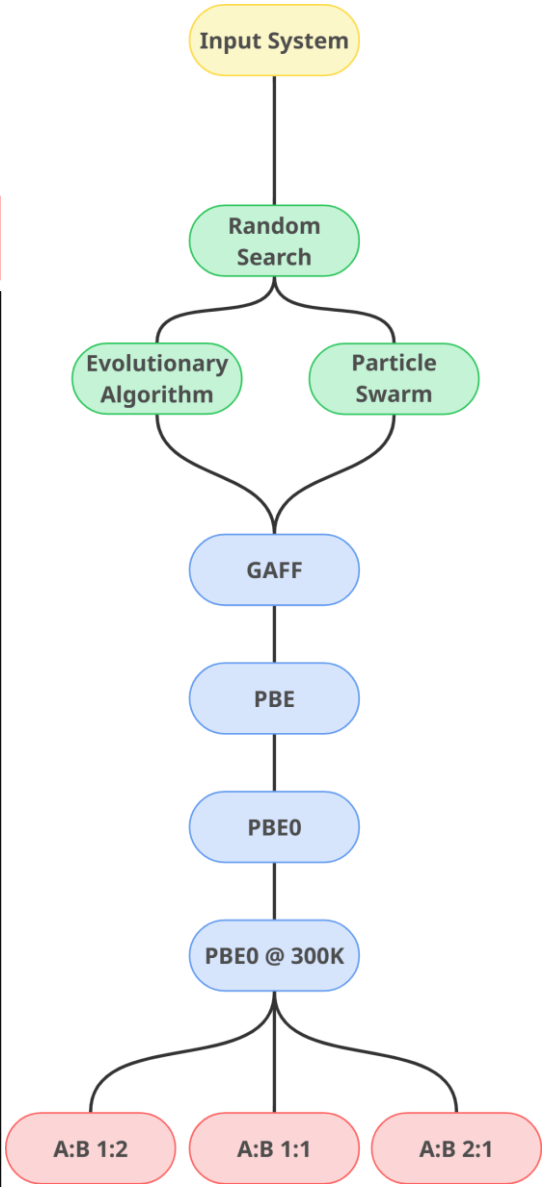
Each data block corresponds to a single output structure at a specific ranking stage (and specific temperature/pressure).

It records:

- The links to the input system, generation method, ranking method, and workflow.
- The physical properties of the structure.
- The crystallographic data.

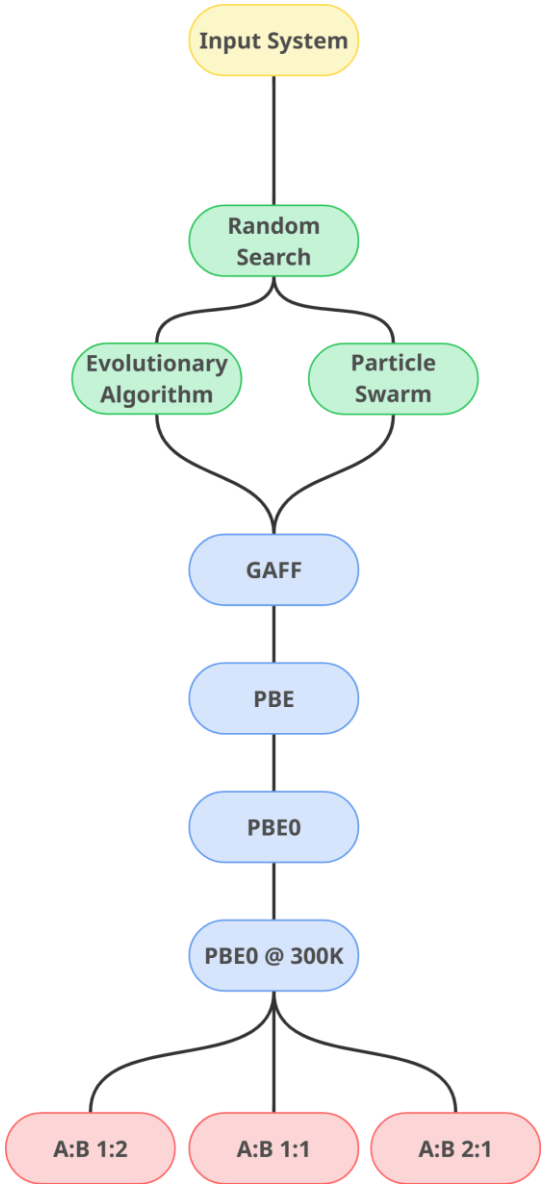
Theoretical Structure

```
data_structure_A1B1_1_step8_t300
    # Datablock Details
    _csp.data_block_class                                          "Theoretical Structure"
    _csp.data_block_description                                    A1B1_1_step4_t300
    _csp.data_block_id                                             3b4de7b5-aed4-43…

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

    # Stage identifiers
    _theoretical_structure.csp_input_system_description            BT-XXX
    _theoretical_structure.csp_input_system_id                     2a2611e3-2021-4b…

    _theoretical_structure.csp_workflow_description                csp_workflow_long
    _theoretical_structure.csp_workflow_id                         425a77e3-a2a4-40…

    _theoretical_structure.csp_generation_stage_description        ea
    _theoretical_structure.csp_generation_stage_id                 af534fed-8153-4a…

    _theoretical_structure.csp_ranking_stage_description           pbe0_qha
    _theoretical_structure.csp_ranking_stage_id                    2b9deed1-116f-45…
    _theoretical_structure.csp_reference_temperature               300.0
    _theoretical_structure.csp_reference_pressure                  100000.0

    _theoretical_structure.csp_previous_stage_structure_description A1B1_1_step3
    _theoretical_structure.csp_previous_stage_structure_id         094f5430-0785-4d…

    # Properties
    _theoretical_structure.temperature                             300
    _theoretical_structure.relative_free_energy                    0.284
    _theoretical_structure.rank                                    4

    # Crystal
    ...
```



Advantages of Using Data Blocks

- Modularity: Each component of a CSP workflow (input, generation method, ranking method, workflow, structure) is isolated in its own data block. This allows mix-and-match workflows without rewriting entire files.
- Clear Linking and Traceability: Every data block has a unique identifier (UUID) and enables the retrieval of the CSP workflow. This prevents ambiguity about what method or parameter set was used.
- Cross-File Organisation: Data blocks can be spread across multiple CIFs.
- Adaptability to new approaches: New CSP or computational chemistry methods can be supported simply by adding new data fields to their specific data block category.



Updates: Stopping Criteria

It is now possible to specify the maximum number of structures to be generated for each space group:

```
# Random Search
data_rs
    _csp.data_block_class                                          "Generation Method"
    _csp.data_block_description                                    rs
    _csp.data_block_id                                             8e0147be-0k54-44a1-a3bb-de7df26ddeer

    _csp.structure_generation_method                               "Random Search"
    _csp.random_search_random_numbers_algorithm_type               "Quasi-random"
    _csp.random_search_random_numbers_algorithm                    "Sobol"

    _csp.structure_generation_space_group_number_list              [14 2 15 61 19 4 33 29 5 1]
    _csp.structure_generation_stopping_criteria_description        "Max Structures"
    loop_
        _csp.structure_generation_stopping_criteria_space_group_number_list
        _csp.structure_generation_stopping_criteria_max_structures_evaluated
        14    1000
        2     1000
        15    1000
        61    1000
        19    1000
        4      500
        33     500
        29     500
        5      500
        1      500
```

Updates: Forcefield Parameters

The parametrisation of intramolecular, coulomb and van der Waals terms can now be described separately.

```
...
loop_
    _forcefield.parameterization_term
    _forcefield.parameterization_description
    _forcefield.parameterization_method
    intra         "Transferable parameters based on atom types" "GAFF"
    electrostatic "Fitting to gas-phase semi-empirical data"    "AM1/BCC"
    vdw           "Transferable parameters based on atom types" "AMBER"
...
```

Additionally, the `_forcefield.multipole_max_rank` data field has now been added to better describe multipoles approches

Updates: Data Fields

A series of data fields have been added or updated to the describe structure ranking methods, workflows and theoretical structures:

- ML potentials data fields:
  - `_ml_potential.method`
  - `_ml_potential.model`
  - `_ml_potential.precision`
- Describe how structures are removed in multistep ranking approaches:
  - `_csp.structure_ranking_relative_energy_cutoff`
  - `_csp.structure_ranking_max_structures_retained`
- Labels of some data fields have been changed:
  - "`predicted_structure`" was changed in the more appropriate "`theoretical_structure`"
  - "`molecule`" in "`molecular_entity`" so to include atoms, molecules, ions in the input definition
- Labels have been standardised to the CIF2/DDLm standard, having a single "." dividing group and subgroup.

Updates: A Computational Chemistry Dictionary

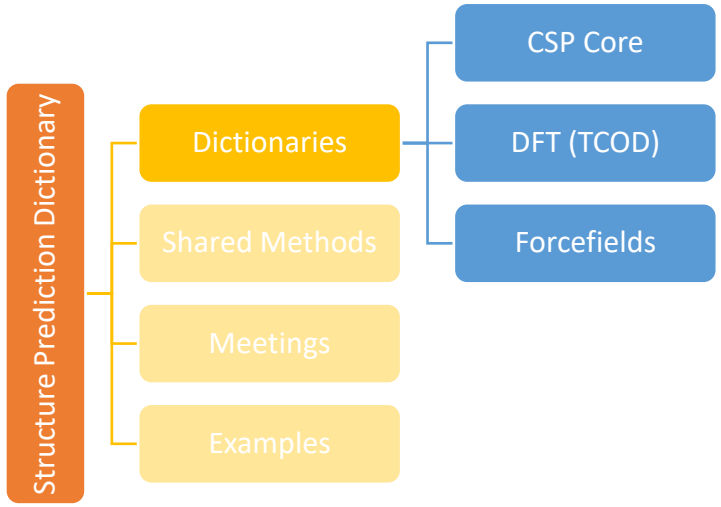
Several "_csp.structure_ranking_[]" data fields have been relabelled to "_compchem.[]".

The resulting data fields could be used as the basis for the development of a more general computational chemistry dictionary.

```
data_optimised_structure

# 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
…

# 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'
…
```

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:

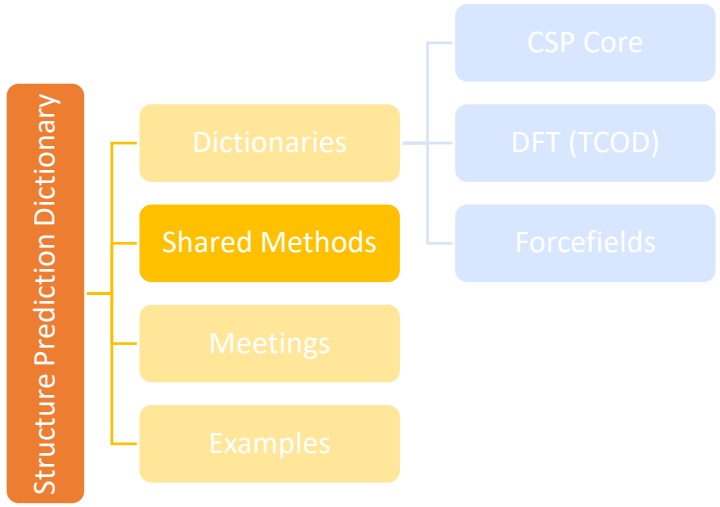
1. Input Chemical System

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

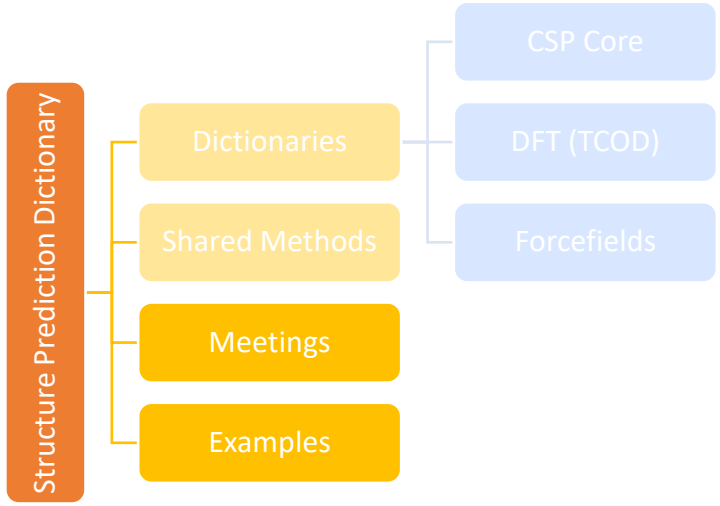


- We are actively seeking contributions from developers and end-users to help shape the final CSP dictionary. If your method requires additional settings, parameters, or properties that you believe should be represented, we warmly encourage you to propose new data fields.
- Our team will provide feedback and propose adjustments, while the development of new data fields will remain driven by your methodology.
- Our goal is to publish the final dictionary in a collaborative paper, and all contributors who participate meaningfully in its development will be included as co-authors.

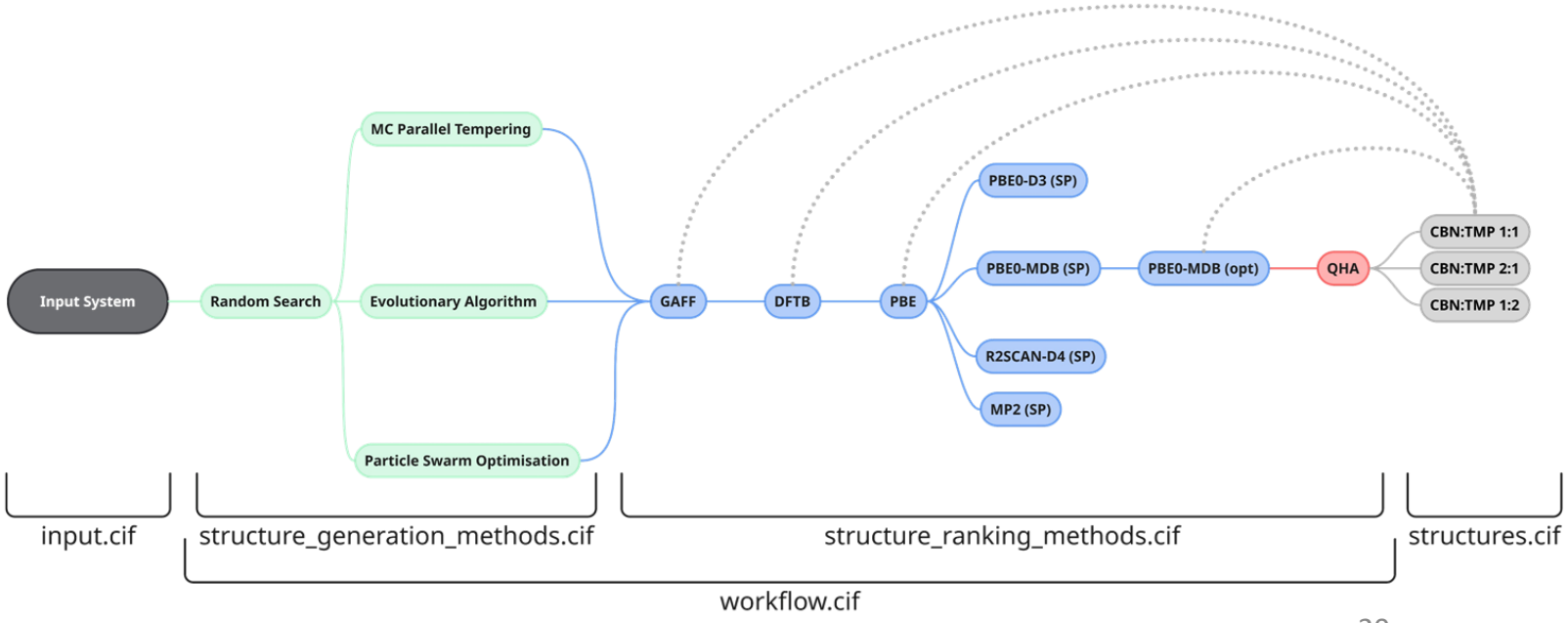
How to contribute:

- Create a New Issue
- Create a Pull Request
- Add comments to the PDF file and/or email to the CSP Data Standards Team
- Upload your method and related CIF files into the Shared Methods folder

The GitHub Repository: New files and folders

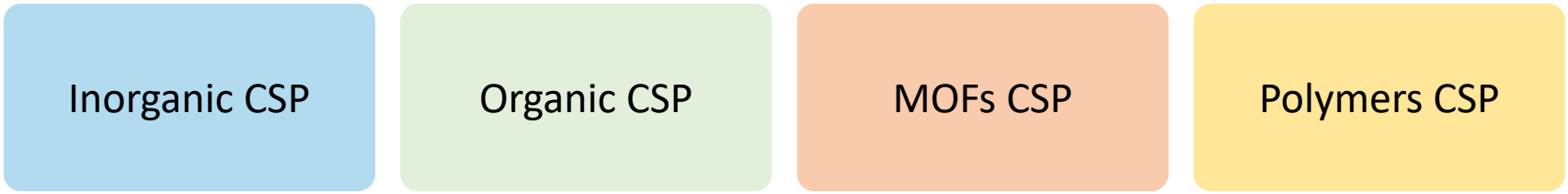


- **Meetings** include slides of previous meetings
- **CHANGELOG.md** file with summary of changes
- **Examples** folder with a series of case studies of CSP protocols



Missing Data Fields

Feedbacks from different CSP Communities:



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

- **ML-based structure generation** methods
- **Free Energy** methods
- **Molecular Dynamics** parameters
- **Conformers** generation
- Output structure **properties**
- **Constraints** (i.e. interatomic/intermolecular distances) and **search variables** in the generation methods (cell parameters, Z', atoms/molecules positions, molecule orientation, flexible torsional angles…)

Discussion: Output Data Block

Would it be worth having a general output data block containing statistical analysis and details on the output set?

- Computational cost of each step
- Statistics of properties across the landscape
- Global Minima and other relevant structures

Discussion: Links between structures

Would it be worth having linking structures based on specific features?

- Common geometric features (i.e. structure similarity, H-bond networks)
- Similarity with structures external to the CSP Dataset (link to experimental structures)

Next steps

3 Months

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

6-9 Months

- 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

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
Artem Oganov
Greg Price
Rui Guo
Jiri Klimes
Marc Meunier
Mihails Arhangelskis
Nikhar Rahul
Benjamin Tan
Jennie Martin