Genarris User Manual

This manual includes the documentation of different procedures and parameters Genarris accepts. For detailed guide to running Genarris, see the tutorial above.

Accepted procedures and their respective parameters sections:

|  |  |  |
| --- | --- | --- |
| Procedures | Parameter Sections | Description |
| Structure\_Generation\_Single | structure\_generation & structure\_generation\_single | Conducts batch structure generation |
| Structure\_Generation\_Batch | structure\_generation & structure\_generation\_batch | Conducts single structure generation |
| Harris\_Approximation\_Single | harris\_approximation\_single & FHI-aims | Conducts HA on a single structure |
| Harris\_Approximation\_Batch | harris\_approximation\_batch & FHI-aims | Conducts HA on a pool of structures |
| Harris\_Single\_Molecule\_Prep | harris\_single\_molecule\_prep &FHI-aims | Conducts single molecule calculations required for HA |
| Aims\_Single\_Run | aims\_single\_run &  FHI-aims | Runs FHI-aims on a single structure |
| Interatomic\_Distance  \_Evaluation | Interatomic\_distance  \_evaluation | Flexible evaluation of interatomic distances. |
| RDF\_Descriptor\_By\_Bin | rdf\_descriptor\_by\_bin | Evaluate structures’ RDF vector using discrete bins |
| RDF\_Descriptor\_By\_Point | rdf\_descriptor\_by\_point | Evaluate structures’ RDF vector using Gaussian distribution |
| K\_Mean\_Clustering | k\_mean\_clustering | Cluster and post-clustering select structures based on a vector property |
| RCD\_Calculation | rcd\_calculation | Gives relative coordinate vectors of closest neighbors to a given molecule |
| RCD\_Difference\_Folder\_Inner | rcd\_difference\_folder\_inner | Gives RCD distances of each structure to the others in vectors and/or a matrix |
| RCD\_Difference\_Compare\_Single | rcd\_difference\_compare\_single | Gives RCD distance vector of a given structure to the others |
| Find\_Duplicates | find\_duplicates | Given a folder of structures, outputs a folder of unique structures |
| FHI\_Aims\_Scavenge | fhi\_aims\_scavenge | Update structures if geometry.in.next\_step exists to restart the aims calculation (if necessary) from where it left off. |
| FHI\_Aims\_Extract | fhi\_aims\_extract | Extracts various data from FHI-Aims calculations (geometry, energy, volume, etc) to create json files of the structures. |
| Niggli\_Reduction\_Batch | niggli\_reduction\_batch | Apply Niggli reduction to many structures |
| Rcd\_difference\_calculation |  | What’s the difference between this and RCD\_Difference\_Folder\_Inner? |
| Vector\_Distance\_Calculation | vector\_distance\_calculation | Gives the distances (of the specified metric) of each structure to the others in vectors and/or a matrix |
| Reverse\_Harris\_Approximation | reverse\_harris\_approximation & FHI-aims | Determine the rotations necessary to obtain a given structure from a molecule and then perform the Harris Approximation (and output cube files if desired) |
| Reverse\_Harris\_Approximation\_Batch | reverse\_harris\_approximation\_batch & FHI-aims | Determine the rotations necessary to obtain given structures from a molecule and then perform the Harris Approximation (and output cube files if desired) |
| Cluster\_Based\_Selection | cluster\_based\_selection | Select desired number of structures with the max or min property evenly across clusters |
| Test\_Launch\_Parallel\_Run\_Single\_Inst | test\_launch\_parallel\_run\_single\_inst | Util to test out Flexible Parallel Parameters |
| Affinity\_Propagation\_Distance\_Matrix | affinity\_propagation\_distance\_matrix | Cluster using Affinity Propagation on a given coordinate system and output exemplar structures |
| Affinity\_Propagation\_Fixed\_Clusters | affinity\_propagation\_fixed\_clusters | AP that explores the setting of preference in order to generate desired number of clusters |
| Affinity\_Propagation\_Fixed\_Silhouette | affinity\_propagation\_fixed\_silhouette | Run AP on the given preference list and return the one within the tolerance of the target\_silhouette |
| Affinity\_Propagation\_Analyze\_Preference | affinity\_propagation\_analyze\_preference | Run AP with a list of preferences |
| Parallel\_Settings | parallel\_settings | Allows the user to select the number of cores per node, to load modules, and call other bash commands |

# Standard Parameter Bundles

This section documents the standard parameters used in many of the modules of Genarris. Instead of verbosely listing them under each section, they are identified and included by a comment next to the section name (e.g., SEMIP-SS applies).

## Standard Generation Parameters (SGP)

Used among generation modules to control the structures being generated.

### Core

**molecule\_path** (required): Path to the single molecule file of the molecule to be filled. The single molecule will have its COM moved to the origin.

**NMPC** (required): Number of molecules per unit cell. Only space groups with compatible symmetry multiplicity can be selected.

**is\_chiral** (optional; Boolean): If set to TRUE, forces the randomly selected space group to be chiral. Otherwise, randomly selected space group will be racemic.

### Unit Cell Volume Control

**ucv\_target (**required): unit cell volume in terms of A3

**ucv\_ratio\_range** (optional; default: [1.0, 1.0]): A list of two floats indicating how much a generated structure’s UCV can deviate from *ucv\_target* by proportion.

**ucv\_std** (optional): If set, enables half-normal sampling to bias unit cell volume selection towards smaller value. This allows more attempts to be made with smaller UCV, where molecules are less likely to fit in.

### Symmetry (Space Group and Wyckoff Selection) Control

In addition to *NMPC* and *is\_chiral*, the following parameters determine the list of space groups from which the generator will randomly select:

**space\_groups\_allowed** (optional): If specified, is a list of integers indicating the space groups that can be selected. Specified space groups will be subjected other compatibility check determined by the other parameters in this section.

**wyckoff\_list** (optional; default: [0]): If specified, is a list of integers indicating the Wyckoff position to be selected for a given space group. Default of [0] forces the molecule to always be placed on the general Wyckoff position. Space groups whose sum of Wyckoff position multiplicity do not add up to NMPC will not be selected.

### Unit Cell Lattice Control

The following parameters determine how much a unit cell can deviate from a perfect cube.

**angle\_range** (optional; default: [30,150]): A list of two floats that indicate the range where angles not 90 degree must fall in.

**ax\_variance** (optional; default: [0.5, 2]): Confines the variance of the x (principle) component of vector. Given as a proportion to the cube root of the unit cell volume. Note that the y and z component ofis by default set to 0.

**by\_variance** (optional; default: [0.5, 2]): Confines the variance of the y (principle) component of vector. Given as a proportion to the cube root of the unit cell volume. Note that the z component ofis by default set to 0.

**cz\_variance** (optional; default: [0.5, 2]): Confines the variance of the z (principle) component of vector. Given as a proportion to the cube root of the unit cell volume.

### Closeness Criteria Control

**com\_dist** (optional): Enables calling structure\_handling.COM\_distance\_check. A lower-bound for the distance between center of mass of the molecules.

**atom\_dist** (optional): Enables calling structure\_handling.atom\_distance\_check\_1. A lower-bound for the distance between atoms of the different molecules.

**specific\_radius\_proportion** (optional): Enables calling structure\_handling.specific\_radius\_check. This value (shortened as *sr*) should provide the proportion of the atomic radius that two atoms from different molecules are required to be. The atomic radius information is read from utilities.radius. Atoms not in the radius list will not be compared. By default, the radius stored is the van der Waals radius of atoms. E.g., the van der Waals radius of carbon is 1.70 Å, and that of nitrogen is 1.55 Å; thus if sr=0.75, then any pair of intermolecular C-N contact must be longer than (1.70+1.55)\*0.75=2.44 Å.

**specific\_radius\_std** (optional, even when *specific\_radius\_proportion* is present): To increase the diversity of the pool, the user may choose to set this value to make the sr for each check fuzzy. This allows a half-Gaussian sampling of sr with *specific\_radius\_proportion* as the center, where sr smaller than *specific\_radius\_proportion* will be selected with probability density based on the left half of a Gaussian distribution. This value determines the standard deviation of the spread. Note that since a structure is more likely to pass a specific radius test if *sr* is small, the distribution of selected *sr* will not reflect the distribution of the final generated pool. E.g., if the desired sr range is 0.5 to 0.9, then then this value can be set as small as 0.1 to bias towards higher sr in order for a more even distribution of sr of the actually generated structures.

**specific\_radius\_lower\_bound** (optional, even when *specific\_radius\_std­* is present): This sets the lower bound beyond which a randomly half-Gaussian sampling of sr will be rejected.

**custom\_radii** (optional; default: {}): A dictionary mapping atom species to radii in Angstrom that overrides the default van der Waal radii of atoms used in specific radius check.

**custom\_radii\_by\_pairs** (optional; default: {}): A dictionary mapping tuples of two atom species to radii in Angstrom that overrides the default van der Waal radii of atoms used in specific radius check.

### Structure ID Control

**struct\_id\_scheme** (optional; default: ["struct\_index","\_","random\_index"] ): Determines the format of automatically assigned struct\_id. Special keywords are: struct\_index, which inserts the index of the generated structure, and random\_index, which inserts a random index. In addition, any entry in the structure’s property dictionary can also be inserted. Any other entry will be simply concatenated. An example will be [“struct\_index”, “\_”, “random\_index”, “\_”, “space\_group”, “\_”, “NMPC”], which can yield “0\_178ab318\_14\_4” as struct\_id.

**struct\_index\_length** (optional; default: 0): Sets the length of automatically filled structure index, which is reached by filling 0 at front.

### Generation Counter Limits

This is a list of parameters that can be specified to override the default counters (number of times Genarris tries to generate a structure that passes the closeness check before re-selecting certain randomized values or considering an attempt as failed)

**attempt\_limit** (optional; default: 1024): Total number of times Genarris tries before quitting an attempt with failure

**attempts\_per\_space\_group\_change** (optional; default: 64): Number of trial structures generated before randomizing a new space group. Note that this will trigger unit cell and Wyckoff list change as well.

**attempts\_per\_unit\_cell\_change** (optional; default: 16): Number of trial structures generated before randomizing a new UCV and a corresponding set of lattice vectors.

**attempts\_per\_wyckoff\_list\_change** (optional; default: 4): Number of trial structures generated before randomizing a new wyckoff\_list

**fill\_molecule\_attempt\_limit** (optional; default: 5): While filling the molecule, Genarris does is recursively through all the Wyckoff positions in the wyckoff\_list. This parameter specifies the number of fails to fill molecules on a specific Wyckoff position without violating closeness criteria before returning and redoing the last step.

## Standard Structure Output Parameters (SSOP)

Used among evaluation modules that outputs a single or a collection of structure(s).

Parameters:

**output\_dir**: Directory to write structure(s) to.

**output\_format**: (optional; must be one of “json”, “geometry” or “both”; default: “both”): Output format. “json” triggers output to output\_dir/struct\_id.json. “geometry” triggers output to output\_dir/struct\_id.in. “both” triggers both.

**SSOP-EXT** is the bundle that includes an additional parameter:

**output\_info\_file** (optional): If present, triggers output of evaluation result to the path specified.

## Standard Evaluation Modules Input Parameters (SEMIP)

Used among evaluation modules that load a single structure or a collection of structures.

Parameters:

**structure\_path**: Path to the structure file

**structure\_dir**: Directory to a collection of structures.

**structure\_dir\_depth** (optional; default: 0): Depth of the structure dir

**structure\_suffix** (optional; default: “”): Suffix of structure files to be read. Make sure to add a necessary dot (“.”) as structures read in from Geometry format will be named by the filename – structure\_suffix.

Note: All structures will be loaded by attempting both Json and Geometry format.

Based on the module nature, the SEMIP parameter bundle is split into three kinds:

**SEMIP-SS** applies to **single-structure evaluation modules** (e.g., Niggli\_Reduction\_Batch). One and only one of structure\_path and structure\_dir is to be specified to avoid confusion.

**SEMIP-PW** applies to **pair-wise evaluation modules** (e.g., RCD\_Difference\_Calculation). structure\_dir must be specified to input a collection. The presence of structure\_path indicates that the operation is only to be done between this structure and the rest of the structures in the collection.

**SEMIP-CW** applies to **collection-wide evaluation modules** (e.g., Affinity\_Propagation\_Distance\_Matrix). Must specify structure\_dir. structure\_path will be ignored.

## Simple Parallelism Parameter (SPP)

Used among modules where simple parallelism through multiprocessing is implemented.

Parameter:

**processes\_limit** (optional; default: Genarris\_master.processes\_limit): Number of parallel processes allowed to multiprocessing.

## Flexible Parallelism Parameters (FPP)

Used among modules that require flexible parallelism, potentially launching Genarris master through mpirun or other batch commands.

Parameter:

**parallel\_run\_command** (optional; default: ["python", "$MASTER", "$CONF"]): Command to launch multiple Genarris master. $MASTER will be replaced with path to Genarris master deduced from python \_\_file\_\_ metadata. $CONF will be replaced with path to temporary configuration file.

**parallel\_run\_instances** (optional; default: 1): Number of times to run the parallel command. If the parallelism is already inherent in the command (e.g., mpirun with -n set to desired number of instances), then this should be set as 1.

**tmp\_dir** (optional; default: inherited from Genarris\_master section): Temporary directory to store ui.conf and log files for parallel instances.

**tmp\_dir\_clean\_up** (optional; Boolean): Whether or not to clean-up the tmp files generated by this run after. The tmp\_dir itself will not be removed for safety.

## Affinity Propagation Parameters (APP)

**dist\_mat\_input\_file**: Path to distance matrix file computed earlier

**affinity**\_**type**: (Optional) A list specifying the type of conversion (either “exponential” or “power” are supported) from the given distance matrix to the Affinity matrix in the first element and a scalar in the second element. The form of “exponential” is affinity\_mat[i][j] = - np.exp(dist\_mat[i][j]\*affinity\_type[1]), and the form of “power” is affinity\_mat[i][j] = - dist\_mat[i][j]\*\*affinity\_type[1]. Default: [“exponential”,1]

**damping**: (Optional) Damping factor (between 0.5 and 1) is the extent to which the current value is maintained relative to incoming values (weighted 1 - damping). This in order to avoid numerical oscillations when updating these values (messages). Default: 0.5

**convergence**\_**iter**: (Optional) Number of iterations with no change in the number of estimated clusters that stops the convergence. Default: 15

**max\_iter**: (Optional) Maximum number of iterations. Default: 200

**preference**: (Optional) float or None. If preference is a float, note that decreasing preference gives more clusters and vice versa. If preference is None or not used, they will be set to the median of the input similarities. Default: None

**property**\_**key**: (Optional) The key at which the cluster number is stored among the structure properties dictionary. Default: AP\_cluster

**output**\_**file**: (Optional) Path and filename of outputted information about the clustering results: number of exemplars selected, average distance to centers, standard deviation of distance to centers, standard deviation of number of structures in clusters, Silhouette score, center and number of structures of each cluster, and assigned cluster label and distance to center. Default: ./AP\_cluster.info

**exemplars**\_**output**\_**dir**: (Optional) The path to the outputted exemplar structures with their cluster number now in the *properties* dictionary.

**exemplars\_output\_format**: (optional; must be one of “json”, “geometry” or “both”; default: “both”): Output format. “json” triggers output to output\_dir/struct\_id.json. “geometry” triggers output to output\_dir/struct\_id.in. “both” triggers both.

# Section Specific Parameters

**All Booleans are expected to function as Flags. When set, the value must be “TRUE” (all caps, without quotation marks). When not set, the Booleans default to False.**

**All lists must be set in Python syntax (e.g., [“a”, “b”]).**

## Section: Genarris\_master

**procedures:** a list of strings that denotes the name of functions callable under Genarris\_master

**working\_dir** (optional):a string of working\_directory. If not present, Genarris\_master will interpret the working directory as the parent directory of the instruction file. The procedures called by Genarris\_master will inherit working\_dir, unless overridden by the option in their respective section.

**master\_log\_path** (Default: Genarris\_master.working\_dir/Genarris.log): Path to the log that records all the activity below.

**master\_err\_path** (Default: Genarris\_master.working\_dir/Genarris.err): Path to the error log for this run.

**info\_level** (Default: 1): Level of verbosity of the information printed to master\_log\_path. 0 = no information; 1 = general high-order information (e.g., successful record of batch generation); 2 = individual incident information included (e.g., a successful generation of a structure); 3 = debug mode (all information included)

**group\_permission** (optional): If present, needs to be set to TRUE, and enables the calling of “chmod g=u” on files or “chmod -R g=u” on folders created by the current incidence of Genarris. Note: certain lower end function that does not interact with this instruction object may create files that by default are granted the same group permission as user permission for job security (e.g., execute.info that records the last time stamp for FHI-aims binary call).

**script\_path** (Default: os.path.realpath(\_\_file\_\_)): Path to the currently running Genarris\_master.py. No need to modify.

**master\_node** (Default: socket.gethostname()): The node that is currently running this Genarris instance

**processes\_limit** (optional; default: 1): Number of parallel processes allowed to multiprocessing. SPP bundles of other procedures inherit from this value if not set.

## Section: FHI-aims

Master section with detailed information about how FHI-aims should be submitted

**aims\_bin\_path**: if any of the procedures requires calling FHI-aims, then the path or alias should be specified here.

**execute\_command**: command used to execute an aims binary. Currently accepting: mpirun, runjob.

**execute\_command\_source** (optional): If enabled, will instead call sh to interpret the execute command’s source file.

**execute\_style** (optional; default: “safe\_subprocess”): “fast\_subprocess” calls the job with subprocess.call, and should be only used if certain that the aims calculation will be successful and will not get hung. “safe\_subprocess” uses subprocess.Popen and carefully monitors the progress of the job through tracking the output. “fast\_system” means that an os.system call will be used to run the binary. “safe\_system” puts an ampersand at the end of the command and tracks the job through output, but unlike “safe\_subprocess”, will not be able to determine if the job was hung, and kill the job if yes (due to inability to track job process). “submit” means that the execution of the binary will be submitted as an individual job, in which case a submission script will be established. Currently only supports “subprocess” and “system”. Both “fast\_subprocess” and “fast\_system” will not allow *write\_active* to be enabled.

**multiple\_launches** (default: 1): If *execute\_style* is set to one where job progress tracking is possible, specifies the number of attempts to launch the job (without seeing any output) before returning failure. Note: *aims\_single\_run* will not relaunch the job if the job has started outputting but got hung and killed. It is up to a higher order process to determine if the calculation has failed, and needs to be rerun.

**call\_interval** (optional): If present, specifies the number of seconds between two consecutive aims binary call. Interval enforced by get\_execute\_clearance(). By default, the last time aims binary is called is recorded in “*Genarris\_master.working\_dir*/execute.info”; this path, however, can be overridden by *execute\_info\_path*.

**execute\_info\_path** (optional, even if *call\_interval* is present; default: “Genarris\_master.working\_dir‌/‌execute.info”): Path where the last time stamp for aims binary call is recorded.

**runjob\_modes** (default: 4): Number of modes per node if using runjob command.

**runjob\_thres** (default: 4): Number of threads if using runjob command.

## Section: aims\_single\_run

A function that specifically carries out one FHI-aims calculation. Should have minimum interaction with user.

**working\_dir**: A necessary working directory to call the FHI-aims calculation at. Will report error if missing; will not be inherited from the Genarris\_master section.working\_dir.

**structure\_path** (optional): if present, will copy the file to *working\_dir*.

**structure\_format** (optional, necessary if structure\_path is present): determines the format of the input structure. “json” for a json file, and “geometry” for an aims file.

**control\_path** (optional): if present, will copy the control.in file to *working\_dir*.

**output\_name** (optional): if present, will be used as the filename of the aims output. Otherwise, the output will be called “aims.out”.

**write\_active\_enabled** (optional): if present, should be set to “TRUE” and will enable output of current time (in seconds from epoch) to a file. By default, outputs the time to “*working\_dir*/active.info”. This path may be overridden by *write\_active\_path*. write\_active will not work if “fast\_subprocess” or “fast\_system” are called (as the commands are run and waited upon).

**write\_active\_path** (optional, even if *write\_active* is enabled): If present, will be path where the current time is written to. Note: should be full path.

**write\_log\_path** (optional): If present, should be the full path where the progress of the aims\_single\_run should be recorded.

**write\_log\_name** (optional): A name used to indicate the current FHI-aims call when writing the log file. If not present, and *write\_log\_path* is specified, use the *working\_dir* to denote this instance of FHI-aims call.

**launch\_time\_out** (optional): If present, and *execute\_style* is set to any option that tracks the progress of an aims job, specifies the number of seconds waited before a launch of aims job is determined to have failed. Otherwise, if monitoring demands, the default time out is 60 seconds.

**update\_poll\_interval** (optional, necessary if job tracking is enabled by FHI-aims.execute\_style): Specifies the number of seconds as the interval between checking the output file. Very necessary to be set appropriately according to the estimated time per job. Recommendations: 1 second for Harris approximation, 5 seconds for fast single point evaluation, 10 seconds for fast relaxation, 30 seconds for long jobs.

**update\_poll\_times** (optional, necessary if job tracking is enabled by FHI-aims.execute\_style): The number of times polling the output file without update before determining the job to be hung. Thus the overall no-update time-out is *update\_poll\_interval\*update\_poll\_times*. Very necessary to be set appropriately according to the estimated time per job.

Below is a list of options underneath this section to specify the correct command to run aims:

**mpirun\_processes** (optional): if present and execute\_command==”mpirun”, will set the value as the number of processes spawned by the mpirun (-n flag).

**mpirun\_hosts** (optional): If present (as a list of strings denoting hostname) and execute\_command==”mpirun”, will specify the host in the mpirun command. WARNING: certain version of mpirun (e.g. the version that comes with anaconda 2.1.0) fails to fully comply with this setting. Format: [“hostname1”, “hostname2”, . . .]

**runjob\_block** (optional, necessary if *execute\_command*==”runjob”): The block on Mira to run the aims instance on.

**runjob\_corner** (optional): The corner in the specific block on Mira to run the aims instance on.

**runjob\_shape** (optional, necessary if *corner* is present): The shape of the corner in the specific block on Mira to run the aims instance on.

**runjob\_nodes (**optional, necessary if *execute\_command*==”runjob”): The size of the block or corner where the aims instance is run.

**additional\_arguments** (optional): additional arguments that will be pended to the end of argument list if present. String will be literally evaluated.

## Section: structure\_generation\_single (SGP applies)

This module makes a single attempt at generating a structure. No additional parameters are required other than the SGP bundle.

## Section: structure\_generation\_batch(SGP, SSOP and FPP applies)

This module handles batch generation of structures.

**number\_of\_structures**: The total number of structures to be generated in this collection

**number\_of\_attempts** (optional; default: None): The total number of attempts to generate the amount of structures specified above. Each attempt consists of multiple sub-attempts (see point 7 of SGP bundle). Default of None sets no attempt limit.

**index\_tracking\_file** (optional; default: None): Path to the index tracking file of a previously started run. Set in order to continue from a previous run. Default of None indicates that this module is called to start from beginning.

## Section: \_structure\_generation\_batch (SGP and SSOP applies)

This is an internal module to parallelize structure\_generation\_batch and should not be called explicitly by user.

## Section: harris\_approximation\_batch

**working\_dir** (optional): The temporary working directory where the calculation is done. By default, Genarris\_master.working\_dir/tmp

**structure\_dir**: The folder where the structure is located in

**structure\_dir\_copy** (Boolean): If present, will enable the structure\_dir to be copied to the temporary working directory with a “tmp\_” attached as a prefix.

**structure\_dir\_depth** (optional; default: 0): Depth to search the structure\_dir. If set to 0, then only search for structure files directly under the structure\_dir.

**structure\_suffix** (optional): Required suffix for the structure to be pulled in, i.e., any structure must have the file name with this suffix in order to be evaluated by this module. Make sure to add a “.” if necessary.

**structure\_format**: The format in which the structure is recorded. “json” for a json file, and “geometry” for an aims file.

**processes\_limit** (optional): The amount of parallel instances of harris\_approximation\_single to be run. By default, inherits the processes\_limit from Genarris\_master. Note: Additional processes will be spawned by the aims calculation. Recommend setting harris\_approximation\_batch.‌mpirun\_processes carefully so that the product of processes\_limit and mpirun\_processes does not exceed the processes the nodes can hold.

**mpirun\_processes** (optional; default: 1): The value used to set the –n flag of mpirun

**structure\_output\_dir** (optional): If present, enables extracting of energy on the structures and the dumping of structures into structure\_output\_dir.

**energy\_name** (optional): See the same option under harris\_approximation\_single

**exec\_control\_path**: See the same option under harris\_approximation\_single

**molecule\_harris\_folders**: See the same option under harris\_approximation\_single

**molecule\_name**: See the same option under harris\_approximation\_single

**enantiomer\_name**: Name of the molecule enantiomer to be outputted for rotations.in. WARNING: make sure this name matches the folder names within molecule\_harris\_folders.

**molecule\_path** (optional): See the same option under harris\_approximation\_single

**molecule\_format** (optional, required if molecule\_path is present): See the same option under harris\_approximation\_single

**full\_reverse\_napm** (optional): See the same option under harris\_approximation\_single.

**full\_reverse\_prep\_control** (optional; required if full\_reverse\_napm is specified): See the same option under harris\_approximation\_single.

**molecule\_match\_tolerance** (optional, even if either any of the reverse scheme is enabled; default: 1): See the same option under harris\_approximation\_single.

**rotations\_title** (optional; default: “molecule”): See the same option under harris\_approximation\_single

**clean\_up** (optional; default: “1”): For this batch module, an additional level “5” is allowed. This cleans up all the directories made to run individual structures. The harris\_approximation\_single module inherits a level “4” if “5” is specified. For “0” through “4”, see the same option under harris\_approximation\_single.

**mpirun\_processes** (optional): See the same option under aims\_single\_run.

**update\_poll\_interval** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run.

**update\_poll\_time** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run.

## Section: harris\_approximation\_single

**working\_dir** (optional): Otherwise interpreted as the os.path.dirname(structure\_path)

**structure\_path**: path to the structure that needs to be reversed.

**structure\_format**: determines the format of the input structure. “json” for a json file, and “geometry” for an aims file.

**structure\_output\_path** (optional): The path where the evaluated structure will be dumped. By default, the energy of the structure evaluated by the HA will be stored as “energy”. Use energy\_name to override that option.

**energy\_name** (optional, even if structure\_output\_path is present; default: “energy”): The key to assign the energy extracted from the HA run in the structure’s properties dictionary.

**exec\_control\_path**: control.in to run the Harris Approximation on the crystal structure

**molecule\_harris\_folders** (required, unless full reverse scheme is triggered): A list of path to folders where the single molecules that are to be called by the harris\_approximation is located. WARNING: upon copying, the original folder name will be preserved. Make sure to match molecule\_name and enantiomer\_name to the desired folder name.

**molecule\_name**: Name of the molecule to be outputted for rotations.in. WARNING: make sure this name matches the folder names within molecule\_harris\_folders.

**enantiomer\_name**: Name of the molecule enantiomer to be outputted for rotations.in. WARNING: make sure this name matches the folder names within molecule\_harris\_folders.

**molecule\_path** (optional): Enables the reverse\_harris\_info\_prep to prepare structures that don’t have the harris\_info ready. If this is not provided, and full reverse scheme is not enabled (see full\_reverse\_napm), then a structure without harris\_info will not be evaluated, and an non-fatal error will be recorded.

**molecule\_format** (optional, required if molecule\_path is present): The format in which the molecule is recorded. “json” for a json file, and “geometry” for an aims file.

**full\_reverse\_napm** (optional): Enables the full reverse scheme when the structure cannot be mapped by the standard given molecule. Under this scheme, the first molecule will be taken out from the structure to be harris\_single\_molecule\_prep and mapped on the other structures.

**full\_reverse\_prep\_control** (optional; required if full\_reverse\_napm is specified): The necessary control.in file to prepare the first molecule taken from the structure as in the full reverse scheme.

**molecule\_match\_tolerance** (optional, even if either any of the reverse scheme is enabled; default: 1): The tolerance in the molecule matching process for any of the reverse scheme.

**rotations\_title** (optional; default: “molecule”): Assigns a name to the title line of the rotations.in file.

**clean\_up** (optional; default: “4”): Determines the level of clean up automated after a Harris Approximation run. If set at “1”, cleans up only the folders containing molecules’ information. If set at “2”, search and clean up the restart binary (by default, titled “restart.new000”). If set at “3”, cleans up everything in the “harris” folder but “aims.out”, “geometry.in”, and “control.in” files. If set at “4”, removes the entire “harris” folder

**mpirun\_processes** (optional): See the same option under aims\_single\_run.

**update\_poll\_interval** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run.

**update\_poll\_time** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run.

## Section: harris\_single\_run

Core of the execution of Harris Approximation, having everything prepared in the working\_dir. Not used by user.

**working\_dir**: If all other optional parameters are absent, enter into automatic mode of looking for rotations.in underneath this working\_dir, calling aimsrotate to creating a folder called harris, and running the calculation in it.

**control\_path**: Path to the control file that runs the rotated restart

**mpirun\_processes** (optional): See the same option under aims\_single\_run.

**update\_poll\_interval** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run. Note that this default is not set in the original module.

**update\_poll\_time** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run. Note that this default is not set in the original module.

## Section: harris\_single\_molecule\_prep

**molecule\_path**: Path to the single molecule file to run the calculation on. The single molecule will have its COM moved to the origin.

**molecule\_format**: The format in which the molecule is recorded. “json” for a json file, and “geometry” for an aims file.

**control\_path**: Path to the control.in file that should have turned on the flags that would ask aims to output the necessary files for Harris Approximation.

**output\_dir**: Directory to run the single molecule calculation in. The resulting files will thus be found in this folder.

**enantiomer\_output\_dir** (optional): if present, will do the preparation on the enantiomer of the molecule as well in this second directory. The enantiomer will be found by first placing the COM of the molecule at the origin and flipping the sign of the z coordinate.

**mpirun\_processes** (optional): See the same option under aims\_single\_run.

**update\_poll\_interval** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run. Note that this default is not set in the original module.

**update\_poll\_time** (optional; necessary if job tracking is enabled by FHI-aims.execute\_style): See the same option under aims\_single\_run. Note that this default is not set in the original module.

**additional\_arguments** (optional): See the same option under aims\_single\_run.

## Section: interatomic\_distance\_evaluation (SEMIP-SS, SSOP-EXT & SPP applies)

The module, evaluation.diversity\_analysis.interatomic\_distance\_evaluation calculates the interatomic distances between different molecules of a certain species pair. Sample output: [[struct\_id\_1, [pair\_1\_atom\_1\_index, pair\_1\_atom\_2\_index, pair\_1\_distance], [pair\_2\_atom\_1\_index, pair\_2\_atom\_2\_index, pair\_2\_distance], [struct\_id\_2, . . .], . . . ]. Atom indexing starts from 0. Returns [-1, -1, -1] for a specific pair if no pairs are found or not enough pairs are found to have the ith closest pair (see atomic\_pairs parameter). When the index is equal or greater than the number of atoms in the entire cell, then it should be found in the super cell. By default, a 3\*3\*3 supercell is constructed.

**napm**: Number of atom per molecule. Set this value to 1 if pairs of atoms from the same molecule are also to be considered.

**atomic\_pairs**:Instructions of atomic pairs to be found. E.g., [[“C”,”N”,1], [“N”, “X”, 2], [12, “H”, 4]] asks for three interatomic distances: 1) the closest C-N pair, 2) the second closest pair between N and an arbitrary element, and 3) the 4th closest H atom to atom number 12 in the structure. Use “X” as a species name for an arbitrary atom. Both element can be set as “X”. Use a number (smaller than total number of atoms in the cell) to fix one atom by index. First atom in the structure is indexed as 0.

**allow\_all\_pairs** (optional; Boolean): Be default, only the first molecule will be selected as the source of the first atom in the pair. The default setting is ideal if all molecules in the unit cell can be considered equivalent. By turning on this flag, all molecules will be considered as the source of the first atom.

**block\_same\_pair** (optional; Boolean): By default, if [“C”, “N”, 3] is found among the atomic\_pairs option, all pairs of C-N contact from the different molecules are sorted to find the 3rd closest contact. However, if this flag is turned on, then each pair of molecule can only contribute one atom pair in sorting.

**structure\_extension** (optional): By default, a 3\*3\*3 supercell is constructed if allow\_all\_pairs is not set to TRUE, and a 2\*2\*2 supercell is constructed if allow\_all\_pairsis set to TRUE. This parameter overrides the default by specifying a list of extensions. Each extension is itself a list of three integers. E.g., [1, 0, 2] indicates an extension through translation by lattice\_vector\_a\*1+​lattice\_vector\_b\*0​+​lattice\_vector\_c\*2. Sample valid input: [[1, 0, 2], . . . ]

**property\_name** (optiona; default: “interatomic\_distance”): Key to store the interatomic distance as a property.

## Section: rdf\_descriptor\_by\_bin (SEMIP-SS, SSOP-EXT & SPP applies)

This module computes the rdf of a structure as a vector, with bin size specified by user.

**atomic\_pairs**: List of atom pairs to sample the radial distribution function. The first in each pair is the reference atom, and the second is the target atom. Sample input: [[“C”,”N”], [12,”S”], [“N”, “X”]] first asks for the radial distribution function of (1) carbon and nitrogen, (2) atom 12 and sulfur, and (3) nitrogen and arbitrary atom. Atom indexing starts from 0.

**distance\_range**: A list of two floats indicating the distance range in angstrom. E.g., [1.0, 5.0] indicates that the rdf vector sets up bins between 1 and 5 angstrom.

**bins** (optional; default: 10): If bins is an int, it defines the number of equal-width bins in the given range (10, by default). If bins is a sequence, it defines the bin edges, including the rightmost edge, allowing for non-uniform bin widths.

**normalized** (optional; Boolean): If present, normalizes each entry of the rdf vector by dividing it by the bin size (in terms of volume).

**property\_name** (optional; default: rdf\_vector): The key at which the rdf function is stored among the structure *properties* dictionary.

## Section: rdf\_descriptor\_by\_point (SEMIP-SS, SSOP-EXT & SPP applies)

This module computes the rdf between two atom type, X and Y, as follows:

Outputs the sampled point as a vector, with leading term as the struct\_id. Concatenates if multiple atomic pairs are specified in *atomic\_pairs*.

**atomic\_pairs**: List of atom pairs to sample the radial distribution function. The first in each pair is the reference atom, and the second is the target atom. Sample input: [[“C”,”N”], [12,”S”], [“N”, “X”]] first asks for the radial distribution function of (1) carbon and nitrogen, (2) atom 12 and sulfur, and (3) nitrogen and arbitrary atom. Atom indexing starts from 0.

**sampled\_distances**: A list of distances r where the rdf function will be calculated. E.g., [2, 3, 4, 5]. User can input using the numpy.linspace function. E.g., numpy.linspace(1,3,5) samples 5 evenly spaced points between 1 and 3 (1, 1.5, 2, 2.5, 3).

**smoothing\_parameter** (optional; default: 0.5): The smoothing parameter *B* in the calculation.

**normalized**: (optional; Boolean): If present, normalizes each entry of the rdf vector by dividing it by the area of a sphere with radius of the sampled distance.

**property\_name** (optional; default: rdf\_vector): The key at which the rdf function is stored among the structure *properties* dictionary.

## Section: k\_mean\_clustering (SEMIP-CW, SSOP-EXT and SPP applies)

This method calls the KMeans method of sklearn.cluster to conduct K-means clustering on a pool according to a vector descriptor.

**n\_clusters**: Number of clusters to generate (k-value).

**init** (optional; default: “k-means++”):Method of initialization. ‘k-means++’: selects initial cluster centers for k-mean clustering in a smart way to speed up convergence. ‘random’: choose k observations (rows) at random from data for the initial centroids.

**n\_init** (optional; default: 10): Number of time the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n\_init consecutive runs in terms of inertia.

**max\_iter** (optional; default: 300): Maximum number of iterations of the k-means algorithm for a single run.

**tol** (optional; default: 1e-4): Relative tolerance with regards to inertia to declare convergence.

**vector\_property\_name**: Specifies the name (key) of the property (observation vector) within the structure properties dictionary to be used for clustering. The property pointed to should be a vector (list).

**cluster\_property\_name** (optional; default: “k\_mean\_cluster\_index”): Specifies the name (key) of the property within the structure properties dictionary to be used for storing of the clustering result. This property will be a list. The latest clustering result is stored in the last entry of the list.

**distance\_property\_name** (optional; default: None): The key under which the distance from centroid will be stored. Default of None disables storing this property.

**pcs\_enabled** (optional; Boolean): If present, enables post clustering selection. This module will then output a tuple of two lists: original collection with the clustering information and the selected collection.

**pcs\_method** (optional; necessary if *pcs\_enabled* is set to TRUE): If present, enables post clustering selection. Specifies the method to be used for selection. Methods include: “property\_max”, “property\_min”. For “property\_max” and “property\_min”, the *pcs\_reserved\_variable* should specify the name (key) of the property used to rank the structures within the same cluster.

**pcs\_number** (optional; necessary if *pcs\_enabled* is set to TRUE): Number of structures to be selected from in the original selection based on the clustering.

**pcs\_reserved\_variable** (optional, even if *pcs\_enabled* is set to TRUE): see *pcs\_method* for what the reserved variable may stand for.

**pcs\_output\_dir**: (optional, even if *pcs\_enabled* is set to TRUE): Directory where the selected structures are printed.

**pcs\_output\_format**: (optional; default: “.json”): Format to output selected structures. See output\_format option in SSOP bundle.

**pcs\_output\_info\_file**: (optional): if presents, allows printing to this file the diversity information (as internal pair comparison) of the selected pool.

## Section: rcd\_calculation (SEMIP-SS, SSOP-EXT & SPP applies)

**NAPM**: Number of atoms per molecule.

**axes**: Define two reference axes with format [(a,b),(c,d)] where a, b, c, and d are 0-based integer indices of atoms in the list of atoms a molecule in the geometry file. The vectors (a,b) and (c,d) form two of the reference axes, and the third is calculated automatically as their cross-product. These axes are orthogonalized and normalized, but should be chosen such that they are as perpendicular as possible.

**close\_picks** (optional; default: 16): Type an integer, N, which is the number of molecules with closest COM positions to a given molecule to construct the RCD vector of the structure. N should be sufficiently large to capture the environment of a molecule in a crystal. Default: 16.

**property\_name** (optional; default: rcd): The key at which the RCD vector is stored among the structure *properties* dictionary.

## Section: rcd\_difference\_calculation

**structure\_dir**: The input path to the structures for the RCD vectors are already calculated.

**structure\_format**: The format in which the structure is recorded. “json” for a json file, and “geometry” for an aims file.

**stored\_property\_key**: The key at which the RCD vector is stored among the structure *properties* dictionary. Default: RCD\_vector

**contribution\_ratio**: Real number, k, defines the weight assigned to the orientation difference (as opposed to the COM position difference). Default: 1

**select\_pairs**: Integer. When calculating the RCD distance between those two structures (v1, and v2), select\_pairs of molecules from v1 will each be compared to each molecule in v2 (len(v2) = close\_picks). Then, the select\_pairs pairs with the smallest RCD distances are used to compute the RCD distance between v1 and v2. Default: 8.

**disable\_enantiomer**: (Optional). Boolean. Default: FALSE. If TRUE, then s2 will be mirror reflected by flipping the sign of its third relative coordinate. This is necessary b/c the third reference axes is generated with cross product, and thus will acquire a sign flip under mirror reflection.

**diff\_list\_output**: (Optional). Path to the output file containing arrays of distance information of every structure to every other structure in the pool: [i, j, distance], where i and j are indices of the structures compared. See diff\_matrix\_output for this output in matrix form.

**diff\_matrix\_output**: (Optional). Path to output the n x n RCD distance matrix to, where where n is the number of structures in the pool. Thus, this matrix contains the distance of every structure to every other structure. Instead of using this option, you could run combine.py after setting structure\_suffix and structure\_dir (they have the same definitions as in this Section) and rcd\_output\_file which is the path to the outputted matrix.

## Section: rcd\_difference\_compare\_single

**structure\_dir**: The input path to the structures for the RCD vectors are already calculated.

**structure\_format**: The format in which the structure is recorded. “json” for a json file, and “geometry” for an aims file.

**stored\_property\_key**: The key at which the RCD vector is stored among the structure *properties* dictionary. Default: RCD\_vector

**contribution\_ratio**: Real number, k, defines the weight assigned to the orientation difference (as opposed to the COM position difference). Default: 1.

**select\_pairs**: Integer. When calculating the RCD distance between those two structures (v1, and v2), select\_pairs of molecules from v1 will each be compared to each molecule in v2 (len(v2) = close\_picks). Then, the select\_pairs pairs with the smallest RCD distances are used to compute the RCD distance between v1 and v2. Default: 4.

**disable\_enantiomer**: (Optional). Boolean. Default: FALSE. If TRUE, then s2 will be mirror reflected by flipping the sign of its third relative coordinate. This is necessary b/c the third reference axes is generated with cross product, and thus will acquire a sign flip under mirror reflection.

**diff\_list\_output**: (Optional). Path to the output file containing arrays of distance information of every structure to every other structure in the pool: [i, j, distance], where i and j are indices of the structures compared. See diff\_matrix\_output for this output in matrix form.

## Section: find\_duplicates (SEMIP-CW and SSOP applies)

## Section: fhi\_aims\_scavenge

**tmp\_folder**: Input folder path where incomplete aims jobs are being run or ran

**wait\_time**: Integer. Do not scavenge if the last update to the aims job was less than wait\_time seconds ago. Default: 15.

**disable\_cleanup**: Boolean. If TRUE, will remove sub-directories of tmp\_folder.

## Section: niggli\_reduction\_batch (SEMIP-SS and SSOP applies)

**NAPM**: Number of atoms per molecule.

## Section: vector\_distance\_calculation

**structure\_dir**: Path to the folder containing structures for which to calculate distance vectors and/or matrices. The vector corresponding to the key, vector\_property\_name, in the structure *properties* dictionary must already have been calculated.

**vector\_property\_name**: The key in the structure *properties* dictionary containing a vector of coordinates (do not have to be euclidean coordinates). E.g. RDF\_vector

**distance\_type**: The distance metric to use. cosine, euclidean, and normalized euclidean are currently supported types. Default: euclidean

**diff\_list\_output**: (Optional). Path to the output file containing arrays of distance information of every structure to every other structure in the pool: [i, j, distance], where i and j are indices of the structures compared. See diff\_matrix\_output for this output in matrix form.

**diff\_matrix\_output**: (Optional). Path to output the n x n RCD distance matrix to, where where n is the number of structures in structure\_dir. Thus, this matrix contains the distance of every structure to every other structure.

**processes\_limit**: (optional; default: 1): Number of parallel processes allowed to multiprocessing.

## Section: reverse\_harris\_approximation

**structure\_path**: Path to the structure geometry file to perform the reverse Harris scheme on.

**structure\_format**: The format of the input structure. “json” for a json file, and “geometry” for an aims file.

**prep\_control\_path**: Path to the control.in file that should have turned on the flags that would ask aims to output the necessary files for Harris Approximation and perform SCF energy evaluation (for a single molecule).

**exec\_control\_path**: control.in to run the Harris Approximation on the crystal structure

**NMPC**: Number of molecules per unit cell

**NAPM**: Number of atoms per molecule.

## Section: reverse\_harris\_approximation\_batch

**working\_dir**: Path to the folder containing all structures to perform the reverse Harris scheme on

**structure\_format**: determines the format of the input structure. “json” for a json file, and “geometry” for an aims file.

**structure\_suffix** (optional): Required suffix for the structure to be pulled in, i.e., any structure must have the file name with this suffix in order to be evaluated by this module. Make sure to add a “.” if necessary.

**prep\_control\_path**: Path to the control.in file that should have turned on the flags that would ask aims to output the necessary files for Harris Approximation and perform SCF energy evaluation (for a single molecule).

**exec\_control\_path**: control.in to run the Harris Approximation on the crystal structures

**NMPC**: Number of molecules per unit cell

**NAPM**: Number of atoms per molecule.

## Section: cluster\_based\_selection (SEMIP-CW and SSOP applies)

number\_selected: The desired number of structures to select with the max or min property (evenly across clusters)

select\_max: TRUE if desire to use the max property, FALSE if the min property

cluster\_property\_name: Set a name for the cluster property in the properties dictionary

preference\_property\_name: Set a name for the preference property in the properties dictionary

## Section: affinity\_propagation\_distance\_matrix (APP, SEMIP-CW, and SSOP-EXT applies)

## Section: affinity\_propagation\_analyze\_preference (APP, SEMIP-CW, SSOP-EXT, and SPP applies)

**preference\_list**: List of preferences to use for AP calculation.

**verbose\_output** (optional; Boolean): Whether to output structure-specific cluster info for each preference analyzed.

## Section: affinity\_propagation\_fixed\_clusters (APP, SEMIP-CW, and SSOP-EXT applies)

**num\_of\_clusters**: The integer number of clusters to have Affinity Propagation find.

**preference\_range**: Initial range of preference values given as a length 2 python list with element 0 being the lower bound and element 1 being the upper bound. E.g. [-1000, 0]

**num\_of\_clusters\_tolerance**: Integer. If the difference between the number of clusters generated and num\_of\_clusters is at most num\_of\_clusters\_tolerance, then treat the clustering as a success. default: 0

**max\_sampled\_preferences**: The number of iterations to attempt finding an acceptable number of clusters before exiting.

**output\_without\_success**: Boolean (e.g. TRUE) to allow clustered pool output even if target num\_of\_clusters is not hit within tolerance

## Section: affinity\_propagation\_fixed\_silhouette (APP, SEMIP-CW, SSOP-EXT and SPP applies)

**target\_silhouette**: The silhouette score to have Affinity Propagation aim for.

**silhouette\_tolerance**: If the difference between the silhouette score achieved and target\_silhouette is at most silhouette\_tolerance, then treat the clustering as a success. default: 0.01

**preference\_list**: List of preferences to use for AP calculation.

**output\_without\_success**: Boolean (e.g. TRUE) to allow clustered pool output even if target num\_of\_clusters is not hit within tolerance

## Section: parallel\_settings

**job\_scheduler**: Only SLURM is a supported option currently. *job\_scheduler* is used to determine a nodelist that the parallel jobs will run on. Default: SLURM

**modules\_launch**: Python-style list of strings which are the modules to load with the “module load” command. E.g. ["anaconda","gcc/4.8.2"]

**additional\_commands**: (Optional) String any additional bash commands to run. E.g. "export LD\_PRELOAD=/share/apps/gcc/4.7.4/lib64/libgfortran.so.3"

**processes\_per\_node**: The integer number of cores per node to utilize.

## Section: fhi\_aims\_extract (SSOP applies)

**calculation\_dir**: Parent directory of aims calculations

**aims\_output\_file**: File to look for in each execution folder as FHI-aims output. default: aims.out

**original\_structure\_file**: File to look for in each execution folder as original structure in JSON format. default: original\_structure.json

**original\_structure\_dir**: Directory where additional original structures are stored outside of calculation folder. Structures are matched according to struct\_id and calc folder's basename. default: None

**fail\_if\_no\_energy** (optional; Boolean): Whether or not to fail a particular subdirectory extraction if no energy update is extracted from aims output

**energy\_property\_name**: Property name to store the structure. default: energy

**fail\_if\_no\_geometry\_next\_step** (optional; Boolean): Whether or not to fail a particular subdirectory extraction if no geometry.in.next\_step file is found.

**check\_execution\_complete**: Whether or not to look for \_execution\_complete file that indicates execution complete

**strict\_success**: Boolean. Whether or not to look for "Have a nice day" to determine execution success (vs. "Leaving FHI-aims")

**clean\_up**: Boolean. Whether or not to clean up successfully extracted folder; files to be cleaned up include: control.in, geometry.in, geometry.in.next\_step, aims output, oiginal structure file, \_execution\_complete. If the folder is empty after removing these files, removes the folder as well.

Here are some parameters that may conflict and the resolution for each:

* Genarris\_master.**working\_dir** and working\_dir under any other section: The working\_dir set underneath the section Genarris\_master will be overridden by any other instance of working\_dir.
* FHI-aims.**execute\_style** and aims\_single\_run.**write\_active**: if the former is set to “fast\_subprocess” or “fast\_system”, the present of the latter will not trigger the writing of active file.