GENOA User's Manual
The GENerator of reduced Organic Aerosol mechanism

Version 1.0

Zhizhao Wang, Florian Couvidat, Karine Sartelet
zhizhao.wang@enpc.fr

April 25, 2022

Contents

List of Code Listings

This manual is part of the electronic supplement of the article "GENerator of reduced Organic Aerosol mechanism (GENOA v1.0): An automatic generation tool" (cite).

1 Installation

1.1 Requirements

GENOA requires Python 3.5 or later[1]. Please also make sure that python3 library numpy 1.11.0 or later[2] has been appropriately installed, and it can be executed with the command "python3".

Furthermore, GENOA is coupled with the 0D aerosol model SSH-aerosol[3] [1]. Please read the User's manual[4] of ssh-aerosol v1.2, and make sure all the requirements for SSH-aerosol have been appropriately installed. For the compilation of SSH-aerosol, the construction tool SCONS[5] is required as well.

Optional python3 libraries matplotlib 1.5.1 or later[6] and basemap 1.2.1 or later[7] are required for post-processing the testing results and draw the error map.

1.2 Download

All codes and data related to GENOA and its application to the reduction of the Beta-caryophyllene mechanism can be downloaded directly from Zenodo[8]. The latest development of GENOA can be found in github[9].

1.3 Compilation of SSH-aerosol

If the user has not previously used SSH-aerosol, we recommend first trying to compile SSH-aerosol before running GENOA. To do that, please copy the genoa version of SSH-aerosol into a test directory and run the commands in Listing 1 after installing all the requirements. This GENOA version of SSH-aerosol is developed from SSH-aerosol v1.2.1 and has been adapted to use

[1] `https://www.python.org/`

[2] `https://numpy.org/`

[3] `https://sshaerosol.wordpress.com/`

[4] `http://cerea.enpc.fr/ssh-aerosol/user_manual.pdf`

[5] `http://www.scons.org/wiki/SconsTutorial1`

[6] `https://matplotlib.org/`

[7] `https://matplotlib.org/basemap/`

[8] `needtoupdate`

[9] `https://github.com/tool-genoa/GENOA`

in GENOA (remove most of the outputs and modify the computation of gas-phase chemistry in order to use different chemical mechanisms).

If SSH-aerosol is successfully compiled, the user should see the following text by the end of the standard output: *"scons: done building targets."*.

```
# copy the GENOA version of SSH-aerosol to the test directory [test]
cp -rf GENOA/src/files/ssh-aerosol-genoa [test]/.
# go to the ssh-aerosol folder for testing
cd [test]/ssh-aerosol-genoa
# clean the previous compilation files (if need)
./clean
# command to compile SSH-aerosol. If success, the last printed message will be "scons: done building targets.".
./compile
```

Listing 1: Compiling SSH-aerosol-GENOA

2 Package Structure

The downloaded package contains three sections: the source code of GENOA, the input data required to run the reduction of Beta-caryophyllene (BCARY) mechanism from the Master Chemical Mechanism (MCM) [2], and one example reduction result generated with the given input data.

The major folders/files include:

- ***GENOA***, the GENOA code
  - ***src***, a folder contains the source code of GENOA.
    - ***files***, a folder contains the files required for generating chemistry files and run SOA simulations.
      - ***ssh-aerosol-genoa*** the GENOA version of ssh-aerosol
  - ***inputs_bcary***, a folder contains the input files required to run the BCARY reduction (except for the input files for conditions, they are saved separately in other folder *conditions_bcary* due to the large size).
    - ***rdc_cfg.ini***, the configuration file of GENOA. The configuration options are detailed in Sect. 4.1.

    - ***BCARYorg***, a chemical folder contains the reference mechanism for BCARY reduction generated from the BCARY degradation mechanism of the Master Chemical Mechanism v3.3.1[10] (hereafter referred to as the MCM mechanism). In this reference mechanism, very fast degraded species with a kinetic rate coefficient of more than 1 s^{-1} have been 'jumped' in the scheme to avoid numerical stiffness. The structure and usage of the chemical files inside the folder are explained in Sect. 2.1.
    - ***dataset.list***, a list of testing dataset for the BCARY reduction, where the testing conditions are presented by identifiers explained in Sect. 2.2.2.
  - ***examples***, a folder contains files for post-processing.
    - ***map_testing.py***, a python script for plotting the error map from the testing result. Python3 libraries *matplotlib* and *basemap* are required to run this script. An example of the generated figure (*Testing_R6R.png*) is provided.
    - ***clean.py***, a python script for cleaning the output of GENOA. The use of this file is explained in Listing 3.

- ***conditions_bcary***, a folder contains all the input files related to the atmospheric conditions to run SOA simulations with SSH-aerosol. See Sect.2.2.2 for more information.

- ***results_bcary_example***, a folder contains the example results generated by GENOA from the BCARY reduction.
  - ***CaseRdc_R***, a record of the entire training process. See more details in Sect. 3.2.
  - ***BCARYR6R***, the final reduced SOA mechanism trained with the input data and the given configuration file. This mechanism is reported in the paper as the "Rdc." mechanism.
  - ***Testing_BCARYR6R_all***, a record of the testing result of the reduced mechanism BCARYR6R. The errors of total SOA concentrations between simulations with the reduced mechanism (*BCARYR6R*) and with the reference mechanism (*BCARYorg*) on the testing dataset are listed in this file. This file can be post-processed with the python script *map_testing.py*.

[10] `http://mcm.york.ac.uk/MCMv3.3.1`

2.1 Chemical mechanism generated by GENOA

2.1.1 Nomenclature

During the reduction, for each potential reduction attempt, GENOA will generate a chemical folder with the name of this mechanism *[chem_name]* (e.g., *inputs_bcary/BCARYorg*). The name for the reduced chemical mechanism [chen_name] is build from **[prefix] + [chem_id] + [suffix]**.

[prefix] and [chem_id] are assigned in the configuration file and fixed for the entire reduction. [suffix] is generated automatically in GENOA. For the first to the 62nd reduction steps, the suffix is one tail letter from the ASCII letters ('a' to 'z' + 'A' to 'Z', in total 52 letters) and digits (0 to 9, in total 10 letters). After 62 steps, the suffix is built up by one middle letter from digits and one tail letter. For each reduction step, the suffix is unique.

For example, in the BCARY reduction ([prefix]="BCARY") with a [chem_id] of "R" (see *rdc_cfg.ini*), the reduced mechanism resulted from the first reduction step will be named after *BCARYRa* (suffix: "a"). All the chemical files generated for this mechanism are stored in the chemical folder *BCARYRa*. Following the same rule, the mechanism resulting from the 63^{rd} reduction step is named after *BCARYR0b* (suffix: "0a", 0 is the middle letter, and "a" is the tail letter).

The user should notice that only the latest reduced mechanism of one reduction step is preserved. Meanwhile, the mechanisms with different names can be identical if no reduction attempt is accepted at one or several reduction steps. To backtrack the mechanism apart from the last one at one reduction step, the user can check the record files of the reduction for more information.

2.1.2 Folder and file structure

The chemical folder contains the following files:

- ***[chem_name].reactions***, a list of the chemical reactions that is used to run SOA simulations in SSH-aerosol.
- ***[chem_name].species***, a list of the gas-phase species and their molar weights that are used to run SOA simulations in SSH-aerosol.
- ***[chem_name].aer.vec***, a list of the aerosol species and their properties that is used to run SOA simulations in SSH-aerosol. The molecular structure of the aerosol species is written as a vector of the number of the AIMOFAC functional groups (Table 1). The saturation vapor pressure is computed from the molecular structure using UManSysprop[11]. The

[11] http://umansysprop.seaes.manchester.ac.uk/

input henry’s law is at zero and will be generated in ssh-aerosol.

- ***[chem_name].RO2***, a list of the peroxy radicals (RO2) species. Those species build up the so-called RO2 pool (referred to as species "RO2" in the species list), used to compute RO2-RO2 reactions.

- ***[chem_name].mol***, a list of all chemical species along with their properties. This file is used in GENOA.

- ***[chem_name].viz***, a file is used to plot the reaction pathways with Graphviz[12]. The file is only generated when the number of aerosols (condensable species) is lower than 20.

2.2 Input files of atmospheric conditions

The geographic information, initial/ background concentrations, and meteorological data of the atmospheric conditions are required to run SOA simulations. That information is stored in one directory (read by "pathInitFiles" in the configuration file). The given input files are extracted from CHIMERE simulations [3].

2.2.1 Nomenclature

In GENOA, conditions (i.e., training, pre-testing, and testing conditions) are written in the format **[**X_y**,**X_x**,**X_m**]**, where X_y and X_x note the location of the condition (X_y for latitude and X_x for longitude), and X_m notes the month (natural month equals to m+1).

To use this format, GENOA needs the coordinates of the conditions given by two files in the directory "pathInitFiles":

- ***latitudes.npy***, a NumPy array file recording the identifier y (the index of the array) and the corresponding latitude of the condition. For the given input data, the range of y is from 0 to 152, corresponding to 32 $^\circ$N to 70 $^\circ$N with a step of 0.25 $^\circ$N.

- ***longitudes.npy***, a numpy array file recording the identifier x (the index of the array) and the corresponding longitude of the condition. For the given input data, the range of x is from 0 to 142, corresponding to 17 $^\circ$W to 39.8 $^\circ$E with a step of 0.4 $^\circ$E.

[12] `http://viz-js.com/`

Index	group name (symbol)	structures
0-3	alkane group (C)	CH3, CH2, CH1, C
4-7	methanol (C[OH])	CH3OH, CH2OH, CHOH, COH
8-11	calcohol between two alcohols ([OH]C[OH])	OHCH3OH, OHCH2OH, OHCH1OH, OHCOH
12-16	calcohol in tails of alcohol ([OH]C)	OHCH3, OHCH2, OHCH, OHC
17-21	alpha-olefin group(C=C)	CH2=CH, CH=CH, CH2-C, CH=C, C=C
22,23	aromatic carbon (AC)	AC-H, AC
24-26	aromatic carbon-alkane (AC-C)	AC-CH3, AC-CH2, AC-CH
27	alcohol (OH)	OH
28	water (H2O)	H2O
29	aromatic carbon-alcohol (ACOH)	AC-OH
30,31	ketone (RCO)	CH3CO, CH3CO
32	aldehyde (HCO)	CHO
33,34	ester (COO)	CH3COO, CH2COO
35-37	ether (COC)	CH3O, CH3O, CHO
38	acid (COOH)	COOH
39	aromatic nitro (ACNO2)	AC-NO2
40-42	nitrate (NO3)	CH2ONO2, CHONO2, CONO2
43-45	hydroxyperoxide (CO-OH)	CH2OOH, CHOOH, COOH
46-54	peroxide (CO-OC	CH3OOCH2, CH3OOCH, CH3OOC, CH2OOCH2, CH2OOCH, CH2OOC, CHOOCH, CHOOC, COOC
55	peroxyacyl nitrates (PAN)	PAN
56	Peroxyacetyl acid (C(O)OOH)	COOOH

Table 1: The UNIFAC Structural Groups and its corresponding vector index in SSH-aerosol. This vector is recorded in file *[chem_name].aer.vec*.

The detailed information for each condition is stored in the subdirectory under the name "m[X_m]/y[X_y]/x[X_x]", where [X] is the index of the corresponding identifier. In the outputs files, the conditions are also recorded in another format, "m[X_m]y[X_y]x[X_x]". When with the suffix "_[#]h", it means that the simulations start at [#] hour.

2.2.2 Folder and file structures

By default, GENOA will automatically load the information from the directory "pathInitFiles" to run SOA simulations with SSH-aerosol (by altering the simulation namelists). With the loaded identifier from *dataset.list* (or from the configuration file for the training dataset), GENOA searches the input files of one condition (identifier is [X_y,X_x,X_m]) in the subdirectory m[X_m]/y[X_y]/x[X_x]. The input files related to this condition include:

- ***init_gas_[#]h.dat***, the initial gas-phase concentration for simulations started at [#] hour. In the given package, data started at 0 h and 12 h are provided.
- ***init_aero_[#]h.dat***, the initial aerosol concentration for simulations started at [#] hour.
- ***gas.cst***, the hourly concentration profiles of gas-phase species. The concentration variations of the species recorded in this file are not simulated with SSH-aerosol, as the chemical mechanisms used do not include the inorganic reactions. During the simulations, GENOA uses the concentrations of those species from the hourly input profiles recorded in this file.
- ***aero.cst***, the hourly concentration profiles of gas-phase species. The same use as *gas.cst* but for aerosol species.
- ***meteo.dat***, a file contains the meteorological information, i.e., the temporal profiles of the relative humidity and the temperature (K).

3 Running the BCARY reduction

All data required to run the BCARY reduction should have been downloaded within the Zenodo package. Please reserve at least 5 GB of storage for the simulations.

To run the BCARY reduction with the predefined configuration file *rdc_cfg.ini* (with a *[chem_id]*: "R"), the user just needs to execute the commands option

I or II in the Listing 2. Since one BCARY reduction may take a few days, we recommend using option II with the command nohup, which allows the user to run the reduction in the background.

Please be aware that due to the influence of numerical noise on SOA concentration, the identical configuration file and initial mechanism may result in different reduced schemes with GENOA when training with different computers (or different version compilers). To exactly reproduce the reported "Rdc." mechanism (i.e., the *BCARYR6R* mechanism in the package), one may need to perform the reduction with Ubuntu 16.04.6 LTS[13], along with gFortran 5.4.0 and gcc 5.4.0.

```
# go to the src folder
cd GENOA/src
# option I: not use the nohup command
python3 __init__.py ../inputs\_bcary/rdc\_cfg.cfg
# option II: use nohup: recommended
# the printed message will be stored in the file nohup.out
nohup python3 __init__.py ../inputs]_bcary/rdc\_cfg.cfg &
```

Listing 2: Running BCARY reduction

3.1 Output structures

After starting the reduction, the output folder named ***rdc*** will be generated automatically. All files related to the reduction will be stored in this directory once it is generated. The output folder *rdc* contains the following folders/files:

- ***chems***, a folder contains all the chemical mechanisms and the records generated from the reduction.
  - ***CaseRdc_[chem_id]***, a file that records all the validated reductions. The content of the record is explained in Sect. 3.2.
  - ***[chem_id]_chems***, a folder contains the details of the generated chemical mechanisms.
    - ***Testing_[final_chem_name]_all***, the record of the testing results of the final reduced mechanism [final_chem_name]. This file is only generated during the GENOA testing process.

[13] http://old-releases.ubuntu.com/releases/16.04.6/

    - ***[chem_id]_recs***, a folder contains the details records for each reduction step, including the changes in the mechanism, reduction parameters, and the SOA simulation results (shown in error) for each condition. The record per reduction step is named after *"Record_" + [chem_name] + "_[strategy]" + "_all"* and *"Record_" + [chem_name] + "_[strategy]" + "_use"*, where the file with "_all" records all the reduction attempt and with "_use" only records the validated reduction attempt. The strategy name uses abbreviation listed in Table 2.

- ***SSHs***, a folder contains the GENOA version of SSH-aerosol models that are used in GENOA reduction. Generally, for one reduction, GENOA will automatically generate two SSH-aerosol models: one is for running the reduced mechanism, and the other is for running the reference mechanism during pre-testing and testing. A [ssh_id] is defined by the user in the configuration file, which is used to assign the name of the SSH-aerosol folders:

    - ***[ssh_id]_rdc***, the SSH-aerosol folder where GENOA simulates SOA concentrations with the reduced mechanism.
    - ***[ssh_id]_ref***, the folder SSH-aerosol where GENOA simulates SOA concentrations with the reference mechanism.
        - ***ref***, a folder contains all the total SOA concentrations simulated by the reference mechanism.

- ***smls***, a folder contains the files related to SSH-aerosol simulations (SOA concentrations, namelists, log files) on training dataset.

    - ***Results_[chem_name]***, a folder contains the total SOA concentration and the ssh-aerosol namelist that simulates the mechanism *[chem_name]*. This folder is generated from each validated reduction step.
    - ***Results_[ref_chem_name]FA***, a folder contains the SSH-aerosol results simulated with the reference mechanism under the complete mode, where the concentrations of each species are recorded. The produced concentrations of radicals are also recorded under the name "FA"+[radical_name]. This reference mechanism, where the production of radicals is noted, is noted by "IDchemFake" in GENOA.

A python script *clean.py* is provided to help the user clean the output files/folders. One can use the command in Listing 3 to run the cleaning process (remove most of the output files), after modifying the option in the script.

Strategy Name	Abbr.
removing reaction	rm
removing elementary-like reaction	rm1
lumping	lp
replacing	rp
jumping	jp
removing species	rs
removing gas-particle partitioning	da

Table 2: Reduction strategies: names and abbreviations

```
# go to the src folder
cd GENOA/src/
# run the script with the applied configuration file
python3 ../examples/clean.py ../inputs_bcary/rdc_cfg.ini
```

Listing 3: Cleaning the output of the reduction

3.2 Records file contents

The total record file *CaseRdc_[chem_id]* tracks the entire reduction process. After one reduction step, GENOA will write down the reduction information as shown in Listing 4.

3.2.1 Reduction parameters

Generally, the record includes the major reduction parameters, including the reduction strategy (line2), the name and the path to the reference mechanism ("IDchemRef" and "refPath"), and the previously validated mechanism ("IDchemPre" and "prePath"), the error tolerance for evaluating the reduction ("err_ref" and "err_pre"), the number of reduction step and stage (1 for late-stage I and 2 for late-stage II.).For the reduction via removing reaction, the threshold of branching ratio is also recorded (line 7).

3.2.2 Training results

The training results are displayed by the change in the size of the mechanism and the pre-testing results after each reduction step. The size of the chemical scheme (in terms of the number of reactions/ gas-phase species/ condensable species) and the pre-testing results before the reduction (line 8 and 9) is given for the first reduction step. For the reduction via removing elementary-like reduction, the scheme before reduction is also given, where the reactions are written with one product (the number of reactions will be difficult from other mechanisms). After the reduction, the chemical scheme (line 13) and the pre-testing results (line 23) are recorded per reduction step. All the validated reduction is listed out in the record (from line 15).

For pre-testing, "err loc" is the condition simulated with the maximum error, "err max" is the maximum error, "err ave" is the average error, and "err ave max" is the maximum average errors simulated at different starting times. The more differences between "err ave" and "err ave max", the mechanism introduces more errors at one period of simulation.

3.2.3 Late-stage reduction

In the last stage of reductions (late-stage I and late-stage II), extra information is recorded in order to help the user choose the best reduction or adjust the user-chosen parameters. The chemical scheme (e.g., [29, 14, 5] in line 10 for the number of [reactions, gas, aerosols]) and the pre-testing results of the tested reduction attempt are recorded, along with the tolerance. A keyword "Sim" is noted for the reduction attempts that have passed the evaluation on the training dataset, while another keyword "NoSimCuzNaer" is applied to the reduction attempts that bypass the evaluation on the training dataset as a result of the aerosol-oriented reduction. If the reduction attempt is accepted in the pre-testing, the record line end with "refuse: 0". Otherwise, it ends with "refuse: 1".

4 Running your own reduction

With the provided input data, the user can run their own BCARY reductions with customized configuration files in Sect. 4.1.

To run the reduction on other chemical mechanisms, the user needs to prepare:

- A folder contains the chemical files (e.g., *BCARYorg*) required for the reference mechanism. The user can also contact the author to generate the reference mechanism with the following information (* indispensable):
  - a reaction list*, containing species names and kinetic rates. The file in SPACK[14] format or KPP[15] format is favourable.

[14] https://www.cerea-lab.fr/dossiers/racine/articles/guide-0.pdf
[15] https://people.cs.vt.edu/asandu/Software/Kpp/

```
===============================================
Training. Reduction strategy: Removing reactions
IDchemPre: BCARYorgP        prePath:
 ↪  GENOA/rdc/smls/R/Results_BCARYorgP
IDchemRef: BCARYorg        refPath:
 ↪  GENOA/rdc/smls/R/Results_BCARYorg
Error Tolerance: err_ref <= 0.010000, err_pre <= 0.010000.
Current reduction step: 1 with tail: R. Reduction stage: 0
Branching ratio for removing reactions: 0.050000        nBRT: 0
Initial scheme No.reaction: 1241        No.gas:
 ↪  493        No.aerosol: 356
Pre-Testing on IDchemPre BCARYorgP: err loc: m4y25x33        err
 ↪  max: 0.0007         err ave: 0.0000         err ave max:
 ↪  0.0000
NoSimCuzNaer        [29, 14, 5]        err loc:
 ↪  m11y50x27        err max: 0.2644 <= 0.2000        err ave:
 ↪  0.0694 (0.0591) <= 0.0300        refuse: 1
Sim        [29, 15, 6]        err loc: m11y50x27        err max:
 ↪  0.1738 <= 0.2000        err ave: 0.0324 (0.0306) <=
 ↪  0.0300        refuse: 1
END, total run: 115 times, valid run: 110
Current scheme No.reaction: 1094        No.gas:
 ↪  473        No.aerosol: 339
------------------------------------------
Valid:
CO24C4CHO + NO3 -> CO2C3CO3 + CO //
   + HNO3
%KNO3AL*5.5
KINETIC ARR2 7.700E-12 1860.00

...

Pre-Testing IDchem BCARYRa: err loc: m11y50x27        err max:
 ↪  0.0082         err ave: 0.0018         err ave max: 0.0023
No.reduced        1        No.round        1        Used time:
 ↪  2480.7
===============================================
```

Listing 4: An example of the recorded information for one reduction step in *CaseRdc_[chem_id]*

  - a species list, containing species name*, chemical formula, molecular structures* (SMILES format or vector format shown in Table 1), and molecular weight*. For condensable species, please also provide saturation vapor pressure, enthalpy of evaporation, and other information that is required in the aerosol species list *[chem_name].aer.vec*.

- The input files of involved atmospheric conditions (training, pre-testing, and testing conditions). Please update both the input files and the identifiers to the dataset.

- A configuration file for GENOA.

For each reduction, a unique chemical ID ([chem_id], "IDchem" in the configuration file) should be assigned in the configuration file in order to avoid confusion with other reductions.

4.1 Configuration Options

The configuration file contains the essential reduction parameters/ options that the user may need to check before each reduction.

4.1.1 Basic settings

As shown in Listing 5, the chemical ID ("IDchem", or [chem_id] in the previous text), prefix for the chemical file (prefix), and a list of the primary VOCs in the mechanism ("primaryVOCs") are listed in the group [chemistry_id]. For each reduction, "IDchem" needs to be unique to avoid overlapping the reduction results. The SOA precursors recorded in "primaryVOCs" will not be removed from the scheme or merged with other species.

In the group [action], the flag "training" is activated for running the training process, while the flag "testing" is activated to run the testing process. If the flag is on, please check the group under the same name to ensure the input is updated.

4.1.2 Input and output

The groups [input] and [output] shown in Listing 6 allow the user to change the input data and the output repositories for the reduction.

"namelist_pre" is the default namelist used for SOA simulations with SSH-aerosol. For each simulation, a namelist is generated based on this file with updated meteorological data, starting simulation time, initial/ background

```
[chemistry_id]
# suffix of the chemical scheme name
IDchem = 'R'
# prefix to save results
prefix = 'BCARY'
# name of the precursor
# should be consistent with the name in the initial file
primaryVOCs = ['BCARY']


# True/False to active the corresponding section
# update if only active
[action]
# 1 for on and 0 for off
# Set both training and testing at 1: reduction is on
training = 1
testing = 1
```

Listing 5: Basic setting

concentration files, and output directory. The rest of the parameters from the default namelist are used in the simulations.

"pathInitFiles" is the path to the initial files of the atmospheric conditions. "speciesfile" is the path to the *[chem_name].mol* file of the reference mechanism.

"pathSSH", "pathNewChem", and "pathNewRes" are the three output directories, which are used to run SOA simulations with SSH-aerosol, store the generated chemical mechanisms, and store SSH-aerosol outputs (mainly the total SOA concentrations and namelists) of the validated reduced mechanisms on the training dataset, respectively.

4.1.3 Settings for testing

If the flag in group [action] "training" is activated, GENOA will run the training process with the setting in the group [training] noted in Listing 7. The training dataset, is assigned in the configuration file in the group [conditions] by the name "locs".

Two options are provided to run the training process, where option I is to run one reduction from the beginning, and option II is to run one reduction from a breakpoint. For a new reduction, one should always start from option I, activating "from_ref" (= 1), "IDchemRef", and "refChemPath" (the chemical name, and the path to the folder that contains the reference mechanism).

```
[input]
# the default namelist used in ssh-aerosol-genoa.
namelist_pre = 'namelist_ssh'
# path to initial concentrations. Folders ordered by
 ↪  m[X_m]/y[X_y]/x[X_x]/
# all input files are extracted from CHIMERE
pathInitFiles = '../../conditions_bcary/'
# species properties in the reference chemical mechanism.
 ↪  Generated by GENOA
speciesfile = '../inputs_bcary/BCARYorg/BCARYorg.mol'

[output]
# directory to ssh-aerosol
# generate two folders [pathSSH]+'_rdc' [pathSSH]+'_ref' for
 ↪  reduction
pathSSH = '../rdc/SSHs/R'
# directory to save generated chemical mechanisms
pathNewChem = '../rdc/chems'
# directory to save SSH-aerosol simulation results on training
 ↪  dataset
pathNewRes = '../rdc/smls'
```

Listing 6: Setting for input and output

If the user has the chemical mechanism and the SOA simulation results on the training dataset of the previously validated mechanism (name: "IDchemPre", the path to the mechanism: "preChemPath", and the path to SOA simulation results: "prePath") and the reference mechanism with produced radical species (name: "IDchemFake", and path to the SOA simulation results (under the complete mode of SSH-aerosol): "fakePath")

The reduction strategies are read in order from their abbreviations (see Table 2) in the list "strategy_types". If the strategy of removing elementary-like reactions ("rm1") is not in "strategy_types", after traversing all possible reductions with the provided strategies, GENOA will automatically add it after removing reactions ("rm") for further reduction.

"BranchRatio" lists the thresholds of the branching ratios used for the reduction via removing reaction. This list should be in the ascending order with a maximum value not larger than 1 (try to remove all reactions one by one). Currently, "BranchRatio" is initialized with [5E-2, 1E-1, 5E-1]. When the error tolerance err_ref reaches 0.03, "BranchRatio" is increased to [1E-1, 5E-1, 1], while "BranchRatio" is set to [1.] for the late reduction stages and for reduction via removing elementary-like reaction.

"err_ref" and "err_pre" are the error tolerances used to compare the SOA simulation results on training dataset simulated with the reduced mechanism to the ones simulated with the reference mechanism and with the previous validated mechanism, respectively.

4.1.4 Settings for pre-testing

The settings for pre-testing is displayed in Listing 8, where "nPreTest" is the number of pre-testing conditions. GENOA will read the pre-testing conditions from the beginning of the list of the testing dataset ("Test_file" in-group [testing], from the first condition to the "nPreTest" number of conditions.

"try_at_err" is the "err_ref" value at which GENOA needs to activate the restrictions on pre-testing results. That means after "err_ref" reaches "try_at_err", GENOA will run the pre-testing simulation after each reduction attempt that has been accepted in the evaluation with the training dataset. The reduction that exceeds the error tolerances of pre-testing will be refused. For efficiency, "try_at_err" for BCARY reduction is set at 0.04. When "try_at_err" is larger than "try_ave_ref", the user should make sure the average error of pre-testing (the pre-testing results are monitored per reduction step) before adding the restriction is no more than "try_ave_ref".

The error tolerances for adding the restrictions on pre-testing results are assigned by the average error "try_ave_ref" and the maximum error "try_max_ref" compared to the reference mechanism. Due to the aerosol-oriented treatment

```
[conditions]
# training dataset
locs = [[55, 65, 11],[0,19,6],[33, 34, 6],[38,7,11],[148,127,0],[144,88,11],[0,134,7],[16,81,6]]

[training]
# Option I: active from_ref IDchemRef refChemPath
# pre chem (IDchemPre,preChemPath,prePath) and
# fake chem (IDchemFake,fakePath) are generated automatically
from_ref = 1
IDchemRef = 'BCARYorg'
refChemPath = '../inputs_bcary'

# Option II: run reduction from a break point
# pre chem
#IDchemPre = 'BCARYorgP'
#preChemPath = '../rdc/chems/R_chem'
#prePath = '../rdc/smls/R/Results_BCARYorgP'
# ref chem
#IDchemRef = 'BCARYorg'
#refChemPath = '../rdc/chems/R_chem'
#refPath = '../rdc/smls/R/Results_BCARYorg'
# fake chem
#IDchemFake = 'BCARYorgFA'
#fakePath = '../rdc/smls/R/Results_BCARYorgFA'

# reduction_strategies
strategy_types = ['rm','rp','lp','jp','rs','ra']

# used for removing reaction
BranchRatio = [5E-2, 1E-1, 5E-1]

# error tolerance compared to ref case
err_ref = [0.01,0.02,0.02,0.03,0.03,0.03,0.04,0.04,0.06,0.06,0.08,0.08,0.10,0.10]
# error tolerance compared to pre case
err_pre = [0.01,0.01,0.02,0.01,0.02,0.03,0.02,0.04,0.04,0.06,0.04,0.08,0.08,0.10]
```

Listing 7: Settings for training

in the late reduction stages, the error tolerance will be larger for reductions that reduce the number of aerosols. (0.01 larger on "try_ave_ref" and 0.10 larger on "try_max_ref".)

```
# setting for pre-testing
# err_ref at which to start the evaluation on pre-testing
try_ave_ref = 0.03

# max err allowed on pre-testing conditions
# an exceed of 0.10 is allowed if number of
# aerosol is reduced in the late stage of reduction (No.aer <= 20)
try_max_ref = 0.20

# number of pre-testing conditions (0-nPreTest) selected from testing dataset
nPreTest = 150
```

Listing 8: Settings for pre-testing

4.1.5 Settings for testing

Finally, for the testing process, the flag "testing" in group [action] need to be activated. "Test_file" is the list of testing dataset. The record file for testing will be generated in the directory "pathNewChem"+[chem_id]+"_chems", named after *Testing_[chem_name]_all*.

```
[testing]
# list of testing dataset
# ../../conditions_bcary/
Test_file = '../inputs_bcary/dataset.list'
```

Listing 9: Settings for testing

References

[1] K. Sartelet, F. Couvidat, Z. Wang, C. Flageul, and Y. Kim, “Ssh-aerosol v1. 1: A modular box model to simulate the evolution of primary and secondary aerosols,” *Atmosphere*, vol. 11, no. 5, p. 525, 2020.

[2] M. Jenkin, K. Wyche, C. Evans, T. Carr, P. Monks, M. Alfarra, M. Barley, G. McFiggans, J. Young, and A. Rickard, “Development and chamber evaluation of the mcm v3. 2 degradation scheme for β-caryophyllene,” *Atmospheric Chemistry and Physics*, vol. 12, no. 11, pp. 5275–5308, 2012.

[3] G. M. Lanzafame, B. Bessagnet, D. Srivastava, J. L. Jaffrezo, O. Favez, A. Albinet, and F. Couvidat, “Modelling aerosol molecular markers in a 3d air quality model: Focus on anthropogenic organic markers,” *Science of The Total Environment*, p. 155360, 2022.