MC-ChemDB

P. Pernot
pascal.pernot@u-psud.fr

January 26, 2017

Abstract

This is a proposal to provide user-friendly support for Monte Carlo Uncertainty Propagation (MCUP) in chemistry databases. A prototype has been implemented in the new MC-ChemDB framework.

Contents

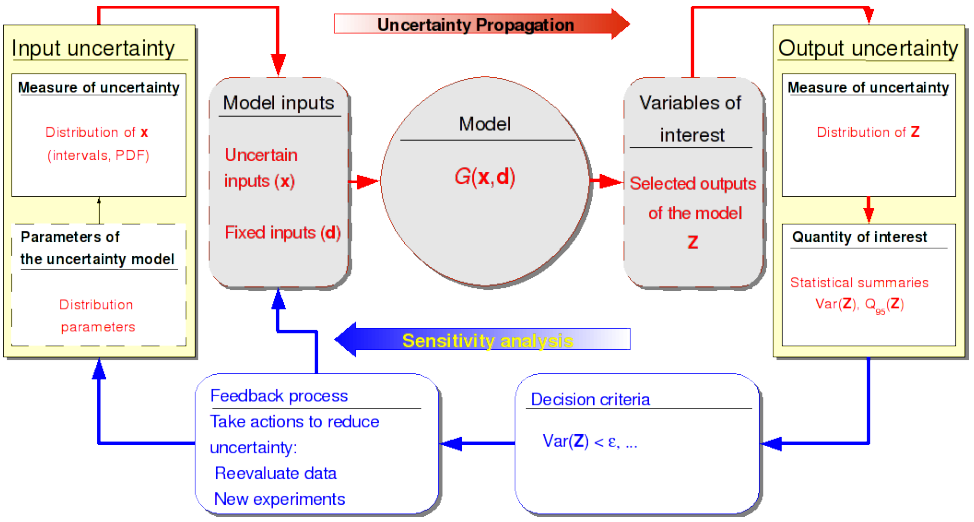

Figure 1: Flowchart of MCUP/SA.

1 Introduction

Monte Carlo Uncertainty Propagation is an optimally parallel problem, where an *uncertainty unaware* code of a chemistry model is run repeatedly for different realizations of its uncertain input data (Fig. 1).

To benefit from this feature and from increasingly available cloud-like computing infrastructures, it is best to keep the uncertainty management separate from the physical model. In the following, one assumes that the uncertain chemistry inputs/databases are provided by a server (MC-ChemDB). This separation has additional advantages for the final user:

- no change to the chemistry code, or minor ones, depending on the uncertainty representation used on the server side (Section 2.1);
- no change to the standard format of chemistry files;
- the complex aspects of uncertainty models (*e.g.* for branching ratios) are implemented on the server side.

In this server-client framework, the chemistry server database generates and stores a large number (say 1000-10000) Monte Carlo samples of chemistry files (Fig. 2). This is done once for each new release of the chemical database. Users download a number of these samples according to their needs. If a subset chemistry is required, the database can be filtered either on the server or on the client side. The user has then only to run her/his code on each of the sample to get a sample of model predictions, to be used for uncertainty estimation and sensitivity analysis (Fig. 1).

The following sections go into more details on the required work on the server and user side.

2 Data and code adaptations

Of course, in the proposed scheme, all the overhead is on the server side and mainly consists in generating the random samples. The main issue is that the server should provide samples that

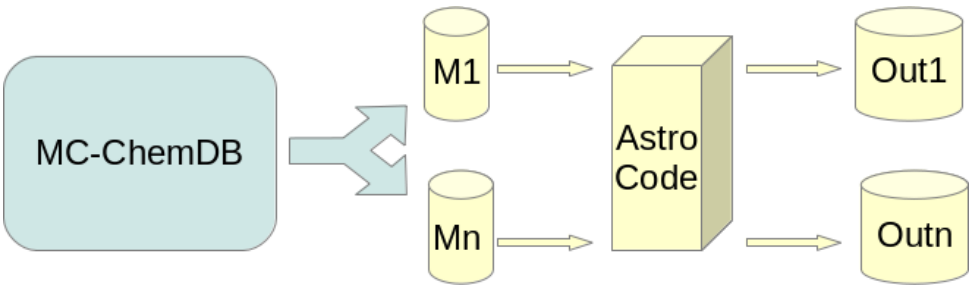

Figure 2: Flowchart of the proposed MCUP scheme.

can be used to generate temperature/pressure-dependent rate constants, while preserving the standard format of chemistry databases. In our implementation, we consider the KiDA-type format [Wakelam et al., 2012], with one line per reaction.

Let's look at a few representative cases.

2.1 Rate constants

When a rate constant (global or partial has been directly measured), two representations of the temperature-dependent uncertainty have been proposed:

- a representation of the uncertainty band by a specific function, called hereafter the "f/g representation" [DeMore et al., 1994, Sander et al., 2006, Hébrard et al., 2006]; and
- the covariance matrix of the rate-law parameters fitted on the experimental data [Hébrard et al., 2009], or on the experimental uncertainties [Nagy and Turányi, 2011].

2.1.1 The f/g representation

The five parameters necessary to calculate the Kooij expression of a bimolecular rate constant $k(T)$ and its T-dependent multiplicative uncertainty $u_k(T)$ are $\{\alpha,\, \beta,\, \gamma,\, f,\, g\}$:

$$\begin{aligned} k'(T) &= k(T) \times u_k(T) && (1)\\ k(T) &= \alpha \times (T/T_0)^{\beta} \times \exp\left(-\gamma/T\right) && (2)\\ u_k(T) &= \exp\left(r \times \log\left(f \times \exp(g\ |1/T - 1/T_0|)\right)\right) && (3) \end{aligned}$$

where $r \sim N(0,1)$ is a standard normal random number, and T_0 is a reference temperature, typically 300 K.

With current databases, the user reads the values of f and g and generates random realizations of $u_k(T)$ curves. In the perspective of transferring the random number generation on the server side, one can take advantage of the fact that Eq. 3 can be rewritten as

$$u_k(T) = f^r \times \exp\left(g \times r\ |1/T - 1/T_0|\right), \qquad (4)$$

which shows that random realizations of $u_k(T)$ curves can still be parameterized by only two numbers $f' = f^r$ and $g' = g \times r$. Without changing the initial database format, one can thus replace the nominal values in the database by random values

$$\{\alpha,\, \beta,\, \gamma,\, f,\, g\} \longrightarrow \{\alpha,\, \beta,\, \gamma,\, f' = f^r,\, g' = g \times r\}\,. \qquad (5)$$

In order to benefit from this solution, the end user has to adapt the rate expression in his code to include the $f * exp(g\,|1/T - 1/T_0|)$ term.

An alternative solution for lazy users who treat systems with uniform temperatures would be to provide them with $\{\alpha', \beta, \gamma, 1, 0\}$ sets, where α' is calculated by the server at the required temperature T_1 as $\alpha' = \alpha \times u_k(T_1)$.

2.1.2 The covariance representation

An alternative description of the rate constants uncertainty is based on the variance-covariance matrix of the rate parameters [Hébrard et al., 2009, Nagy and Turányi, 2011].

In this representation, a random perturbation is obtained as $\{\alpha', \beta', \gamma', 1, 0\}$, where the primed parameters are generated by random sampling from the mean values $\{\alpha_0, \beta_0, \gamma_0\}$ and the variance-covariance matrix $\Sigma_{\alpha,\beta,\gamma}$ of the rate-law parameters

$$\{\alpha', \beta', \gamma'\} = \mathscr{N}\left(\{\alpha_0, \beta_0, \gamma_0\}, \Sigma_{\alpha,\beta,\gamma}\right) \tag{6}$$

This case does not require any modification of the standard chemistry files format, nor any change to the user's code, even for non-uniform T systems.

Note. This representation differs from the f/g representation by the inter-temperature correlation of the random rate constants. One has $\mathrm{corr}\,(k(T_1), k(T_2)) = 1$ in the f/g case, while $\mathrm{corr}\,(k(T_1), k(T_2))$ is a function of both temperatures in the covariance matrix case. To my knowledge, nobody has demonstrated that this correlation difference might have an impact on uncertainty quantification in realistic systems. In order to see an effect, there has to be a strong T gradient coupled with an efficient transport. Maybe in models of chemical explosions ???

2.2 Branching ratios

When branching ratios (BR) have been measured instead of partial rate constants, one has to use a specific representation separating the global rate constant from the BRs. For the global rate constant, the uncertainty model can be any of the two cases detailed previously.

Random values of the BRs, b_i, are sampled from Dirichlet-type distributions [Plessis et al., 2010]. This way, sets of BRs are generated consistently with the constraint $\sum b_i = 1$, avoiding the spurious effects demonstrated in [Carrasco and Pernot, 2007, Carrasco et al., 2008]. For many ion processes, notably dissociative recombination, the generation of BRs is complex and is better handled on the server side.

To derive a partial rate constant, one has thus to generate on the server side $\{\alpha' = \alpha * b'_i, \beta, \gamma, f', g'\}$ (for instance), where b'_i is a realization of the BR for the i^{th} pathway.

Here again, no change of database format is required. On the user side, the changes are the same as described above, depending on the description of the global rate's uncertainty.

2.3 Photo-process

A special treatment has to be considered for photo-processes which are described by wavelength-dependent properties (cross-sections, branching ratios) [Gans et al., 2013]

The server has to provide random samples of wavelength-dependent cross-sections and branching ratios as sets of files.

2.3.1 Cross-sections

$$\sigma'(\lambda) = \sigma(\lambda) * u_\sigma \tag{7}$$

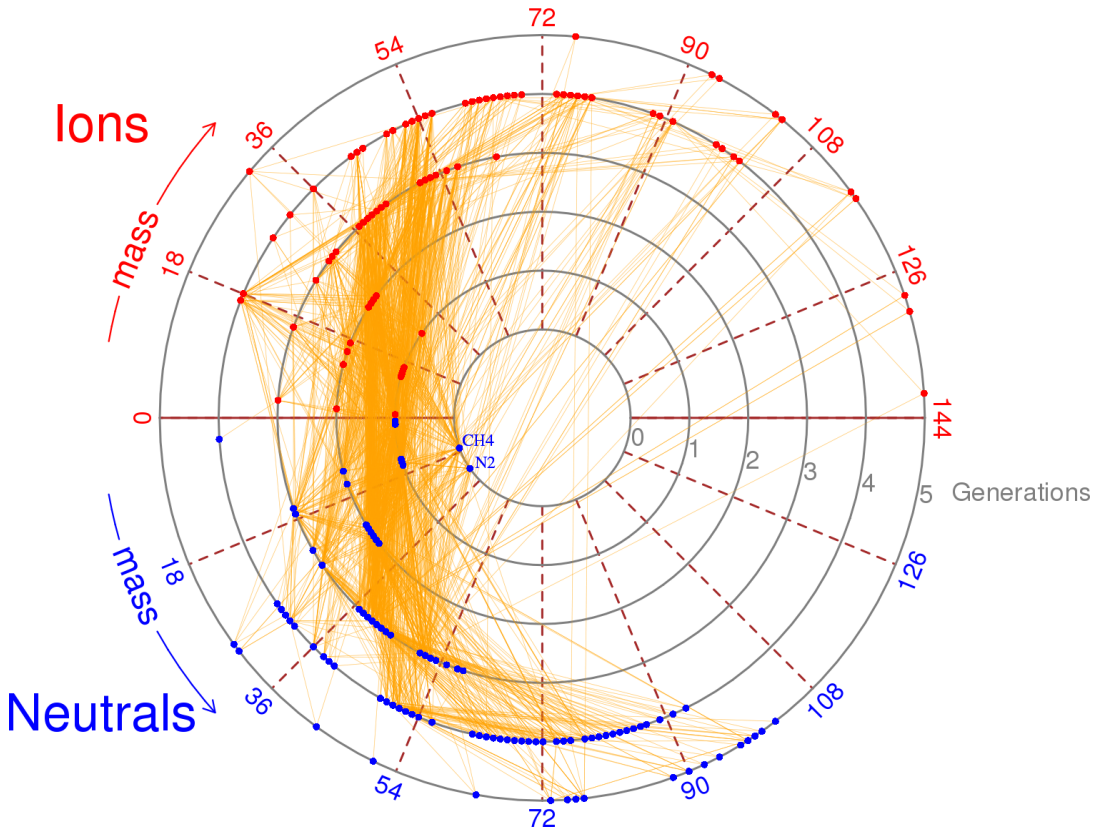

Figure 3: Generating scheme of the species list in a N_2/CH_4 photochemical plasma.

2.3.2 Branching ratios

3 Automatic model building

A consistent set of reactions can be iteratively generated from a set of species. For instance, in a model N_2/CH_4 photochemical plasma (1) one would start from $N_2, CH_4, h\nu$; (2) select all the reactions involving these 3 species (in fact, photo-processes of N_2 and CH_4); (3) update the list of species with the generated products; and (4) iterate until no new species is produced.

Other generating schemes might be provided.

4 MC-ChemDB structure

On the server side, one considers a `Source` database, which contains all the relevant information to generate random samples of the reaction rates. After processing by the `R` script, a summary sheet and database samples are generated in temporary storage (Table 2). Then the samples for all reactions are gathered in global samples, which are stored in the `Public` database. They can be downloaded by users, or used to generate specific models databases.

There is presently a discrepancy on the uncertainty representations of reaction rates for reactions involving neutrals and the ones involving ions. The sum-to-one of branching ratios has been extensively implemented in ionic reactions by Pernot and coworkers [Carrasco et al., 2007, Carrasco and Pernot, 2007, Plessis et al., 2010]. This has not yet been done for neutrals [Hébrard, 2006, Hébrard et al., 2009]. Moreover, a special treatment has to be considered

C2H2+ + E			
ALPHA	Logu	2.7E-7	10e-7
REF_ALPHA	Florescu2006	physto	
BETA	Unif	0.5	1.0
REF_BETA	Florescu2006	physto	
REF_BR	Derkatch1999		
BR	Diri	Dirg	
C2H + H		0.5/0.06	
C2 + H + H		0.3/0.05	
C2 + H2		0.02/0.03	
1CH2 + C	1/1	0.05/0.01	
3CH2 + C	1		
CH + CH		0.13/0.01	

Table 1: Exemple of Source data for ionic reactions.

Reactants		**Products**			α	β	γ	f	g	**Type**
C2H2+	E	C2H	H		2.591e-07	-0.752	0	1	0	dr
C2H2+	E	C2	H	H	1.563e-07	-0.752	0	1	0	dr
C2H2+	E	C2	H2		8.563e-09	-0.752	0	1	0	dr
C2H2+	E	1CH2	C		1.323e-08	-0.752	0	1	0	dr
C2H2+	E	3CH2	C		1.251e-08	-0.752	0	1	0	dr
C2H2+	E	CH	CH		6.638e-08	-0.752	0	1	0	dr

Table 2: One of the Public database samples generated from Table 1.

for photo-processes which are described by wavelength-dependent properties (cross-sections, branching ratios) [Gans et al., 2013]. The `Source/Public` databases consist therefore of three sections (`Ions`, `Neutrals` and `Photo`).

4.1 Ion reactions

For ion reactions, one describes separately the T-dependence of the global rate constant (Kooij or ionpol1/2), and the set of branching ratios (BR).[1]

The `Source` database consists of a set of tab-delimited '.csv' files, which are processed by a `R` script (`parser.R`) to produce the adequate samples. One example is shown below for the dissociative recombination of $C_2H_2^+$ (Table 1). The ALPHA, BETA sections define the probability density functions of the (independent) DR rate parameters, *i.e.* $p(\alpha, \beta) = p(a)p(\beta)$; the BR section codes for the probabilistic tree in Fig. 4; and the REF_XXX sections provide `bibtex` labels for bibliography on property XXX.

[1] At the moment, there is practically no information on T-dependence of BRs, and no provision has been made in the database for this opportunity. The database contains thus only informations to generate T-independent BRs.

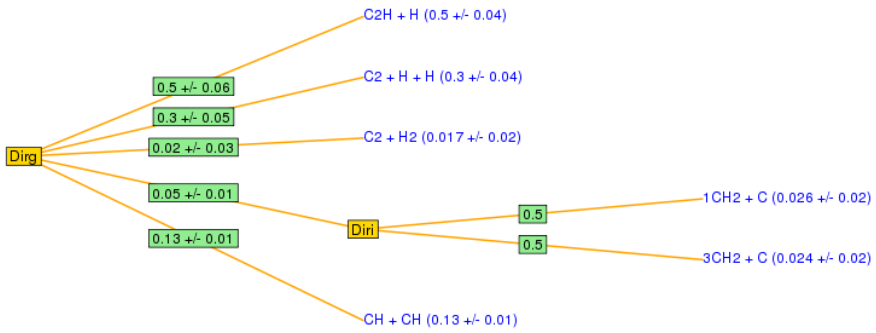

Figure 4: Probabilistic tree for the fragments of the dissociative recombination of $C_2H_2^+$.

4.2 Neutral reactions

The branching ratios formalism has not yet been implemented for reactions between neutral species, and the `Source` databases are in the 'standard' format with 1 line per reactions. They are presently imported as '.csv' files from the `GoogleDocs` lists maintained by J.-Ch. Loison et M. Dobrijevic (Titan).

From these `Source` files, samples of databases are generated in the `Public` directory using the f/g representation (Eq. 5). Samples of T-dependent rate curves are also generated as a mean to check for errors and provided as summary sheets (Fig. 5).

4.3 Photo-processes

The `Photo` section of the database extends the `Ions` model for the use of wavelength-dependent data. The uncertainty on the cross-section is modeled by a lognormal distribution for which only the uncertainty factor is used. The mean value is replaced by values read in a `se_XXX` file, where `XXX` is the process descriptor. Similarly, for BRs, only the structure and uncertainty information is used, and the mean values are extracted from `qy_YYY` files, where `YYY` is the channel descriptor.

As an example, consider Table 3 for $N_2+h\nu$. The `N2 + HV` directory contains the following files: `data.csv, se_N2_+_HV.dat, qy_N4S_+_N2D.dat, qy_N2+_+_E.dat, and qy_N4S_+_N3P+_+_E.dat`. The `se_` and `qy_` files ascii files with two columns: wavelength and value. They have to be on a common wavelength scale. The sampler will generate similar random sets of `se_` and `qy_` files from the information in `data.csv`.

For most photodissociations, one does not have extensive information about the tree structure or the channels uncertainties [Huebner and Mukherjee, 2015], and the main distinction is made between ionic channels and neutral channels [Peng et al., 2014].

The present scheme assumes only systematic, wavelength-independent, uncertainty.

5 Conclusions and perspectives

One can therefore envision to provide a user-friendly MCUP service.

9: H+C2H6->C2H5+H2
Rate law: kooij
Parameters: 1.22e-11 / 1.500 / 3720.00 / 2.00 / 100.00
! Baulch et al. [1992]

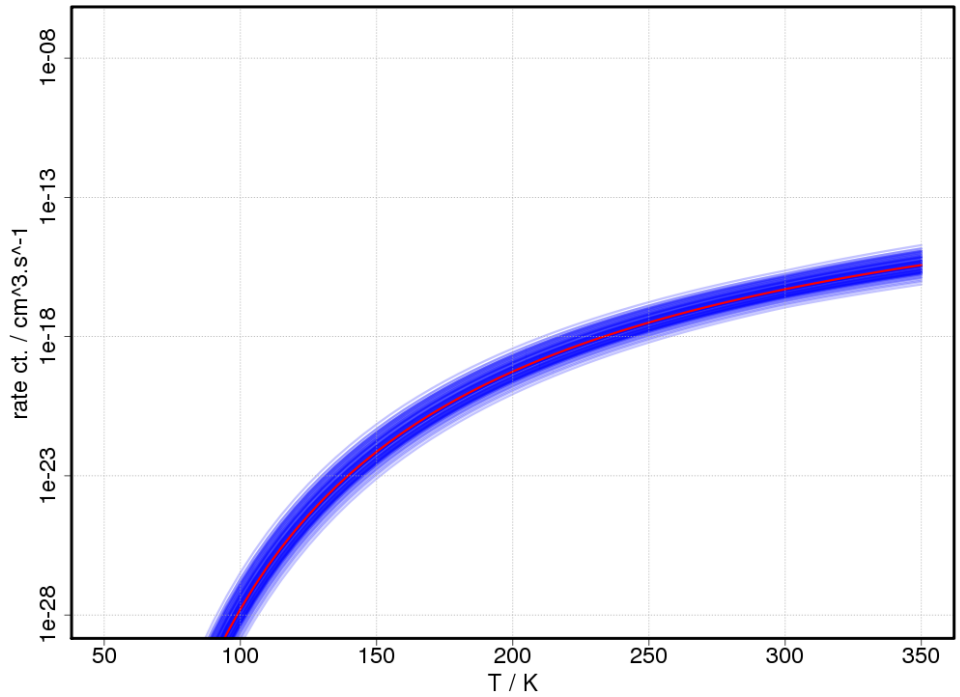

Figure 5: Summary sheet for neutral reaction $H + C_2H_6$.

References

N. Carrasco and P. Pernot. Modeling of branching ratio uncertainty in chemical networks by Dirichlet distributions. *J. Phys. Chem. A*, 11(18):3507–3512, 2007.

N. Carrasco, O. Dutuit, R. Thissen, M. Banaszkiewicz, and P. Pernot. Uncertainty analysis of bimolecular reactions in Titan ionosphere chemistry model. *Planet. Space Sci.*, 55:141–157, 2007. doi: 10.1016/j.pss.2006.06.004.

N. Carrasco, S. Plessis, and P. Pernot. Towards a reduction of the bimolecular reaction model for Titan ionosphere. *International Journal of Chemical Kinetics*, 40(11):699–709, 2008. URL http://dx.doi.org/10.1002/kin.20374.

W. B. DeMore, S. P. Sander, D. M. Golden, R. F. Hampson, M. J. Kurylo, C. J. Howard, A. R. Ravishankara, C. E. Kolb, and M. J. Molina. Chemical kinetics and photochemical data for use in stratospheric modeling. Evaluation number 11. *JPL Publication*, 94-26:1–273, 1994.

B. Gans, Z. Peng, N. Carrasco, D. Gauyacq, S. Lebonnois, and P. Pernot. Impact of a new wavelength-dependent representation of methane photolysis branching ratios on the modeling of Titan's atmospheric photochemistry. *Icarus*, 223:330 – 343, 2013. ISSN 0019-1035. doi: 10.1016/j.icarus.2012.11.024. URL http://www.sciencedirect.com/science/article/pii/S0019103512004782.

N2 + HV			
ALPHA	Logn	1	1.2
REF_ALPHA	Hebrard2006	Huebner2015	
REF_BR	Peng2014	Huebner2015	
BR	Dirg	Dirg	
N4S + N2D		0.5/0.01	
N4S + N3P+ + E	0.5/0.1	0.5/0.01	
N2+ + E	0.5/0.1		

Table 3: Example of `Source` data for the photo-processes of N_2.

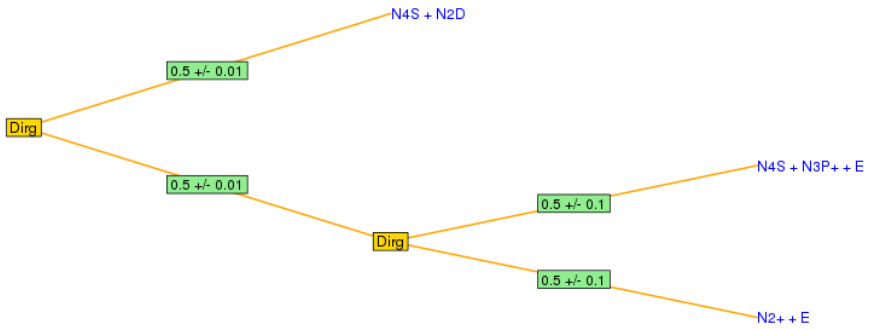

Figure 6: Probabilistic tree for the neutral and ionic fragments of the photodissociation of N_2.

E. Hébrard. *Incertitudes photochimiques dans les modèles de l'atmosphère de Titan : revue et conséquences*. PhD thesis, Université Paris VII - Denis Diderot, 2006.

E. Hébrard, M. Dobrijevic, Y. Bénilan, and F. Raulin. Photochemical kinetics uncertainties in modeling Titan's atmosphere: A review. *J. Photochem. Photobiol., A*, 7:211–230, 2006. doi: 10.1016/j.jphotochemrev.2006.12.004.

E. Hébrard, M. Dobrijevic, P. Pernot, N. Carrasco, A. Bergeat, K. M. Hickson, A. Canosa, S. D. Le Picard, and I. R. Sims. How Measurements of Rate Coefficients at Low Temperature Increase the Predictivity of Photochemical Models of Titan's Atmosphere. *J. Phys. Chem. A*, 113:11227–11237, 2009. ISSN 1089-5639. doi: 10.1021/jp905524e. URL http://dx.doi.org/10.1021/jp905524e.

W. Huebner and J. Mukherjee. Photoionization and photodissociation rates in solar and blackbody radiation fields. *Planet. Space Sci.*, 106:11–45, 2015. doi: 10.1016/j.pss.2014.11.022. URL https://doi.org/10.1016%2Fj.pss.2014.11.022.

T. Nagy and T. Turányi. Uncertainty of Arrhenius parameters. *Int. J. Chem. Kinet.*, 43:359–378, 2011. ISSN 1097-4601. doi: 10.1002/kin.20551. URL http://dx.doi.org/10.1002/kin.20551.

Z. Peng, N. Carrasco, and P. Pernot. Modeling of synchrotron-based laboratory simulations of Titan's ionospheric photochemistry. *GeoResJ*, 1-2:33–53, 2014. doi: 10.1016/j.grj.2014.03.002. URL http://dx.doi.org/10.1016/j.grj.2014.03.002.

S. Plessis, N. Carrasco, and P. Pernot. Knowledge-based probabilistic representations of branching ratios in chemical networks: the case of dissociative recombinations. *J. Chem. Phys.*, 133:134110, 2010. doi: 10.1063/1.3479907.

S. P. Sander, A. R. Ravishankara, D. M. Golden, C. E. Kolb, M. J. Kurylo, M. J. Molina, G. K. Moortgat, B. J. Finlayson-Pitts, W. P. H., and R. E. Huie. Chemical kinetics and photochemical data for use in atmospheric studies. Evaluation number 15. *JPL Publication*, 06-2:1–522, 2006.

V. Wakelam, E. Herbst, J.-C. Loison, I. W. M. Smith, V. Chandrasekaran, B. Pavone, N. G. Adams, M.-C. Bacchus-Montabonel, A. Bergeat, K. Béroff, V. M. Bierbaum, M. Chabot, A. Dalgarno, E. F. van Dishoeck, A. Faure, W. D. Geppert, D. Gerlich, D. Galli, E. Hébrard, F. Hersant, K. M. Hickson, P. Honvault, S. J. Klippenstein, S. L. Picard, G. Nyman, P. Pernot, S. Schlemmer, F. Selsis, I. R. Sims, D. Talbi, J. Tennyson, J. Troe, R. Wester, and L. Wiesenfeld. A KInetic Database for Astrochemistry (KIDA). *The Astrophysical Journal Supplement Series*, 199:21, 2012. URL http://stacks.iop.org/0067-0049/199/i=1/a=21.

A Reference for Source data for ions and photo-processes

The files are tab-delimited `data.csv` files, stored in repertories named with the reaction tag. The structure is based on keywords, followed by data (Table 1).

The first lines contains the reaction tag, *i.e.* reactants separated by the '+' symbol. This is used *inter alia* to check the mass balance with the products.

The order of the following lines is irrelevant, except for the 'BR' keyword, which has to be the last one, followed by the BR infos. The BR infos have tab-delimited tree structure starting left from the deeper branches.

Keywords	**Information**	**Note**
CHECKED	initials of checker	if present, the notice has been verified
RQ	comments	
ALPHA	pdfName, pdfParams	pdf of α parameter
REF_ALPHA	bibKeys	bibtex keys of reference for α data
BETA	pdfName, pdfParams	pdf of β parameter
REF_BETA	bibKeys	bibtex keys of reference for β data
GAMMA	pdfName, pdfParams	pdf of γ parameter
REF_GAMMA	bibKeys	bibtex keys of reference for γ data
REF_BR	bibKeys	bibtex keys of reference for BR data
BR	[pdfNamesBR]	followed by N lines :
Products	[pdfParamsBR]	1 line per product

Table 4: Keywords

pdfNames	**pdfParams**	**Note**
Delta	x_0	fixed value x_0, no uncertainty
Unif	x_{min}, x_{max}	uniform distribution over interval
Logu	x_{min}, x_{max}	log-uniform distribution over interval
Norm	x_0, u_x	normal distribution, mean x_0, stdev u_x
Logn	x_0, f_x	log-normal distribution, mean x_0, uncertainty factor f_x

pdfNamesBR	**pdfParamsBR**	**Note**
Diri	$b_1/\Gamma, \ldots, b_N$	Dirichlet distribution, mean $b_1, \ldots, b_N$, precision Γ
Dirg	$b_1/u_{b_1}, \ldots, b_N/u_{b_N}$	Generalized Dirichlet distribution, mean b_i, stdev u_{b_i}
Mlgn	$b_1/f_{b_1}, \ldots, b_N/f_{b_N}$	Multivariate lognormal (no correlation), mean b_i, uncert. factor f_{b_i}

Table 5: Available distributions for rate constants (top) and for branching ratios (bottom).