

The chemidentifier package

Aliocha SKRZYPCZAK

Version 2.0.1

2026-09-20

Abstract

`chemidentifier` numbers the compounds of a synthesis-chemistry paper in *the logical order of the synthesis* — the order in which you declare them — rather than the order in which they appear in the text. Every printed number can be made clickable, jumping to the scheme showing the compound (`links=true`, off by default, §8.3). The package is written in L^AT_EX3 (expl3); like `\ref/\label`, everything settles over **two compilations**.

Contents

1	Installation and loading	2
1.1	Which version gets loaded	2
2	The three commands	2
2.1	<code>\chemid*</code> — declaring	3
2.2	<code>\chemid</code> — using	4
2.3	<code>\chemidhere</code> — anchoring	4
3	What happens in the table of contents and bookmarks	5
4	Numbering	5
4.1	Keys	5
4.2	Beyond 26 children	5
4.3	Resetting	6
5	Templated families	6
5.1	The options of a family	6
5.2	Children numbered by the family	7
5.3	One family, one kind of child	7
6	Options	7
7	Multilingual documents	8
8	Two points worth knowing	9
8.1	The raw name derived automatically	9
8.2	A number is a link only when it has somewhere to go	9
8.3	Turning links on	10
9	Declaring after using	10

10 Text substitution in .pdf_tex figures (LuaLaTeX only)	10
11 Assumed limitations	12

1 Installation and loading

Copy `chemidentifier.sty` next to your document, or run `make install` to install it into `~/texmf`.

```
\usepackage{hyperref}           % preferably before chemidentifier
\usepackage[links=true]{chemidentifier}
```

`hyperref` is not loaded by the package: it is *detected*, leaving you full control over its options and the loading order. Clickable numbers and anchors are an opt-in on top of that: `links` defaults to `false`, so add `links=true` — as above, or later via `\chemidsetup` — to get them (§8.3). Either way, if `hyperref` is missing, everything is still numbered normally, nothing is clickable, and a warning reports it once.

1.1 Which version gets loaded

`chemidentifier.sty` is a loader: it reads the `version` option and nothing else, then hands over to the implementation asked for. Version 2 is the default and the one described here; version 1 is kept unchanged, for good, for the documents written against it.

```
\usepackage{chemidentifier}      % version 2
\usepackage[version=1]{chemidentifier} % version 1, frozen
```

Version 2 changes three things a version 1 document may notice, all three deliberate:

- `\herechemid` is now called `\chemidhere`. The old name still works and will keep working; it says once per document that it has been renamed.
- A compound with no anchor is printed *without* a link rather than linked to nothing (§8.2). Links therefore settle on the second compilation, as `\ref` does.
- `\chemidhere{parent}` now anchors the whole family, children included (§2.3).

Everything else is unchanged, and the version 1 manual is installed next to this one as `chemidentifier-doc-v1.pdf`. The regression suite compiles every version 1 test document through the frozen implementation and checks it against its original references, byte for byte.

2 The three commands

Command	Role	Output
<code>\chemid*{key}</code>	declaration	none
<code>\chemid{key}</code>	use	the number, clickable
<code>\chemidhere{key}</code>	anchor	none

2.1 \chemid* — declaring

```
\chemid*{key}
\chemid*{key}{rich name}
\chemid*{key}{rich name}[raw name]
\chemid*{key}[rich name][raw name]      % equivalent variant
```

The declaration prints nothing. **The order of declarations sets the numbering:** place them in the preamble, in the order of your synthesis scheme.

```
\chemid*{benz}          % 1
\chemid*{benz.cl}       % 1a
\chemid*{benz.br}       % 1b
\chemid*{amine}          % 2
```

The main counter advances on every new *parent*. Each parent has its own letter counter, so you can declare children later without breaking the numbering: **1** and **1a,b**, then **2**.

A *rich name* entirely replaces the automatic number:

```
\chemid*{target}{target\textsubscript{final}}[final target]
```

gives **target_{final}**. The second argument, in brackets, is the plain-text version used in PDF bookmarks and metadata, where no formatting is allowed. If you omit it, it is derived automatically from the rich name (see §8.1).

Declaring several keys at once

A comma-separated list of keys declares each of them, auto-numbered in turn — the same convention as \chemid:

```
\chemid*{mol1,mol2,mol3}      % three parents: 1, 2, 3
\chemid*{mol1.a,mol1.b}       % two children of mol1: 1a, 1b
```

A custom name cannot accompany a multiple declaration — \chemid*{a,b}{name} would be ambiguous: which key would it belong to? This is an error; declare that key on its own instead. A comma *inside* a single key (the result of a malformed key, e.g. one built by a macro) is also rejected.

A child without its parent

Declaring children in bulk without ever declaring their parent on its own also works: the parent is created implicitly (§4.1), with its own number, and \chemid of its name alone prints it normally, without error:

```
\chemid*{molz.a,molz.b}      % molz was never declared on its own
...
\chemid{molz}                 % -> its number
```

2.2 `\chemid` — using

`\chemid` accepts a comma-separated list of keys and works everywhere: body text, `\caption`, section titles.

Call	Result	Rule
<code>\chemid{benz}</code>	1	plain key
<code>\chemid{benz.cl}</code>	1a	child key
<code>\chemid{benz.cl,benz.br,benz.i}</code>	1a-c	3 contiguous: range
<code>\chemid{benz.cl,benz.br}</code>	1a,b	2 contiguous
<code>\chemid{benz.cl,benz.i}</code>	1a,1c	non-contiguous
<code>\chemid{benz,amine.a}</code>	1 and 2a	different families
<code>\chemid{target}</code>	<code>target_{final}</code>	custom name
<code>\chemid{seriesa,seriesb,seriesc}</code>	4-6	3 consecutive parents: range

The order you give is respected as-is: nothing is sorted. An undeclared key produces a compilation error and prints `??`, like an unresolved `\ref`.

Several plain parents cited in a row contract into a range exactly like several children of the same parent do: `\chemid{seriesa,seriesb}` (only two, below `range-threshold`) still gives 4 and 5, but a gap breaks the range rather than being papered over, e.g. 4 and 6 for `\chemid{seriesa,seriesc}` (`seriesb` not cited in between). A custom name, an undeclared key, or a child key never joins such a run — it simply ends it, exactly as a different family would.

2.3 `\chemidhere` — anchoring

`\chemidhere{key}` places the target of the link. Put it in the figure that shows the compound, typically near the `\caption`:

```
\begin{figure}
  \centering
  \includegraphics{schema}
  \chemidhere{benz}\chemidhere{benz.cl}\chemidhere{benz.br}
  \caption{Halogenation of \chemid{benz}}
\end{figure}
```

It prints nothing and *never* affects the layout, in horizontal mode as well as in vertical mode (this is checked by the test suite). A click brings the reader to the figure; the PDF format does not allow targeting a sub-part of an image, an anchor there being a point, not an area.

Several keys, and whole families, at once

`\chemidhere` takes a list, with the same comma convention as `\chemid` and `\chemid*`. And since a scheme usually shows a parent with all of its children, naming the *parent* anchors the family:

```
\chemidhere{cmbr.4,cmbr.6,cmbr.8} % three keys, three anchors
\chemidhere{cmbr}                 % the same three, plus cmbr itself
\chemidhere{cmbr,cmal}            % two families in one call
```

A child that carries an anchor of its own is left alone by the family anchor — the explicit one wins, and says nothing about it. Children declared *after* the scheme are anchored too, being read back from the `.aux`. Set `anchor-children` to `false` to keep `\chemidhere{parent}` to the parent alone.

Diagnostics

Placing the anchor of the same key twice triggers a warning: the PDF target would be ambiguous. Using a key that was never declared is an error. Finally, a key that is used but never anchored is reported at the end of compilation, like an unresolved `\ref` — and, version 2 having nowhere to send that link, it is printed without one (§8.2).

`\herechemid`, the old name

Version 1 called this command `\herechemid`. It still works, and will keep working, but it was the odd one out among `\chemid`, `\chemidkey`, `\chemidnote`, `\chemidscheme` and `\chemidsetup`; using it says so once per document.

3 What happens in the table of contents and bookmarks

A compound number in a section title must behave differently depending on where that title gets reused.

Context	Behaviour
Body text, caption, displayed title	clickable number
Table of contents, list of figures	number shown, <i>not</i> clickable
PDF bookmarks, metadata	plain text, no formatting

The link is dropped from the table of contents because the entry is *already* a link to the section: two nested links cannot be represented in PDF. The three renderings are produced by the same, fully expandable internal code, which guarantees they cannot diverge.

4 Numbering

4.1 Keys

A key is written `parent` or `parent.child` — a single level of hierarchy; `a.b.c` is an error. Keys are case-sensitive and accept digits, hyphens and underscores (`2nd`, `my_key`, `compound-3`), but no comma: that is the list separator of `\chemid` and `\chemid*`.

Declaring `parent.child` without having declared `parent` creates the latter automatically, with its own number, and reports it in the `.log`. The option `implicit-parent=false` turns this into an error instead.

4.2 Beyond 26 children

Letters keep going past `z`: `aa`, `ab`, ... There is no limit at 26.

4.3 Resetting

`\chemidreset` resets the main counter to zero: the next declaration will start again from 1. This is the only reset offered — no resetting to an arbitrary value, no partial reset. Compounds already declared keep the number they were given.

5 Templated families

A parent numbers its children with letters: **1a**, **1b**, **1c**. That is right for a series of analogues, and wrong for a series whose members are told apart by something the chemist already has a name for — a chain length, an oxidation state, a substituent. A *family* lets the parent hold the stem, and each child carry only what varies:

```
\chemidfamily{lcoum}{ name = Lcoum-C , placeholder = n }
\chemid*{lcoum.four}{4}
\chemid*{lcoum.eight}{8}
\chemid*{lcoum.twelve}{12}
```

`\chemidfamily` declares the parent, exactly as `\chemid*` would — it takes the same place in the numbering — and gives it a shape its children inherit. Inside such a family, the argument that would otherwise *name* a key is read as its *value*: the second argument of `\chemid*{lcoum.four}{4}` is the 4, not a compound called “4”.

Written	Printed
<code>\chemid{lcoum.four}</code>	Lcoum-C₄
<code>\chemid{lcoum.four,lcoum.eight}</code>	Lcoum-C_{4,8}
<code>\chemid{lcoum.four,lcoum.eight,lcoum.twelve}</code>	Lcoum-C_{4,8,12}
<code>\chemid{lcoum}</code>	Lcoum-C_n

The stem is printed *once* for a whole group and the values are gathered in a single subscript, just as **1a,b** factors out the 1. Each value keeps its own link, and the stem links to the first of them.

5.1 The options of a family

Option	Default	Role
<code>name</code>	(empty)	the stem; without one, the parent’s number is used
<code>raw</code>	derived	its PDF-string form, as for <code>\chemid*</code>
<code>sub-style</code>	<code>sub-style</code>	numbering of the children that carry no value
<code>child-format</code>	<code>subscript</code>	<code>subscript</code> , <code>superscript</code> , <code>plain</code> , or <code>code</code>
<code>placeholder</code>	(empty)	what <code>\chemid{parent}</code> shows in place of a value
<code>placeholder-raw</code>	derived	its PDF-string form

`child-format` takes one of the three keywords or, for anything they cannot express, the code itself, with `\chemidvalue` standing where the value belongs:

```
\chemidfamily{cu}{ name = Cu , child-format = \textsuperscript{\chemidvalue} }
```

5.2 Children numbered by the family

A child declared with no value of its own is numbered by the family's **sub-style** — including **lower-greek**, which is what such series are usually labelled with:

```
\chemidfamily{ser}{ name = Series , sub-style = lower-greek ,  
                    child-format = plain }  
\chemid*{ser.a}\chemid*{ser.b}\chemid*{ser.c}
```

`\chemid{ser.a}` gives *Series* α , and `\chemid{ser.a,ser.b,ser.c}` gives *Series* α - γ : automatically numbered children are consecutive, so they contract into a range like any other run. In PDF bookmarks, where there is no maths to be had, the Greek letter appears as the character itself.

Values given by hand never contract: 4, 8 and 12 have no notion of “consecutive”, and 4-12 would claim something nobody wrote. They are always listed one by one.

5.3 One family, one kind of child

Inside a family, either every child carries a value of its own or none does. Mixing the two is an error: the ones in between would have to be numbered by a rule the others visibly do not follow, and no reading of the list would be honest. The child is still declared — a diagnostic must not shift the numbering of everything after it.

6 Options

They are given at load time, or at any point with `\chemidsetup{...}`; in the latter case they follow the scope of the current $\text{\TeX}$ group.

Option	Default	Role
<code>list-sep</code>	<code>{, }</code>	between two families
<code>last-sep</code>	<code>{ and }</code>	before the last family
<code>sub-sep</code>	<code>{,}</code>	between letters of the same family
<code>range-sep</code>	<code>{-}</code>	inside a range <code>1a-c</code>
<code>range-threshold</code>	<code>3</code>	size from which a range is contracted
<code>format</code>	(empty)	formatting applied to each identifier, e.g. <code>\textbf</code>
<code>main-style</code>	<code>arabic</code>	<code>arabic</code> , <code>alph</code> , <code>Alph</code> , <code>roman</code> , <code>Roman</code>
<code>sub-style</code>	<code>alph</code>	same, plus <code>lower-greek</code> , for the letter
<code>prefix</code>	<code>{chemid.}</code>	prefix of PDF anchors
<code>unknown-text</code>	<code>{??}</code>	shown for an undeclared key
<code>purify</code>	<code>true</code>	derive the raw name from the rich name
<code>implicit-parent</code>	<code>true</code>	create a missing parent automatically
<code>strict-anchors</code>	<code>true</code>	no anchor, no link — see §8.2
<code>links</code>	<code>false</code>	<code>true</code> adds hyperlinks and anchors
<code>anchor-children</code>	<code>true</code>	<code>\chemidhere{parent}</code> anchors the family
<code>auto-lang</code>	<code>true</code>	see §7
<code>version</code>	<code>2</code>	which implementation to load, §1.1

The chemistry-journal typographic convention — bold numbers — is obtained as follows:

```
\usepackage[format=\textbf]{chemidentifier}
```

7 Multilingual documents

The separator before the last item is, by default, the English `and` (`last-sep`). To choose a fixed language, without `babel`:

```
\usepackage[last-sep={~et~}]{chemidentifier} % at load time
\chemidsetup{last-sep={~et~}}                % or at any time
```

In a thesis mixing several languages with `babel`, the separator automatically follows the current language — the one that `\selectlanguage` just set — without anything further needed:

```
\usepackage[french,ngerman,spanish,italian,english]{babel}
...
\selectlanguage{french}   \chemid{a,b}   % ... et ...
\selectlanguage{ngerman} \chemid{a,b}   % ... und ...
\selectlanguage{spanish} \chemid{a,b}   % ... y ...
\selectlanguage{italian} \chemid{a,b}   % ... e ...
\selectlanguage{english} \chemid{a,b}   % ... and ...
```

The mechanism reads `\language`, which `babel` updates on every `\selectlanguage`, in a fully expandable way: it therefore works identically in the text, the table of contents and PDF bookmarks. Without `babel` loaded, or for an unregistered language, the package silently falls back to the `last-sep` option.

Five languages are known by default, with their usual `babel` variants:

Language	Separator	Recognized babel names
English	and	english, american, british, australian, UKenglish, USenglish
French	et	french, francais, acadian, canadien
German	und	german, ngerman, austrian, naustrian
Spanish	y	spanish, mexican
Italian	e	italian

To add or redefine a language:

```
\chemidaddlanguage{portuguese}{ e }
```

To disable the automatic switch — a fixed separator, whatever the current language — set `auto-lang=false`; `last-sep` then becomes the sole source again, as before this feature existed. Like any option, this can be done locally inside a group:

```
\begingroup
  \chemidsetup{auto-lang=false,last-sep={ or else }}
  \chemid{a,b} % ... or else ...
\endgroup
\chemid{a,b} % the language takes over again
```

8 Two points worth knowing

8.1 The raw name derived automatically

When you give a rich name without its raw counterpart, the latter is derived using `expl3`'s `\text_purify:n`: `H\textsubscript{2}O` becomes `H2O`. The result is good for common cases (subscripts, superscripts, bold, italics) but **is not guaranteed** for nested mathematics. For a non-trivial rich name, provide the raw version yourself:

```
\chemid*{k}{\alpha$-D-glucopyranose}[alpha-D-glucopyranose]
```

8.2 A number is a link only when it has somewhere to go

A compound that is never anchored is printed as a plain number: no link at all, rather than one landing wherever the reader's viewer decides to put it. This is `strict-anchors`, and it is the default in version 2.

For that, the package must know, at the moment it prints a number, whether an anchor for it exists *anywhere* in the document — which it cannot know on the way in. Anchors are therefore recorded in the `.aux` and read back on the next run: the first compilation links only what it has already seen anchored, the second links everything, and it stays that way afterwards. The same trade-off as `\ref/\label`, for the same reason.

A parent is a case worth knowing about. A scheme often anchors only the children — `\chemidhere{parent.a}`, `\chemidhere{parent.b}` — while the text also cites the parent on its own. Rather than drop that link, the package sends it to the *first anchored child*: `\chemid{parent}` lands on the same figure, which is where the reader wanted to go. With `\chemidhere{parent}` (§2.3) the question does not even arise, the parent being anchored as well.

Set `strict-anchors` to `false` to go back to version 1's behaviour: the link is always written, and only the missing anchor is reported at the end of the compilation.

8.3 Turning links on

`links` defaults to `false`: compounds are plain numbers, with no `\hyperlink` and no `\hypertarget` either, not even where a `\chemidhere` asks for one. The end-of-document report about compounds without an anchor stays quiet as well, having nothing left to protect. Numbering, names and every other rendering are unaffected either way.

Set `links=true` — as a package option or via `\chemidsetup` — to get the whole hyperlinking layer described above: clickable numbers, anchors, and the end-of-document warning for a compound used without one.

```
\usepackage[links=true]{chemidentifier}  
\chemidsetup{links=true}      % or from here on, or inside a group
```

9 Declaring after using

`\chemid` and `\chemidhere` may appear *before* the `\chemid*` that declares the key — useful when you don't want to be forced to declare everything at the top of the document, in particular when the table of contents precedes the text:

```
Compound \chemid{mol1} is mentioned here, before its declaration.  
...  
\chemid*{mol1}
```

Every declared key is recorded in the `.aux` at the end of compilation (number, letter, rich name, raw name), and read back on the next `\begin{document}` — the same mechanism as `\label/\ref`, with the same trade-off: the first compilation shows ?? for any reference ahead of its declaration, the second resolves it, and the result stays stable afterwards, since the order of declarations never depends on where they are used. A key still unknown at the end of the second pass is genuinely undeclared: the error then keeps showing.

If you have no particular constraint, declaring at the top of the preamble remains the simplest approach: the number is then known as soon as the key is read, with no forward reference to resolve.

10 Text substitution in `.pdf_tex` figures (LuaLaTeX only)

A scheme exported by Inkscape (a `.pdf_tex` + `.pdf` pair) can have its compound labels wired to `\chemid`, the same way `psfrag` once patched text into `.eps` figures — without the `.eps/psfrag` baggage: a `.pdf_tex` is just L^AT_EX text calling `\includegraphics`, so a plain, line-by-line substitution is enough. This requires compiling with `lualatex` (or another engine providing `\directlua`): the substitution is implemented in Lua, reading the file as plain text and handing the result back to T_EX once the placeholders are replaced. The file on disk is never modified, so re-exporting from Inkscape loses nothing.

A plain-T_EX catcode-trick alternative (à la `psfrag`) was considered and rejected: a `.pdf_tex` is genuine T_EX/PGF code, full of `\`, `{`, `}`, `%`, `#`, `_`, `~`... Reading it verbatim and then re-executing it as code requires the same character to carry two contradictory catcodes — inert while captured, active while replayed — for arbitrary content, which is precisely what made `psfrag` fragile. Lua sidesteps this entirely by treating the file as a plain string throughout, and handing the result to T_EX only at the very end, read under T_EX's own, ordinary catcode regime.

Put a plain-text placeholder in the drawing for each compound label (TMP1, TMP2... anything, as long as it does not also occur as a substring elsewhere in the figure text), then:

```
\usepackage{graphicx}           % needed by the .pdf_tex itself
\usepackage{chemidentifier}
\chemidsetup{ pdftex-font = \sffamily\small } % optional, once

\chemid*{precursor}
\chemid*{product}

\begin{figure}
  \centering
  \chemidkey{TMP1}{precursor}      % TMP1 -> \chemid{precursor}
  \chemidkey[1.3]{TMP2}{product}  % 1.3x bigger than the rest
  \chemidnote{TMPCOND}{K$_2$CO$_3$, acetone, 70~\textcelsius}
  \chemidscheme[0.8]{figures/scheme.pdf_tex} % scale is optional
  \chemidhere{precursor}\chemidhere{product}
  \caption{Synthesis of \chemid{product} from \chemid{precursor}.}
\end{figure}
```

Command	Role
<code>\chemidkey[factor]{motif}{key}</code>	placeholder → the compound's current number, e.g. <code>\chemid{key}</code>
<code>\chemidnote[factor]{motif}{text}</code>	placeholder → any other text, free-form
<code>\chemidscheme[scale]{path}</code>	reads the file, substitutes, typesets; also extends <code>\graphicspath</code>

`\chemidscheme` consumes the pending `\chemidkey`/`\chemidnote` list as it reads the file, so the next figure automatically starts from an empty list — there is no separate "clear" step to remember. An empty list is perfectly valid: a scheme with no substitutable label at all (a lone structure, a graphical abstract, labels already typeset in the drawing) is simply read and typeset verbatim, and is still worth passing through `\chemidscheme` for its `\graphicspath` and font handling.

The line including the companion PDF is never substituted, only the label text around it. The natural motif is the compound's own name, and Inkscape names the exported PDF after the compound too, so `\chemidkey{KaBMD244}{...}` on a `KaBMD244.pdf_tex` would otherwise rewrite the file name itself and break the figure; when a motif does occur in the image path, the package says so by name rather than letting it fail obscurely. The warning is harmless if the figure looks right, but a motif that cannot collide (TMP1, cpdA...) remains the safer habit.

The size of a substituted label is the product of three independent factors, so a figure scaled down still reads fine without retouching every label by hand:

1. `pdftex-font` (`\chemidsetup`) – the base style, set once, independently of the body text (Inkscape otherwise reinjects the document's own family, at whatever size the drawing used, typically too large), e.g. `\chemidsetup{ pdftex-font = \sffamily\fontsize{9}{11}\}`
2. the `[scale]` of `\chemidscheme` – the same number handed to the figure's own `\svgscale`, so labels shrink/grow together with the drawing.
3. the optional `[factor]` of `\chemidkey`/`\chemidnote` – one label singled out, relative to the others in the same figure; defaults to 1.

Under a non-Lua engine, `\chemidscheme` raises a clear error instead of silently doing nothing or mis-rendering. `\chemidkey` takes the compound's *current* number: it does not declare or renumber anything, so keys still need a matching `\chemid*` declared elsewhere.

11 Assumed limitations

- A single level of hierarchy (`parent.child`), families included.
- Lists are neither sorted nor deduplicated: what you write is what gets printed — `\chemid{a,b,a}` prints 1, 2 and 1, without error.
- `cleveref` is not handled (out of scope).
- A key removed from the document keeps resolving to its last known value until the `.aux` file is cleared — the same limitation as `\ref/\label`.
- The optional arguments of `\chemid*` are only recognized when *glued* to the key, with no space or line break at all: `\chemid*{k}{rich}[raw]`. Any space before `{` or `[` excludes them from being read; the content that follows is then treated as ordinary text, never absorbed.