Project ChEMU

Guidelines for Annotating Chemical Reaction References in Chemical Patents

Version <1.00>

Revision History

Version Number, Date	Revision made by	Change Summary
1.00, 4 Dec 2020	Jiayuan He	Created
1.01, 8 Dec 2020	Christian Druckenbrodt	Revised
1.02, 9 Dec 2020	Jiayuan He	Revised
1.03, 10 Dec 2020	Christian Druckenbrodt	Revised
1.04, 11 Dec 2020	Jiayuan He	Revised
1.05, 01 Feb 2020	Christian Druckenbrodt	Revised

Contents

SCOPE OF THE GUIDELINE

This document is the guideline for the Full-Text Patent Annotation (FTPA) project and should advise how to annotate manually chemical reactions and their general conditions in patents correctly. In addition, the reference relationships amongst the annotated chemical reactions and general conditions should be marked. It will provide definitions, and also examples, in which cases and how a string of characters embedded in text of patents must be annotated accordingly. The resulting manually annotated BRAT files can be used as a so-called “gold standard” in order to determine Recall, Precision and F1-scores regarding corresponding automatic annotations, produced by state-of-the-art text mining tools.

OVERVIEW

Chemical reactions described in chemical patents contain valuable information about the detailed synthesis processes of new chemical compounds filed in the patent. Recognizing and understanding chemical reactions is a crucial task for information extraction over chemical patents. However, it is often the case in chemical patents that some chemical reactions (e.g., reaction A) are not fully described, and readers need to refer to their contextual reactions (e.g., reaction B) within the same patent to understand these reactions. In such cases, the reaction A is considered as *referring to* reaction B. Identification of chemical reactions and more importantly, sorting out the referring relationships amongst the chemical reactions, is the goal of this project. Note that in this project, the focus is on chemical reactions in textual and tabular form. Any drawn chemical entities/structures are excluded, as state-of-the-art technologies are currently not able to reach sufficient quality in such regions of documents.

Identifying the exact text spans of chemical reactions in patents is challenging. This is because many chemical reactions are not fully described. Instead, they are presented in simplified/structured ways to keep the patents short and concise. For example, some reaction may only be presented as a set of key compounds which can be used to complete a *general condition*, i.e., a certain synthesis procedure for preparing many similar reactions. The combination of the general condition and the provided key compounds form a complete description of the reaction. In this task, these simplified forms of chemical reactions and general conditions need to be annotated and connected, as our goal is to provide complete understanding of all reactions in patents.

The challenge for annotating reference relationships amongst chemical reactions and general conditions lies in the fact that such reference relationships are not always expressed explicitly. In some

cases, the reference relationships are only implied by the document structures, such as tables and reactions indices. Another challenge is that the reference relationships amongst chemical reactions may span very long texts. Capturing reference relationships between distant reactions pose challenges to both annotation and the design of information extraction systems.

INTRODUCTION TO PATENTS

A patent is a right granted to the owner of an invention that prevents others from making, using, importing or selling the invention without his permission. A patentable invention can be a product or a process that gives a new technical solution to a problem. It can also be a new method of doing things, the composition of a new product, or a technical improvement on how certain objects work. For an invention to be patentable, it must, in general, satisfy three key criteria: New, inventive step, and industrial application.

The sections of patents are quite conserved: title, bibliographic information (patent number, dates, inventors, assignees, IPC classes, …), abstract, description, and claims. Most of the chemical data will be found in the experimental section of the description, while compounds claimed (protected) are available in the claims section. Drawings, sequences, or other additional information will normally be found at the very end of a patent.

This particular project focuses on the free texts in the patents and do not include any drawings. Tables in original patents are included but are converted in a simplified form (see Section “Reactions in Tables”).

HIGH LEVEL RULES

In general, there are two major types of chemical reaction references: (1) analogous reactions, and (2) general conditions.

ANALOGOUS REACTIONS. It is often the case in chemical patents that a reaction procedure (e.g., to synthesize Compound A1) is described first and then a number of similar reactions (e.g., to synthesize Compound A2, A3, ...) are listed afterwards without full description. The former reaction (for Compound A1) is called the *parent reaction*, and the latter (for Compound A2, A3, ...) is called the *child reaction*. The pair of parent and child reactions, or this type of reference, is called *analogous reactions*.

To give an example, Figure 1 presents the snapshot of a patent in the annotation interface, where the parent text is presented with paragraph segmentation and the backgrounds of consecutive paragraphs

are toggled between grey and white. In the figure, there are two chemical reactions highlighted in blue, **T1** and **T2**. **T2** describes how to prepare the chemical compound "1(S)-Benzyl-6-methoxy-1(R)-(3-oxo-butyl)-3,4-dihydro-1H-naphthelen-2-one". However, since its preparation method is similar with that in "Preparation 2", **T2** refers to the reaction **T1** and only describes its differences from **T1**. In this figure, **T1** and **T2** form a pair of analogous reactions with **T1** being the parent reaction and **T2** being the child reaction, respectively.

T1

181 Preparation 2 1(R)-Benzyl-6-methoxy-1(S)-(3-oxo-butyl)-3,4-dihydro-1H-naphthelen-2-one

182 A solution of 62 g (0.23 mol) of the title product of Preparation 1 and 28 mL, (0.23 mol) of freshly distilled (S)-(-)-alpha-methyl benzylamine in 100 mL of toluene was heated to reflux, over a Dean-Stark trap, overnight. After removal of the azeotroped water, the imine solution was cooled to 0 °C and 21 mL (0.26 mol) of freshly distilled methylvinylketone was added dropwise to the solution. The solution was stirred at 0 °C for 30 min then heated to 40 °C overnight. The reaction solution was cooled to 0° C and 17 mL of acetic acid and 14 mL of H2O were added and the resultant solution was allow to warm to RT for 2 h. The solution was poured into H2O and extracted three times with EtOAc. The combined organic layers were washed with 1 N HCl, H2O, saturated NaHCO3, then dried over Na2SO4, filtered, and evaporated to dryness. The crude product was purified by chromatography over SiO2 using 15% EtOAc to 35% EtOAc in hexanes as the gradient eluant to give 48 g of the title product of this preparation as a yellow solid. 1HNMR (400 MHz, CDCL3) δ 1.38 (s, 3H), 1.40-1.51 (m, 2H), 1.64 (ddd, 1H, J = 2.1, 4.5, 13), 1.97 (broad s, 1H), 2.20 (dt, 1H, J = 4.5, 13), 2.59 (d, 1H, J = 6.6), 3.08 (d, 1H, J = 18), 3.16 (d, 1H, J = 16), 3.33 (dd, 1H, J = 6.6, 18), 3.62 (d, 1H, J = 16), 3.72 (s, 3H), 6.57 (d, 1H, J = 2.5), 6.67 (dd, 1H, J = 2.5, 8.8), 7.00-7.23 (m, 6H); 13C NMR (100 MHz, CDCl3) δ 27.90, 32.79, 34.40, 38.43, 41.49, 53.51, 55.12, 58.47, 79.06, 112.05, 113.09, 125.37, 127.63, 127.69, 130.27, 132.21, 135.45, 138.65, 157.88, 213.49; MS m/z 337 (M+H)+, 319 (M-OH)+.

T2

183 Preparation 3 1(S)-Benzyl-6-methoxy-1(R)-(3-oxo-butyl)-3,4-dihydro-1H-naphthelen-2-one

184 The title product of this preparation was prepared using a method analogous to Preparation 2, using (R)-(+)-alphamethyl benzylamine in the initial imine formation. Starting with 4.64 g 1-benzyl-6-methoxy-3,4-dihydro-1H-naphthalen-2-one produced 3.58 g of the title product of this preparation as a yellow solid. 1HNMR (400 MHz, CDCL3) δ 1.38 (s, 3H), 1.40-1.51 (m, 2H), 1.64 (ddd, 1H, J = 2.1, 4.5, 13), 1.97 (broad s, 1H), 2.20 (dt, 1H, J = 4.5, 13), 2.59 (d, 1H, J = 6.6), 3.08 (d, 1H, J = 18), 3.16 (d, 1H, J = 16), 3.33 (dd, 1H, J = 6.6, 18), 3.62 (d, 1H, J = 16), 3.72 (s, 3H), 6.57 (d, 1H, J = 2.5), 6.67 (dd, 1H, J = 2.5, 8.8), 7.00-7.23 (m, 6H); 13C NMR (100 MHz, CDCl3) δ 27.90, 32.79, 34.40, 38.43, 41.49, 53.51, 55.12, 58.47, 79.06, 112.05, 113.09, 125.37, 127.63, 127.69, 130.27, 132.21, 135.45, 138.65, 157.88, 213.49; MS m/z 337 (M+H)+, 319 (M-OH)+.

Figure 1. An example of analogous reactions.

GENERAL CONDITIONS. In some cases, the conditions to synthesize many similar compounds are described without specifying the actual name of compounds. Such a description is referred as a description of a general condition. In chemical patents, general conditions are often presented using Markush structures. Thereafter, any child reactions that follow this general procedure are usually briefly summarized and pointed to the general conditions.

In Figure 2, a general condition is first given including several reaction steps, i.e., "(1) Step (A)", "(2) Steps (B) and (C)", and "(3) Preparation of Salt". The general condition can be applied to all its following examples such as Example 1 and Example 2 in the figure. The difference from the case of analogous reactions is that the description to be referred to here is not associated with any specific reaction. In

this particular example, the general conditions are described using an image with Markush structures, with a variable *X* that can be replaced with several substructures, as shown in Scheme 1a. Such Markush structures are common in patents.

T1

(1) Step (A)

4-Bromobenzaldehyde and boronic acid were subjected to Suzuki cross coupling reaction using a palladium catalyst as shown in [Scheme 1a]. Specifically, 4-bromobenzaldehyde (3 g, 16.21 mmol), boronic acid (1.28 equivalents),tetrakis(triphenylphosphine)palladium(0) (4-8 mol%) andsodium carbonate (4.86 equivalents) were refluxed in degassedtoluene/distilledwater (150 mL/21.6 mL) for18 hours while heating. The reaction mixture was filtered through celite and the filtrate was washed twice with ethyl acetate (200 mL) and water (200 mL). The organic layer was combined and dried with sodium sulfate, concentrated in vacuo and then separated and purified by silica gel column chromatography.<img> id-imgb0011.tif </img>

T2

(2) Steps (B) and (C)

An imine compound was obtained by subjecting the compound of the step (A) to reductive amination using L-alaninamide hydrochloride or D-alaninamide hydrochloride (step (B), Scheme 1b). Then, an amine compound was obtained by reducing the imine compound with sodium cyanoborohydride (step (C), Scheme 1c).

After adding 1.2 equivalents of glycinamide hydrochloride or L-alaninamide hydrochloride or D-alaninamide hydrochloride or L-valinamide hydrochloride or L-leucinamide hydrochloride to anhydrousmethanol to a concentration of 0.92 M, 1.5 equivalents of triethylamine was added. When the solution became transparent, 1.0 equivalent of the aldehyde synthesized in the step (A) was added. Two hours later, the solution was washed with ethyl acetate and distilled water. After drying the organic layer with sodium sulfate and drying in vacuo, the concentrated reaction solution was dissolved in anhydrous methanol to a concentration of 1.0 M and then 4.0 equivalents ofsodium cyanoborohydride was added at0 °C. After performing reaction atroom temperature for18 hours, the reaction solution was washed with ethyl acetate and distilled water. The organic layer was dried with sodium sulfate, concentrated in vacuo and then separated and purified by silica gel column chromatography.<img> id-imgb0012.tif </img>

T3

(3) Preparation of salt

The salt preparation step is an optional step that can be either performed, if necessary, or omitted. A compound in salt form is synthesized to improve the solubility of the amine compound synthesized in the preceding step. The compound in salt form may be synthesized using an acid. The acid that can be used is described above but is not limited thereto.

Specifically, a compound in salt form was synthesized usingmethanesulfonic acid. After heatingethyl acetate to50-55 °C and completely dissolving 1.0 equivalent of the compound of the step (C), 1.25 equivalents of methanesulfonic acid was added.1 hour later, the reaction mixture was cooled toroom temperature and filtered using a vacuum filtration device. The filtrate was washed with ethyl acetate and dried without a purification process.<img> id-imgb0013.tif </img>

T4

Example 1: Synthesis of (S)-2-(((2'-fluorobiphenyl-4-yl)methyl)amino) propanamide methanesulfonate

<img> id-imgb0014.tif </img>

White solid; yield: 90%; 1H NMR (300 MHz, DMSO-d6) δ 9.17 (br s, 2H), 7.94 (br s, 1H), 7.30-7.94 (m, 9H), 4.16 (m, 2H), 3.80 (q, J = 6.54 Hz, 1H), 2.30 (s, 3H), 1.45 (d, J = 6.93 Hz, 3H); 13C NMR (75 MHz, DMSO-d6) δ 170.9 (C(O)), 161.2, 157.9, 136.2, 131.7, 131.2, 131.1, 130.8, 130.5, 130.3, 129.5, 129.4, 128.1, 127.9, 125.5, 125.4, 116.8, 116.5 (ArC), 55.1 (C(O)CH+NH2), 48.7 (+NH2CH2Ph), 16.4 (CH3). SCH3 signal overlapping with DMSO signal.

T5

Example 2: Synthesis of (S)-2-(((3'-fluorobiphenyl-4-yl)methyl)amino) propanamide methanesulfonate

<img> id-imgb0015.tif </img>

White solid; yield: 97%; 1H NMR (300 MHz, DMSO-d6) δ 9.15 (br s, 2H), 7.92 (br s, 1 H), 7.81 (d, J = 8.25 Hz, 2ArH), 7.68 (br s, 1 H), 7.49-7.60 (m, 5ArH), 7.20-7.27 (m, 1ArH), 4.15 (s, 2H), 3.76 (q, J = 9.24 Hz, 1 H), 2.30 (s, 3H), 1.44 (d, J = 9.28 Hz, 3H); 13C NMR (75 MHz, DMSO-d6) δ 171.0 (C(O)), 164.9, 161.8, 161.6, 142.4, 142.3, 139.8, 132.0, 131.5, 131.4, 131.2, 127.5, 123.3, 115.2, 114.9, 114.1, 113.8 (ArC), 55.0 (C(O)CH+NH2), 48.6 (+NH2CH2Ph), 16.4 (CH3). SCH3 signal overlapping with DMSO signal.

[Scheme 1a]

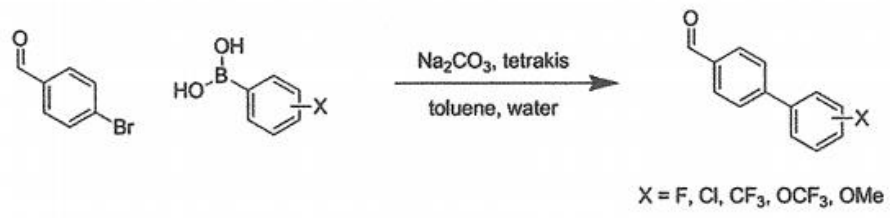


Figure 2. An example of general conditions.

This example also uses the name *boronic acid* not for the compound itself but indicating its chemical class; the actual name of the acid is not used, and an inference process is needed to identify the specific chemical compounds used in each reaction, requiring expert knowledge in chemistry.

SUMMARY

The target of this annotation is to capture the reaction references in the above examples. The objectives of the annotation include: (1) identifying all chemical reactions/general conditions in patents; and (2) linking chemical reactions to all relevant chemical reactions/general conditions that are required to fully specify the reactions.

This document describes the annotation of chemical reaction references in chemical patents. Specifically, the task requires annotating the text spans of all chemical reactions and general reactions occurring in given patents, the reference relationships between analogous reactions/general conditions, and the cue statements that are associated with the reference relationships.

In what follows, guidelines for annotating chemical reactions/general conditions, reference relationships, and cue statements are described.

Part 1: Reaction Span Annotation

CHEMICAL REACTION

The text span of a chemical reaction is the reaction full text with its attached reaction title, but without its analytical data.

This means reaction title such as “Example 1”, “Step 1”, or “Compound A” need to be included. For example, the reaction title “Step 1” needs to be included in the annotated span in Figure 3.

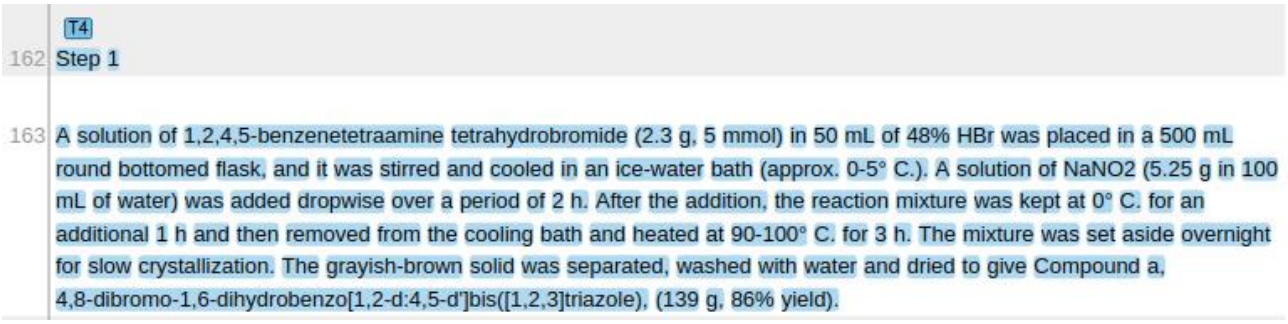


Figure 3. The title of a reaction must be included in the reaction span.

However, any analytical data must not be included. For example, the analytical data of **T1** in Figure 4, i.e., “1H NMR … (d, J=6.5 Hz, 18H)”, should not be included in the text span of **T1**.

T1
248 A mixture of the obtained ester, pentaerythrityl tetrabromide (5.82 g, 15 mmol), potassium carbonate (10.35 g, 75 mmol), and DMF (40 mL) was stirred under argon and heated at 110° C. for 24 h. After cooling, the mixture was poured into ice/water (200 mL) and extracted with hexane/ethyl acetate (200+200 mL). The extract was washed with water (200 mL), dried over anhydrous magnesium sulfate, and the solvent was removed under reduced pressure. Column chromatography of the residue (silica gel-hexane/ethyl acetate, 9:1) gave Compound B11, diisobutyl 3,3'-((2-(bromomethyl)-2-((3-(isobutoxycarbonyl)phenoxy)methyl)propane-1,3-diyl)bis(oxy))dibenzoate, (4.27 g, 39% yield). 1H NMR (500 MHz, CCCl3): δ 7.65 (d, J=8.0 Hz, 3H), 7.59 (m, 3H), 7.33 (t, J=8.0 Hz, 3H) 7.11 (dd, J=2.0 and 7.5 Hz, 3H), 4.31 (s, 6H), 4.10 (d, J=6.5 Hz, 6H), 3.90 (s, 2H), 2.08 (m, 3H), 0.97 (d, J=6.5 Hz, 18H).

Figure 4. Analytical data must not be included in reaction spans.

There may exist non-textual data within reaction spans, e.g., images. This data needs to be included as long as it is part of the reactions. In this project, images in the original patents are converted and displayed in a plain textual form. As shown in Figure 5, the image named “id-imgb0011.tif” is displayed as its figure name enclosed by a pair of tags “<img>” and “</img>”, without the actual content of the image. When annotating the span of **T1**, such non-textual information should be **included**.

T1
61 (1) Step (A)
62 4-Bromobenzaldehyde and boronic acid were subjected to Suzuki cross coupling reaction using a palladium catalyst as shown in [Scheme 1a]. Specifically, 4-bromobenzaldehyde (3 g, 16.21 mmol), boronic acid (1.28 equivalents),tetrakis(triphenylphosphine)palladium(0) (4-8 mol%) andsodium carbonate (4.86 equivalents) were refluxed in degassedtoluene/distilledwater (150 mL/21.6 mL) for18 hours while heating. The reaction mixture was filtered through celite and the filtrate was washed twice with ethyl acetate (200 mL) and water (200 mL). The organic layer was combined and dried with sodium sulfate, concentrated in vacuo and then separated and purified by silica gel column chromatography.<img> id-imgb0011.tif </img>

Figure 5. Non-textual data (image in this figure) should not be included in the reaction span.

REACTION SEQUENCE

In some cases, the synthesis process of a chemical compound is an ordered sequence of chemical reactions. The whole sequence of reactions called a *reaction sequence* and each reaction in the sequence called a *reaction step*.

Figure 6 presents an example of a reaction sequence. When annotating a reaction sequence, each reaction step should be treated as a chemical reaction with an independent text span. In Figure 6, for example, the preparation for Example 1 consists of three reaction steps, “1.1”, “1.2”, and “1.3”. The three reaction steps must be annotated as independent chemical reactions.

168 Example 1 The preparation of 2-fluoro-5- (pyridin-2-ethynyl) -N- (4-fluorophenyl) benzamide (ZD001)

T1

169 1.1 Synthesis of 2-fluoro-5-iodo-benzoyl chloride

170 500 mg of 2-fluoro-5-iodobenzoic acid was added to a 50 ml eggplant flask, and then 3 ml of thionyl chloride was added, and heated at 77 °C for 2 hours. The reaction was monitored by thin layer chromatography (TLC). After the reaction was completed, the mixture was cooled to room temperature and dried by rotary evaporation to remove thionyl chloride to give 524 mg 2-fluoro-5-iodo-benzoyl chloride as colorless liquid.

T2

171 1.2 Synthesis of 2-fluoro-N-(4-fluorophenyl)-5-iodobenzamide

172 200 mg of 4-fluoroaniline was dissolved in 5 ml of ethyl acetate, and 260 μl of triethylamine was added. Then 2-fluoro-5-iodo-benzoyl chloride in ethyl acetate was added dropwise under ice-cooling, and the reaction was completed after 1.5 hours. 10ml ethyl acetate was added to dilute, and 20 ml water was added to extract. The mixture was extracted with ethyl acetate for three times, washed once with saturated brine, dried over anhydrous sodium sulfate, and dried by rotary evaporation to obtain 620 mg 2-fluoro-N-(4-fluorophenyl)-5-iodobenzamide as light yellow solid.

T3

173 1.3 Synthesis of final product ZD001

174 625 mg 2-fluoro-N- (4-fluorophenyl) -5-iodobenzamide was dissolved in toluene, 1.5 eq 2-ethynylpyridine and 2.2 eq triethylamine were added followed by 0.2 eq cuprous iodide, 0.2 eq bis(triphenyl-phosphine)palladium dichloride. The mixture was heated and stirred at 100 °C for 6 hours under an inert atmosphere. The reaction liquid was dried by rotary evaporation and purified to give 460 mg ZD001 as tan solid, yield 79%. 1H NMR (400 MHz, CDCl3) δ 10.61 (s, 1H), 8.62 (d, J = 4.1 Hz, 1H), 7.95 - 7.80 (m, 3H), 7.78 - 7.71 (m, 2H), 7.68 (d, J = 7.8 Hz, 2H), 7.51 - 7.41 (m, 2H), 7.22 (t, J = 8.9 Hz, 2H). LRMS (EI) m/z 335(M+).

T4

175 Example 2 The preparation of (2-chloro-5- (pyridin-2-ethynyl) phenyl) (7-oxa-2-aza-spiro [3.5] nonan-2-yl) methanone (ZD002)

176 2-fluoro-5-iodobenzoic acid was replaced by 2-chloro-5-iodobenzoic acid, and 4-fluoroaniline was replaced by 7-oxa-2-azaspiro [3.5] nonane, while the remaining raw materials, reagents and the preparation method were the same as those in Example 1 to give the product ZD002, yield 80%. 1H NMR (400 MHz, CDCl3) δ 8.62 (m, J = 4.9, 1.8, 0.9 Hz, 1H), 7.70 (m, J = 7.7, 1.8 Hz, 1H), 7.56 - 7.55 (m, 1H), 7.54 - 7.50 (m, 2H), 7.40 (dd, J = 8.1, 0.7 Hz, 1H), 7.30 - 7.26 (m, 1H), 3.94 (s, 4H), 3.65 - 3.49 (m, 4H), 1.89 - 1.70 (m, 4H). LRMS (EI) m/z367(M+).

Figure 6. An example of a reaction sequence.

GENERAL CONDITIONS

General conditions are chemical reactions/conditions which do not specify detailed compound names and serve as general procedures to synthesize many similar compounds. The rules for annotating general conditions are the same as other chemical reactions. In the cases where a general condition is made up by a series of reaction steps, follow the instructions of annotating reaction sequences.

REACTIONS IN TABLES

The source of the relevant information of the reaction is often explicitly mentioned in the description of the child reaction (as in Figure 1). However, it is also sometimes expressed implicitly by the document structure such as in the form of images and tables.

In this task, only chemical reactions in *tables* are considered. All tables are converted presented in a simplified textual form. A table starts with the tag “<table>” and ends with the tag “</table>”. Each row in the table is presented as one line, where the elements in the row are separated by the delimiter “&”.

A header row is enclosed by the pair of tags “<header>” and “</header>”. An example of a table in this simplified textual form is presented in Figure 7. Note that this table does not contain any reactions.

```
<table>
<header>
& Stereo & R & X & MAO-B (IC50, μm) & MAO-A (IC50, μm)
</header>
Example 1 & S & CH3 & 2'-F & > 10 & > 100
Example 2 & S & CH3 & 3'-F & > 10 & > 100
Example 3 & S & CH3 & 4'-F & > 10 & > 100
Example 4 & S & CH3 & 2'-Cl & > 10 & > 100
Example 5 & S & CH3 & 3'-Cl & 0.442 & > 100
Example 6 & S & CH3 & 4'-Cl & 0.416 & > 100
Example 7 & S & CH3 & 2'-CF3 & > 10 & > 100
Example 8 & S & CH3 & 3'-CF3 & 0.316 & > 100
Example 9 & S & CH3 & 4'-CF3 & 0.042 & >500
Example 10 & S & CH3 & 3'-OCF3 & 0.216 & > 100
Example 11 & S & CH3 & 4'-OCF3 & 0.098 & > 100
Example 12 & S & CH3 & 3'-OCH3 & 3.33 & > 100
Example 13 & S & CH3 & 4'-OCH3 & 1.06 & > 100
Example 14 & R & CH3 & 3'-F & > 10 & > 100
Example 15 & R & CH3 & 4'-CF3 & 0.082 & > 100
Example 16 & S & H & 4'-CF3 & 0.126 & > 100
Example 17 & S & CH(CH3)2 & 4'-CF3 & 4.073 & > 100
Example 18 & S & CH2CH(CH3)2 & 4'-CF3 & 5.302 & > 100
S-Safinamide & - & - & - & 0.12 & > 100
Selegilne & - & - & - & 0.009 & ~1
</table>
```

Figure 7. An example of a table in input texts.

When annotating reaction spans in tables, any reaction related texts must be covered: the name or identifier of the product, starting material given together with their amount, conditions, and yield of the product must be included. However, any analytical data or non-textual data must not be covered here as well. For example, in Figure 8, “Example 192” and “Example 193” are prepared by procedures analogous to those for “Example 136”. Thus, their synthesis processes are not described in the patent, and are only briefly summarized in a table. The two examples should be annotated as chemical reactions including their titles and reaction products.

```
The title compounds of Examples 192-193 were prepared by procedures analogous to those described above in Example 136.
<table>
T4
Example 192 & 2,7-Phenanthrenediol, 1,2,3,4,4a,9,10,10a-octahydro-2,4a-dipropyl-, [2R-(2α,4aα,10aα)]-, MS: 285 (M-17)+.
T5
Example 193 & 2,7-Phenanthrenediol, 1,2,3,4,4a,9,10,10a-octahydro-2,4a-dipropyl-, [2S-(2α,4aβ,10aβ)]-, MS: 285 (M-17)+.
</table>
```

Figure 8. Annotating chemical reactions in a table.

Part 2: Reaction Reference Annotation

ANNOTATION INTERFACE

To facilitate annotation of reaction references, "*reference mode*" is provided via "Data -> Reference -> Reference mode". In this mode, the full patent text with reaction span annotations is displayed on the left (read-only) with the index tags of annotated reactions displayed on the right (to be annotated). The annotation for reaction reference relationships is conducted in the document on the right by linking the indices (e.g., the indices **T1**, **T2**, etc.) of reactions.

LINK TO A CHEMICAL REACTION

A pair of reaction steps should be linked if one refers to the other for the details of the reaction. Here, the details of a reaction include its starting materials, product, reagents, catalysts, solvent, and the conditions of the reaction such as temperature and time. An example of analogous reactions is shown in Figure 9.

In the reaction span of **T2**, there is a statement showing that the preparation process of "1(S)-Benzyl-6-methoxy-1(R)-(3-oxo-butyl)-3,4-dihydro-1H-naphthelen-2-one" is analogous to "Preparation 2". Thus, the detailed procedure of "Preparation 3" is omitted in the parent and readers/information extraction systems need to refer to the description of "Preparation 2" to obtain the complete details. In this case, the reaction **T1** is the parent reaction of the child reaction **T2**. A link with the label **REF** from **T2** to **T1** should be added in the right window. Note that the link is *directed*: the link should start from the child (*referrer*) and point to the parent (*referent*).

T1

181 Preparation 2

1(R)-Benzyl-6-methoxy-1(S)-(3-oxo-butyl)-3,4-dihydro-1H-naphthelen-2-one

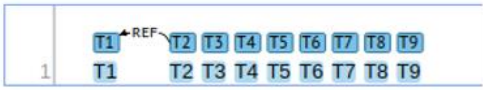
182 A solution of 62 g (0.23 mol) of the title product of Preparation 1 and 28 mL, (0.23 mol) of freshly distilled (S)-(-)-alpha-methyl benzylamine in 100 mL of toluene was heated to reflux, over a Dean-Stark trap, overnight. After removal of the azeotroped water, the imine solution was cooled to 0 °C and 21 mL (0.26 mol) of freshly distilled methylvinylketone was added dropwise to the solution. The solution was stirred at 0 °C for 30 min then heated to 40 °C overnight. The reaction solution was cooled to 0° C and 17 mL of acetic acid and 14 mL of H2O were added and the resultant solution was allow to warm to RT for 2 h. The solution was poured into H2O and extracted three times with EtOAc. The combined organic layers were washed with 1 N HCl, H2O, saturated NaHCO3, then dried over Na2SO4, filtered, and evaporated to dryness. The crude product was purified by chromatography over SiO2 using 15% EtOAc to 35% EtOAc in hexanes as the gradient eluant to give 48 g of the title product of this preparation as a yellow solid. 1HNMR (400 MHz, CDCL3) δ 1.38 (s, 3H), 1.40-1.51 (m, 2H), 1.64 (ddd, 1H, J = 2.1, 4.5, 13), 1.97 (broad s, 1H), 2.20 (dt, 1H, J = 4.5, 13), 2.59 (d, 1H, J = 6.6), 3.08 (d, 1H, J = 18), 3.16 (d, 1H, J = 16), 3.33 (dd, 1H, J = 6.6, 18), 3.62 (d, 1H, J = 16), 3.72 (s, 3H), 6.57 (d, 1H, J = 2.5), 6.67 (dd, 1H, J = 2.5, 8.8), 7.00-7.23 (m, 6H); 13C NMR (100 MHz, CDCl3) δ 27.90, 32.79, 34.40, 38.43, 41.49, 53.51, 55.12, 58.47, 79.06, 112.05, 113.09, 125.37, 127.63, 127.69, 130.27, 132.21, 135.45, 138.65, 157.88, 213.49; MS m/z 337 (M+H)+, 319 (M-OH)+.

T2

183 Preparation 3

1(S)-Benzyl-6-methoxy-1(R)-(3-oxo-butyl)-3,4-dihydro-1H-naphthelen-2-one

184 The title product of this preparation was prepared using a method analogous to Preparation 2, using (R)-(+)-alphamethyl benzylamine in the initial imine formation. Starting with 4.64 g 1-benzyl-6-methoxy-3,4-dihydro-1H-naphthalen-2-one produced 3.58 g of the title product of this preparation as a yellow solid. 1HNMR (400 MHz, CDCL3) δ 1.38 (s, 3H), 1.40-1.51 (m, 2H), 1.64 (ddd, 1H, J = 2.1, 4.5, 13), 1.97 (broad s, 1H), 2.20 (dt, 1H, J = 4.5, 13), 2.59 (d, 1H, J = 6.6), 3.08 (d, 1H, J = 18), 3.16 (d, 1H, J = 16), 3.33 (dd, 1H, J = 6.6, 18), 3.62 (d, 1H, J = 16), 3.72 (s, 3H), 6.57 (d, 1H, J = 2.5), 6.67 (dd, 1H, J = 2.5, 8.8), 7.00-7.23 (m, 6H); 13C NMR (100 MHz, CDCl3) δ 27.90, 32.79, 34.40, 38.43, 41.49, 53.51, 55.12, 58.47, 79.06, 112.05, 113.09, 125.37, 127.63, 127.69, 130.27, 132.21, 135.45, 138.65, 157.88, 213.49; MS m/z 337 (M+H)+, 319 (M-OH)+.



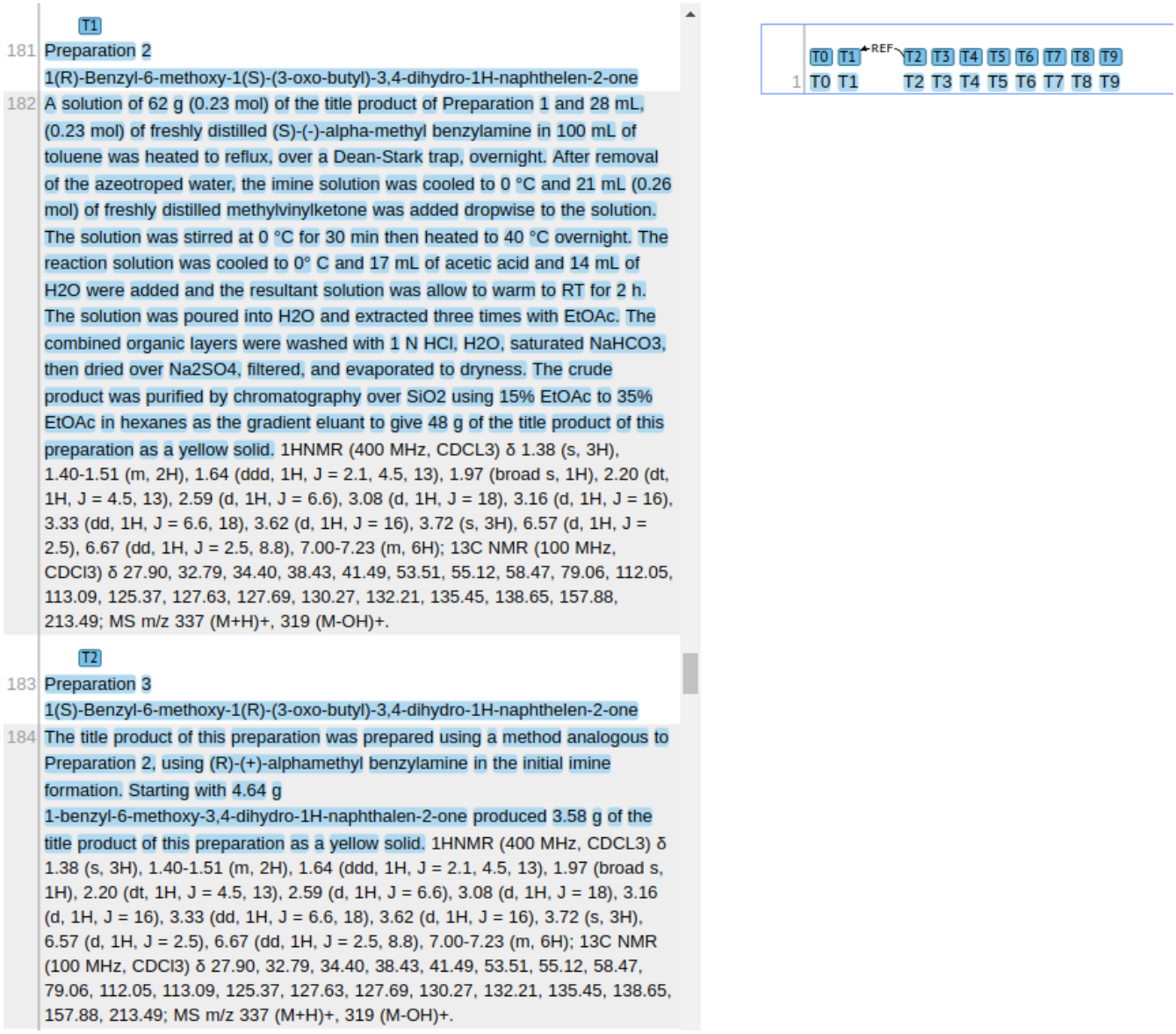


Figure 9. Annotating a pair of analogous reactions.

LINK TO A REACTION SEQUENCE

Sometimes a child reaction does not refer to a single reaction step but a sequence of several reaction steps. In this case, annotators should consider the reference relationship between the child reaction to each reaction step in the reaction sequence, and only link the child reaction to those steps that it actually refers to. Figure 10 gives an example where a child reaction refers to a whole reaction sequence.

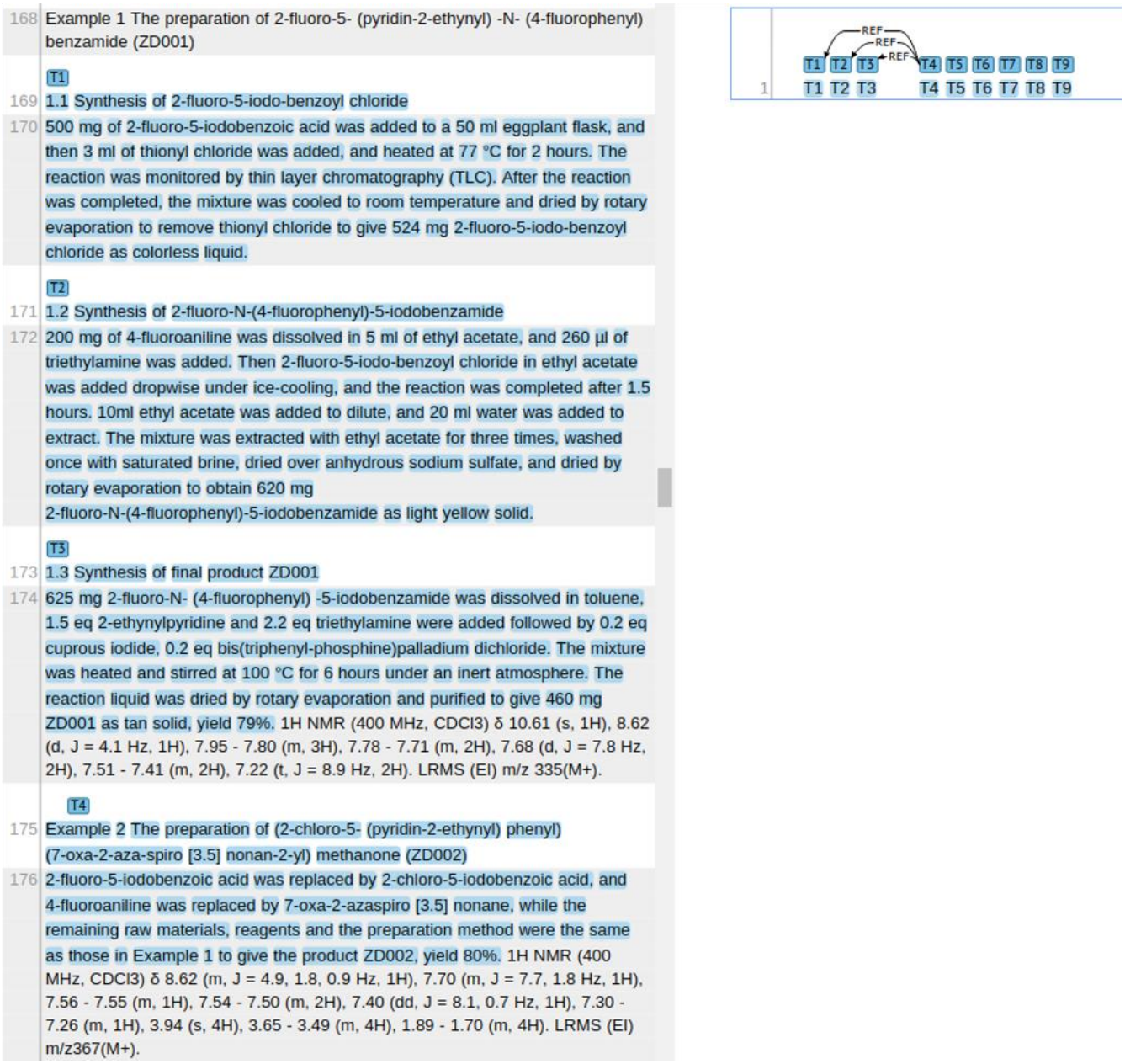


Figure 10. Linking a chemical reaction to a reaction sequence.

In the above figure, the reaction in Example 2 refers to Example 1 for the detail. Example 1 has a sequence of three reaction steps including the first two steps to produce the intermediate compounds.

It is clear from the description of Example 2 that it refers to all the three steps, as it mentions “2-fluoro-5-iodobenzoic acid” and “4-fluoroaniline”, which are used in 1.1 and 1.2 of Example 1, respectively. In this case, three links should be added starting from **T4** and ending at the three steps of Example 1 (**T1**, **T2**, and **T3**) with the label **REF**, respectively.

Sometimes, a reaction only refers to a subset of reaction steps in a reaction sequence. In such cases, the child reaction must be linked only to the reaction steps that it refers to. The figure below gives an example.

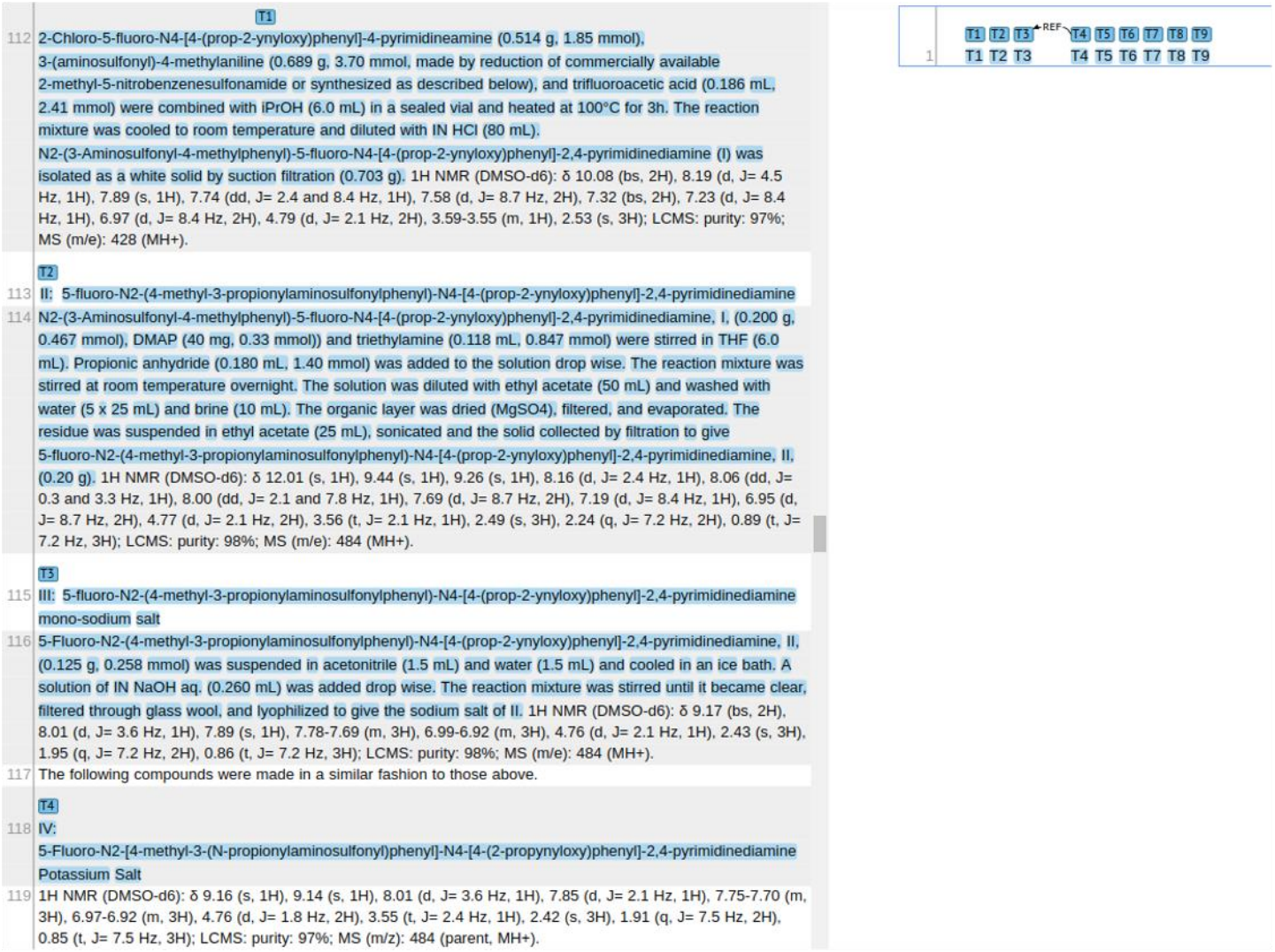


Figure 11. Linking a chemical reaction and one reaction step of a reaction sequence.

Example 1 presents the synthesis procedures of several different kinds of salts. Preparation of compound I and II are common for all the salts, and each salt in III--IV is made starting from compound II. Note that only parts of the synthesis procedure for compound I (**T1**) are included for ease of presentation.

In this case, even though the reactions to synthesize compound I, II and III constitutes a sequence, synthesis of salt IV (**T4**) is considered as analogous to only the final step to synthesize compound III (**T3**).

LINK TO A GENERAL CONDITION

The annotation rules for annotating references to general conditions are similar as those for analogous reactions. When a general condition contains only one chemical reaction, annotators should follow the rules for annotating LINK TO A CHEMICAL REACTION. In the cases where a general condition consists of a sequence of reaction steps, annotators should follow the rules for annotating LINK TO A REACTION SEQUENCE. The following figure depicts an example for annotating the reference relationship between a chemical reaction to a general condition.

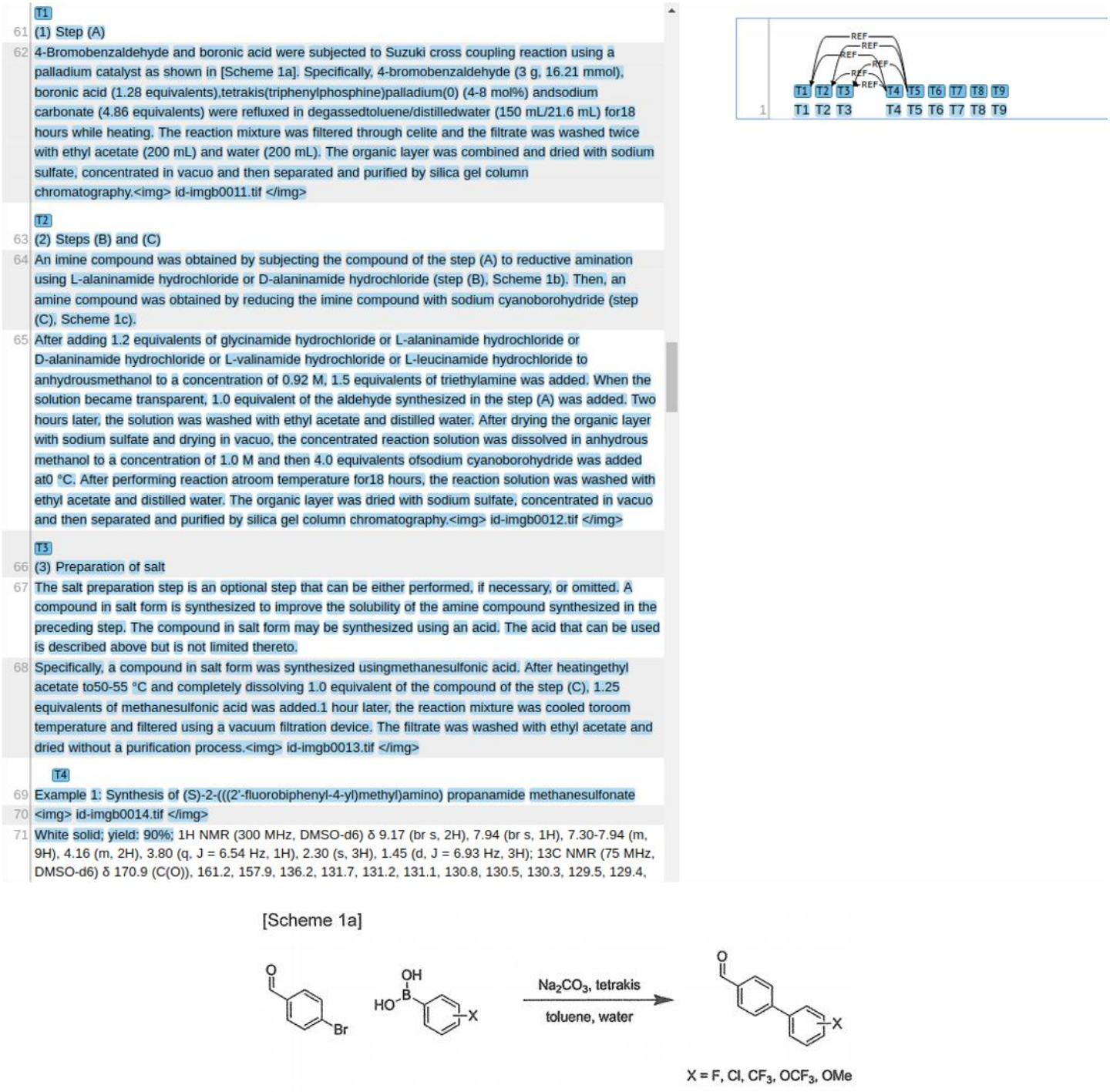


Figure 12. Annotating the reference relationship between a reaction to a general condition.

In the figure above, a general condition consisting of a number of reactions is described in the first part of the Examples section including the steps titled by “Step (1) Step (A)”, “(2) Steps (B) and (C)”, and “(3) Preparation of salt”. Following this general condition, numbered examples with only the product name and analytical information are presented such as “Example 1” and “Example 2”. Readers need to refer to the description of the general condition to obtain complete details of these actual examples.

In this case, all actual reactions (e.g., **T4** and **T5**) must be linked to the general condition: for each actual reaction, three links labeled with **REF** are created pointing to *each step* in the general description.

MULTIPLE LINKS

It is possible that a reaction description has multiple referents (multiple parent reactions and/or general conditions). In this case, links should be given between the child reaction and all parent reactions and/or general conditions.

REPEATED REACTION DESCRIPTION

Sometimes description of the same reaction is repeated multiple times in a document. If a parent reaction has repeated descriptions, the link should be given to the first occurrence that contains complete details of the reaction.

REFERENCE TO EXTERNAL RESOURCES

Sometimes a reaction description refers to external documents or resources for the complete details of the synthesis. These cases need to be ignored, and no links should be given from the description.

Part 3: Cue Annotation

In general, for each pair of parent-child reactions, there should be a *cue statement* that indicates the analogy between them. Annotators are asked to mark the cue statement and associate it with the analogous reaction pair.

The cue statements often contain a phrase such as “analogous to”, “in the same manner” or “in a similar fashion”. For example, in Figure 1, the cue statement is “The title product of this preparation was prepared using a method analogous to Preparation 2, …, formation”. This statement clearly indicates a reference relationship from Preparation 3 to Preparation 2 via the phrase “analogous to”.

In ideal cases, cue statements explicitly refer to the parent reactions by the example labels (e.g., “Step 4 of Example 1”). However, sometimes they do not contain any explicit references to example labels. For example, in Figure 11, the cue statement is: “The following compound were made in a similar fashion to those above”, which does not contain any explicit example labels.

When annotating cue statements, full sentences need to be annotated wherever applicable. A cue statement must be a sentence or a sequence of consecutive sentences. Continue with the example in Figure 1, the cue statement in Figure 1 needs to be annotated as **T3** in the figure below, i.e., the text span “The title product of this preparation, …, formation.”.

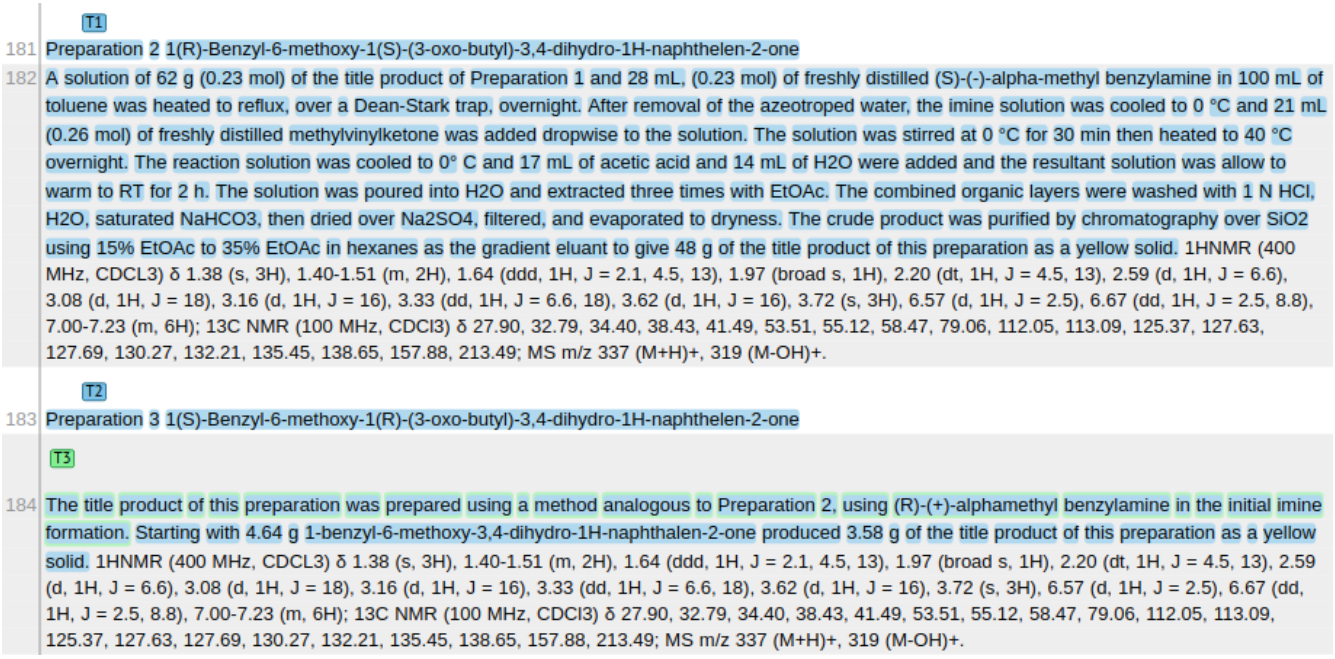


Figure 13. An example of cue annotation.

In some cases, there are no cue statements and the reference relationships amongst reactions can only be inferred by the document structure. In such cases, the text spans that reflect the document structure need to be annotated. Continue with the example in Figure 2. Since the reference relationship is indicated by the figure Schema 1a, Schema 1a (presented in a simplified textual form) needs to be annotated, i.e., the text span **T6** in Figure 14. In addition, the figures corresponding to the structures of Example 1 and Example 2 also reflect the reference relationships. As such, they should be annotated as cue statements as well. Note that the original figure is not included in the annotation interface. Annotators need to refer to the original PDF document of the patent to identify which “img” text span corresponds to Schema 1a.

/reaction_span/EP3202759A1 brat

T1
(1) Step (A)
4-Bromobenzaldehyde and boronic acid were subjected to Suzuki cross coupling reaction using a palladium catalyst as shown in [Scheme 1a]. Specifically, 4-bromobenzaldehyde (3 g, 16.21 mmol), boronic acid (1.28 equivalents),tetrakis(triphenylphosphine)palladium(0) (4-8 mol%) andsodium carbonate (4.86 equivalents) were refluxed in degassedtoluene/distilledwater (150 mL/21.6 mL) for18 hours while heating. The reaction mixture was filtered through celite and the filtrate was washed twice with ethyl acetate (200 mL) and water (200 mL). The organic layer was combined and dried with sodium sulfate, concentrated in vacuo and then separated and purified

T2

by silica gel column chromatography.<img> id-imgb0011.tif </img>

T3
(2) Steps (B) and (C)
An imine compound was obtained by subjecting the compound of the step (A) to reductive amination using L-alaninamide hydrochloride or D-alaninamide hydrochloride (step (B), Scheme 1b). Then, an amine compound was obtained by reducing the imine compound with sodium cyanoborohydride (step (C), Scheme 1c).
After adding 1.2 equivalents of glycinamide hydrochloride or L-alaninamide hydrochloride or D-alaninamide hydrochloride or L-valinamide hydrochloride or L-leucinamide hydrochloride to anhydrousmethanol to a concentration of 0.92 M, 1.5 equivalents of triethylamine was added. When the solution became transparent, 1.0 equivalent of the aldehyde synthesized in the step (A) was added. Two hours later, the solution was washed with ethyl acetate and distilled water. After drying the organic layer with sodium sulfate and drying in vacuo, the concentrated reaction solution was dissolved in anhydrous methanol to a concentration of 1.0 M and then 4.0 equivalents ofsodium cyanoborohydride was added at0 °C. After performing reaction atroom temperature for18 hours, the reaction solution was washed with ethyl acetate and distilled water. The

T4

organic layer was dried with sodium sulfate, concentrated in vacuo and then separated and purified by silica gel column chromatography.<img> id-imgb0012.tif </img>

T5
(3) Preparation of salt
The salt preparation step is an optional step that can be either performed, if necessary, or omitted. A compound in salt form is synthesized to improve the solubility of the amine compound synthesized in the preceding step. The compound in salt form may be synthesized using an acid. The acid that can be used is described above but is not limited thereto.
Specifically, a compound in salt form was synthesized usingmethanesulfonic acid. After heatingethyl acetate to50-55 °C and completely dissolving 1.0 equivalent of the compound of the step (C), 1.25 equivalents of methanesulfonic acid was added.1 hour later, the reaction mixture was cooled toroom temperature and filtered using a vacuum

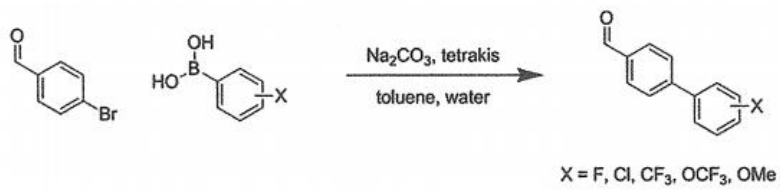
T6

filtration device. The filtrate was washed with ethyl acetate and dried without a purification process.<img> id-imgb0013.tif </img>

T7
Example 1: Synthesis of (S)-2-(((2'-fluorobiphenyl-4-yl)methyl)amino) propanamide methanesulfonate
<img> id-imgb0014.tif </img>
White solid; yield: 90%; 1H NMR (300 MHz, DMSO-d6) δ 9.17 (br s, 2H), 7.94 (br s, 1H), 7.30-7.94 (m, 9H), 4.16 (m, 2H), 3.80 (q, J = 6.54 Hz, 1H), 2.30 (s, 3H), 1.45 (d, J = 6.93 Hz, 3H); 13C NMR (75 MHz, DMSO-d6) δ 170.9 (C(O)), 161.2, 157.9, 136.2, 131.7, 131.2, 131.1, 130.8, 130.5, 130.3, 129.5, 129.4, 128.1, 127.9, 125.5, 125.4, 116.8, 116.5 (ArC), 55.1 (C(O)CH+NH2), 48.7 (+NH2CH2Ph), 16.4 (CH3). SCH3 signal overlapping with DMSO signal.

T8
Example 2: Synthesis of (S)-2-(((3'-fluorobiphenyl-4-yl)methyl)amino) propanamide methanesulfonate
<img> id-imgb0015.tif </img>
White solid; yield: 97%; 1H NMR (300 MHz, DMSO-d6) δ 9.15 (br s, 2H), 7.92 (br s, 1 H), 7.81 (d, J = 8.25 Hz, 2ArH), 7.68 (br s, 1 H), 7.49-7.60 (m, 5ArH), 7.20-7.27 (m,

[Scheme 1a]



X = F, Cl, CF_3, OCF_3, OMe

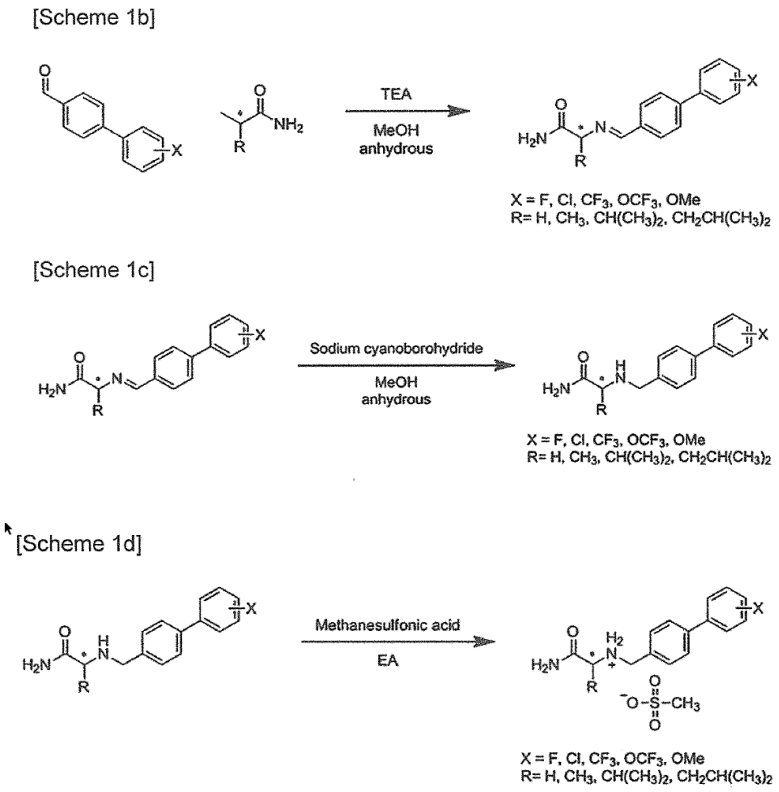


Figure 14. Annotating cues when the reference relationships are reflected by the document structure.

Part 4: Linking Cues to Reaction Spans

Each annotated cue reflects the reference relationships between at least one pair of reaction spans. Reference links need to be annotated to bridge reaction pairs with their associated cue statements. Given a pair of reaction spans and their associated cue statement, two links need to be added: one from the child reaction to the cue statement, and another from the cue statement to the parent reaction. Similar with reaction reference annotation, linking cue statements to reactions is done in the "Reference mode" of BRAT.

For example, in Figure 15, **T3** is the cue statement that implies the reference relationship from **T2** to **T1**. Thus, two links are added in the right window: (1) from **T2** to **T3** with the label **CHILD_CUE**; and (2) from **T3** to **T1** with the label **CUE_PARENT**.

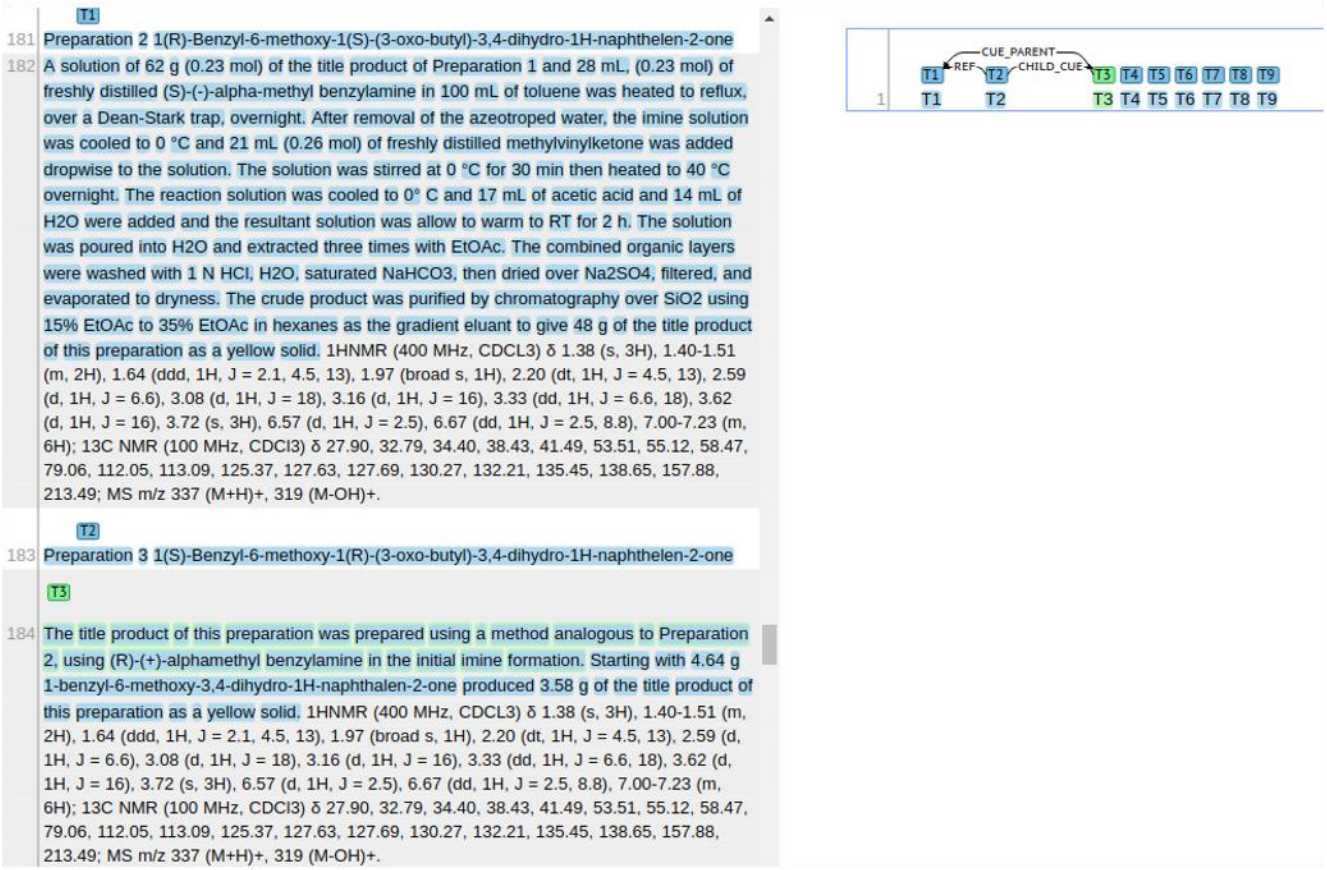


Figure 15. Linking a cue statement to its associated reaction spans.