
High-Fidelity Counterfactual Explanations through Conformal Prediction

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 We propose Conformal Counterfactual Explanations: an effortless and rigorous way
2 to produce realistic and faithful Counterfactual Explanations for Black-Box models
3 using Conformal Prediction. To address the need for realistic counterfactuals,
4 existing work has primarily relied on surrogate models to learn the data-generating
5 process. While this is an effective way to produce plausible counterfactual ex-
6 planations, it effectively reallocates the task of learning realistic representations
7 of the data from the model itself to the surrogate. Consequently, the generated
8 explanations may look compelling to human inspectors but not necessarily accu-
9 rately reflect the behaviour of the Black-Box model. Recent work has shown that
10 instead of relying on surrogate models to generate plausible explanations, we can
11 leverage predictive uncertainty for this task. Unfortunately, this approach works
12 under the assumption that models can produce predictive uncertainty estimates. By
13 leveraging a novel approach to differentiable Conformal Prediction we show that it
14 is possible to relax that assumption.

15 1 Submission of papers to NeurIPS 2022

16 Please read the instructions below carefully and follow them faithfully.

17 1.1 Style

18 Papers to be submitted to NeurIPS 2022 must be prepared according to the instructions presented
19 here. Papers may only be up to **nine** pages long, including figures. Additional pages *containing only*
20 *acknowledgments and references* are allowed. Papers that exceed the page limit will not be reviewed,
21 or in any other way considered for presentation at the conference.

22 The margins in 2022 are the same as those in 2007, which allow for $\sim 15\%$ more words in the paper
23 compared to earlier years.

24 Authors are required to use the NeurIPS L^AT_EX style files obtainable at the NeurIPS website as
25 indicated below. Please make sure you use the current files and not previous versions. Tweaking the
26 style files may be grounds for rejection.

27 1.2 Retrieval of style files

28 The style files for NeurIPS and other conference information are available on the World Wide Web at

29 <http://www.neurips.cc/>

30 The file `neurips_2022.pdf` contains these instructions and illustrates the various formatting re-
31 quirements your NeurIPS paper must satisfy.

32 The only supported style file for NeurIPS 2022 is `neurips_2022.sty`, rewritten for \LaTeX 2 ϵ .
33 **Previous style files for \LaTeX 2.09, Microsoft Word, and RTF are no longer supported!**

34 The \LaTeX style file contains three optional arguments: `final`, which creates a camera-ready copy,
35 `preprint`, which creates a preprint for submission to, e.g., arXiv, and `nonatbib`, which will not
36 load the `natbib` package for you in case of package clash.

37 **Preprint option** If you wish to post a preprint of your work online, e.g., on arXiv, using the
38 NeurIPS style, please use the `preprint` option. This will create a nonanonymized version of your
39 work with the text “Preprint. Work in progress.” in the footer. This version may be distributed as
40 you see fit. Please **do not** use the `final` option, which should **only** be used for papers accepted to
41 NeurIPS.

42 At submission time, please omit the `final` and `preprint` options. This will anonymize your
43 submission and add line numbers to aid review. Please do *not* refer to these line numbers in your
44 paper as they will be removed during generation of camera-ready copies.

45 The file `neurips_2022.tex` may be used as a “shell” for writing your paper. All you have to do is
46 replace the author, title, abstract, and text of the paper with your own.

47 The formatting instructions contained in these style files are summarized in Sections 2, 3, and 4
48 below.

49 **2 General formatting instructions**

50 The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long.
51 The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points.
52 Times New Roman is the preferred typeface throughout, and will be selected for you by default.
53 Paragraphs are separated by $\frac{1}{2}$ line space (5.5 points), with no indentation.

54 The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal
55 rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow $\frac{1}{4}$ inch
56 space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the
57 page.

58 For the final version, authors’ names are set in boldface, and each name is centered above the
59 corresponding address. The lead author’s name is to be listed first (left-most), and the co-authors’
60 names (if different address) are set to follow. If there is only one co-author, list both author and
61 co-author side by side.

62 Please pay special attention to the instructions in Section 4 regarding figures, tables, acknowledgments,
63 and references.

64 **3 Headings: first level**

65 All headings should be lower case (except for first word and proper nouns), flush left, and bold.

66 First-level headings should be in 12-point type.

67 **3.1 Headings: second level**

68 Second-level headings should be in 10-point type.

69 **3.1.1 Headings: third level**

70 Third-level headings should be in 10-point type.

71 **Paragraphs** There is also a `\paragraph` command available, which sets the heading in bold, flush
72 left, and inline with the text, with the heading followed by 1 em of space.

73 4 Citations, figures, tables, references

74 These instructions apply to everyone.

75 4.1 Citations within the text

76 The natbib package will be loaded for you by default. Citations may be author/year or numeric, as
77 long as you maintain internal consistency. As to the format of the references themselves, any style is
78 acceptable as long as it is used consistently.

79 The documentation for natbib may be found at

80 `http://mirrors.ctan.org/macros/latex/contrib/natbib/natnotes.pdf`

81 Of note is the command `\citet`, which produces citations appropriate for use in inline text. For
82 example,

83 `\citet{hasselmo}` investigated\dotso

84 produces

85 Hasselmo, et al. (1995) investigated...

86 If you wish to load the natbib package with options, you may add the following before loading the
87 neurips_2022 package:

88 `\PassOptionsToPackage{options}{natbib}`

89 If natbib clashes with another package you load, you can add the optional argument nonatbib
90 when loading the style file:

91 `\usepackage[nonatbib]{neurips_2022}`

92 As submission is double blind, refer to your own published work in the third person. That is, use “In
93 the previous work of Jones et al. [4],” not “In our previous work [4].” If you cite your other papers
94 that are not widely available (e.g., a journal paper under review), use anonymous author names in the
95 citation, e.g., an author of the form “A. Anonymous.”

96 4.2 Footnotes

97 Footnotes should be used sparingly. If you do require a footnote, indicate footnotes with a number¹
98 in the text. Place the footnotes at the bottom of the page on which they appear. Precede the footnote
99 with a horizontal rule of 2 inches (12 picas).

100 Note that footnotes are properly typeset *after* punctuation marks.²

101 4.3 Figures

102 All artwork must be neat, clean, and legible. Lines should be dark enough for purposes of reproduction.
103 The figure number and caption always appear after the figure. Place one line space before the figure
104 caption and one line space after the figure. The figure caption should be lower case (except for first
105 word and proper nouns); figures are numbered consecutively.

106 You may use color figures. However, it is best for the figure captions and the paper body to be legible
107 if the paper is printed in either black/white or in color.

108 4.4 Tables

109 All tables must be centered, neat, clean and legible. The table number and title always appear before
110 the table. See Table 1.

¹Sample of the first footnote.

²As in this example.

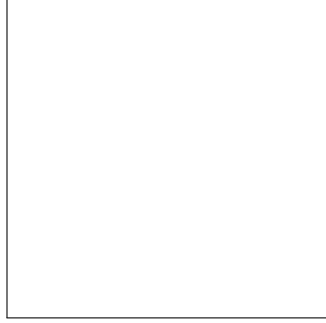


Figure 1: Sample figure caption.

Table 1: Sample table title

Part		
Name	Description	Size (μm)
Dendrite	Input terminal	~ 100
Axon	Output terminal	~ 10
Soma	Cell body	up to 10^6

111 Place one line space before the table title, one line space after the table title, and one line space after
 112 the table. The table title must be lower case (except for first word and proper nouns); tables are
 113 numbered consecutively.

114 Note that publication-quality tables *do not contain vertical rules*. We strongly suggest the use of the
 115 booktabs package, which allows for typesetting high-quality, professional tables:

116 <https://www.ctan.org/pkg/booktabs>

117 This package was used to typeset Table 1.

118 5 Final instructions

119 Do not change any aspects of the formatting parameters in the style files. In particular, do not modify
 120 the width or length of the rectangle the text should fit into, and do not change font sizes (except
 121 perhaps in the **References** section; see below). Please note that pages should be numbered.

122 6 Preparing PDF files

123 Please prepare submission files with paper size “US Letter,” and not, for example, “A4.”

124 Fonts were the main cause of problems in the past years. Your PDF file must only contain Type 1 or
 125 Embedded TrueType fonts. Here are a few instructions to achieve this.

- 126 • You should directly generate PDF files using `pdflatex`.
- 127 • You can check which fonts a PDF files uses. In Acrobat Reader, select the menu
 128 Files>Document Properties>Fonts and select Show All Fonts. You can also use the program
 129 `pdf fonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- 130 • The IEEE has recommendations for generating PDF files whose fonts are also ac-
 131 ceptable for NeurIPS. Please see [http://www.emfield.org/icuwb2010/downloads/](http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf)
 132 `IEEE-PDF-SpecV32.pdf`
- 133 • `xfig` “patterned” shapes are implemented with bitmap fonts. Use “solid” shapes instead.
- 134 • The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS
 135 Fonts:

136 `\usepackage{amsfonts}`
 137 followed by, e.g., `\mathbb{R}`, `\mathbb{N}`, or `\mathbb{C}` for \mathbb{R} , \mathbb{N} or \mathbb{C} . You can also
 138 use the following workaround for reals, natural and complex:

139 `\newcommand{\RR}{\mathbb{R}}` %real numbers
 140 `\newcommand{\Nat}{\mathbb{N}}` %natural numbers
 141 `\newcommand{\CC}{\mathbb{C}}` %complex numbers

142 Note that `amsfonts` is automatically loaded by the `amssymb` package.

143 If your file contains type 3 fonts or non embedded TrueType fonts, we will ask you to fix it.

144 6.1 Margins in L^AT_EX

145 Most of the margin problems come from figures positioned by hand using `\special` or other
 146 commands. We suggest using the command `\includegraphics` from the `graphicx` package.
 147 Always specify the figure width as a multiple of the line width as in the example below:

148 `\usepackage[pdftex]{graphicx} ...`
 149 `\includegraphics[width=0.8\linewidth]{myfile.pdf}`

150 See Section 4.4 in the graphics bundle documentation ([http://mirrors.ctan.org/macros/](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf)
 151 [latex/required/graphics/grfguide.pdf](http://mirrors.ctan.org/macros/latex/required/graphics/grfguide.pdf))

152 A number of width problems arise when L^AT_EX cannot properly hyphenate a line. Please give LaTeX
 153 hyphenation hints using the `\-` command when necessary.

154 References

155 References follow the acknowledgments. Use unnumbered first-level heading for the references. Any
 156 choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font
 157 size to `small` (9 point) when listing the references. Note that the Reference section does not count
 158 towards the page limit.

159 References

160 Checklist

161 The checklist follows the references. Please read the checklist guidelines carefully for information on
 162 how to answer these questions. For each question, change the default **[TODO]** to **[Yes]**, **[No]**, or
 163 **[N/A]**. You are strongly encouraged to include a **justification to your answer**, either by referencing
 164 the appropriate section of your paper or providing a brief inline description. For example:

- 165 • Did you include the license to the code and datasets? **[Yes]** See Section 2.
- 166 • Did you include the license to the code and datasets? **[No]** The code and the data are
 167 proprietary.
- 168 • Did you include the license to the code and datasets? **[N/A]**

169 Please do not modify the questions and only use the provided macros for your answers. Note that the
 170 Checklist section does not count towards the page limit. In your paper, please delete this instructions
 171 block and only keep the Checklist section heading above along with the questions/answers below.

172 1. For all authors...

- 173 (a) Do the main claims made in the abstract and introduction accurately reflect the paper's
 174 contributions and scope? **[TODO]**
- 175 (b) Did you describe the limitations of your work? **[TODO]**
- 176 (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**

- 177 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
178 them? **[TODO]**
- 179 2. If you are including theoretical results...
- 180 (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
- 181 (b) Did you include complete proofs of all theoretical results? **[TODO]**
- 182 3. If you ran experiments...
- 183 (a) Did you include the code, data, and instructions needed to reproduce the main experi-
184 mental results (either in the supplemental material or as a URL)? **[TODO]**
- 185 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
186 were chosen)? **[TODO]**
- 187 (c) Did you report error bars (e.g., with respect to the random seed after running experi-
188 ments multiple times)? **[TODO]**
- 189 (d) Did you include the total amount of compute and the type of resources used (e.g., type
190 of GPUs, internal cluster, or cloud provider)? **[TODO]**
- 191 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 192 (a) If your work uses existing assets, did you cite the creators? **[TODO]**
- 193 (b) Did you mention the license of the assets? **[TODO]**
- 194 (c) Did you include any new assets either in the supplemental material or as a URL?
195 **[TODO]**
- 196 (d) Did you discuss whether and how consent was obtained from people whose data you're
197 using/curating? **[TODO]**
- 198 (e) Did you discuss whether the data you are using/curating contains personally identifiable
199 information or offensive content? **[TODO]**
- 200 5. If you used crowdsourcing or conducted research with human subjects...
- 201 (a) Did you include the full text of instructions given to participants and screenshots, if
202 applicable? **[TODO]**
- 203 (b) Did you describe any potential participant risks, with links to Institutional Review
204 Board (IRB) approvals, if applicable? **[TODO]**
- 205 (c) Did you include the estimated hourly wage paid to participants and the total amount
206 spent on participant compensation? **[TODO]**

207 **A Appendix**

208 Optionally include extra information (complete proofs, additional experiments and plots) in the
209 appendix. This section will often be part of the supplemental material.