Conformal Counterfactual Explanations

Patrick Altmeyer*
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology
2628 XE Delft, The Netherlands
p.altmeyer@tudelft.nl

Abstract

We propose Conformal Counterfactual Explanations: an effortless and rigorous way to produce plausible and conformal Counterfactual Explanations for Black Box Models using Conformal Prediction. To address the need for plausible explanations, existing work has primarily relied on surrogate models to learn the data-generating process. This effectively reallocates the task of learning realistic representations of the data from the model itself to the surrogate. Consequently, the generated explanations may look plausible to humans but not necessarily conform with the behaviour of the Black Box Model. We formalise this notion through the introduction of new evaluation measures. In order to still address the need for plausibility, we build on a recent approach that works by minimizing predictive model uncertainty. Using differentiable Conformal Prediction, we relax the previous assumption that the Black Box Model can produce predictive uncertainty estimates.

1 Introduction

Counterfactual Explanations are a powerful, flexible and intuitive way to not only explain Black Box Models but also enable affected individuals to challenge them through the means of Algorithmic Recourse. Instead of opening the black box, Counterfactual Explanations work under the premise of strategically perturbing model inputs to understand model behaviour [20]. Intuitively speaking, we generate explanations in this context by asking simple what-if questions of the following nature: 'Our credit risk model currently predicts that this individual's credit profile is too risky to offer them a loan. What if they reduced their monthly expenditures by 10%? Will our model then predict that the individual is credit-worthy'?

This is typically implemented by defining a target outcome $t \in \mathcal{Y}$ for some individual $x \in \mathcal{X}$, for which the model $M_\theta : \mathcal{X} \mapsto \mathcal{Y}$ initially predicts a different outcome: $M_\theta(x) \neq t$. Counterfactuals are then searched by minimizing a loss function that compares the predicted model output to the target outcome: $\text{yloss}(M_\theta(x), t)$. Since Counterfactual Explanations (CE) work directly with the Black Box Model, they always have full local fidelity by construction. Fidelity is defined as the degree to which explanations approximate the predictions of the Black Box Model. This arguably one of the most important evaluation metrics for model explanations, since any explanation that explains a prediction not actually made by the model is useless [10].

In situations where full fidelity is a requirement, CE therefore offers a more appropriate solution to Explainable Artificial Intelligence (XAI) than other popular approaches like LIME [15] and SHAP [8], which involve local surrogate models. But even full fidelity is not a sufficient condition for ensuring that an explanation adequately describes the behaviour of a model. That is because two

*Use footnote for providing further information about author (webpage, alternative address)—*not* for acknowledging funding agencies.

Preprint. Under review.

very distinct explanations can both lead to the same model prediction, especially when dealing with heavily parameterized models:

> [...] deep neural networks are typically very underspecified by the available data, and [...] parameters [therefore] correspond to a diverse variety of compelling explanations for the data. — Wilson [21]

When people talk about Black Box Models, this is usually the type of model they have in mind.

In the context of CE, the idea that no two explanations are the same arises almost naturally. Even the baseline approach proposed by Wachter et al. [20] can yield a diverse set of explanations if counterfactuals are initialised randomly. This multiplicity of explanations has not only been acknowledged in the literature but positively embraced: since individuals seeking Algorithmic Recourse (AR) have unique preferences, Mothilal et al. [11], for example, have prescribed *diversity* as an explicit goal for counterfactuals. More generally, the literature on CE and AR has brought forward a myriad of desiderata for explanations, which we will discuss in more detail in the following section.

2 From Adversarial Examples to Plausible Explanations

Most state-of-the-art approaches to generating Counterfactual Explanations rely on gradient descent to optimize different flavours of the same counterfactual search objective,

$$\mathbf{s}^\prime = \arg \min_{\mathbf{s}^\prime \in \mathcal{S}} \{ \text{yloss}(M_\theta(f(\mathbf{s}^\prime)), y^*) + \lambda \text{cost}(f(\mathbf{s}^\prime)) \} \tag{1}$$

where yloss denotes the primary loss function already introduced above and cost is either a single penalty or a collection of penalties that are used to impose constraints through regularization. Following the convention in Altmeyer et al. [1] we use $\mathbf{s}^\prime = \{s_k\}_K$ to denote the vector K-dimensional array of counterfactual states. This is to explicitly account for the fact that we can generate multiple counterfactuals, as with DiCE [11], and may choose to traverse a latent representation $\mathcal{Z}$ of the feature space $\mathcal{X}$, as we will discuss further below.

Solutions to Equation 1 are considered valid as soon as the predicted label matches the target label. A stripped-down counterfactual explanation is therefore little different from an adversarial example. In Figure 1, for example, we have the baseline approach proposed in Wachter et al. [20] to MNIST data (centre panel). This approach solves Equation 1 through gradient-descent in the feature space with a penalty for the distance between the factual x and the counterfactual $x^\prime$. The underlying classifier M_θ is a simple Multi-Layer Perceptron (MLP) with good test accuracy. For the generated counterfactual $x^\prime$ the model predicts the target label with high confidence (centre panel in Figure 1). The explanation is valid by definition, even though it looks a lot like an Adversarial Example [3]. Schut et al. [16] make the connection between Adversarial Examples and Counterfactual Explanations explicit and propose using a Jacobian-Based Saliency Map Attack to solve Equation 1. They demonstrate that this approach yields realistic and sparse counterfactuals for Bayesian, adversarially robust classifiers. Applying their approach to our simple MNIST classifier does not yield a realistic counterfactual but this one, too, is valid (right panel in Figure 1).

The crucial difference between Adversarial Examples (AE) and Counterfactual Explanations is one of intent. While an AE is intended to go unnoticed, a CE should have certain desirable properties. The literature has made this explicit by introducing various so-called *desiderata*. To properly serve both AI practitioners and individuals affected by AI decision-making systems, counterfactuals should be sparse, proximate [20], actionable [18], diverse [11], plausible [5, 14, 16], robust [17, 13, 1] and causal [7] among other things. Researchers have come up with various ways to meet these desiderata, which have been surveyed in [19] and [6].

Finding ways to generate *plausible* counterfactuals has been one of the primary concerns. To this end, Joshi et al. [5] were among the first to suggest that instead of searching counterfactuals in the feature space $\mathcal{X}$, we can instead traverse a latent embedding $\mathcal{Z}$ that implicitly codifies the data generating process (DGP) of $x \sim \mathcal{X}$. To learn the latent embedding, they introduce a surrogate model. In particular, they propose to use the latent embedding of a Variational Autoencoder (VAE) trained to generate samples $x^* \leftarrow \mathcal{G}(z)$ where $\mathcal{G}$ denotes the decoder part of the VAE. Provided the surrogate

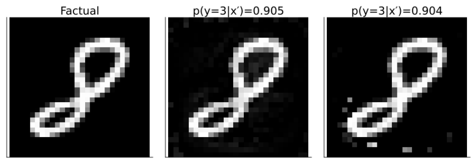

Figure 1: You may not like it, but this is what stripped-down counterfactuals look like. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using Wachter et al. [20] (centre); and a counterfactual produced using JSMA-based approach introduced by [16].

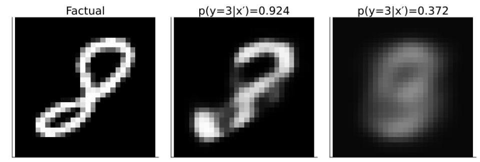

Figure 2: Using surrogates can improve plausibility, but also increases vulnerability. Counterfactuals for turning an 8 (eight) into a 3 (three): original image (left); counterfactual produced using REVISE [5] with a well-specified surrogate (centre); and a counterfactual produced using REVISE [5] with a poorly specified surrogate (right).

model is well-trained, their proposed approach —REVISE— can yield compelling counterfactual explanations like the one in the centre panel of Figure 2.

Others have proposed similar approaches. Dombrowski et al. [2] traverse the base space of a normalizing flow to solve Equation 1, essentially relying on a different surrogate model for the generative task. Poyiadzi et al. [14] use density estimators ($\hat{p} : \mathcal{X} \mapsto [0, 1]$) to constrain the counterfactual paths. Karimi et al. [7] argue that counterfactuals should comply with the causal model that generates the data. All of these different approaches share a common goal: ensuring that the generated counterfactuals comply with the true and unobserved DGP. To summarize this broad objective, we propose the following definition:

Definition 2.1 (Plausible Counterfactuals)**.** *Let $\mathcal{X}|y = t$ denote the true conditional distribution of samples in the target class t. Then for x' to be considered a plausible counterfactual, we need: $x' \sim \mathcal{X}|y = t$.*

Note that Definition 2.1 is consistent with the notion of plausible counterfactual paths, since we can simply apply it to each counterfactual state along the path.

Surrogate models offer an obvious solution to achieve this objective. Unfortunately, surrogates also introduce a dependency: the generated explanations no longer depend exclusively on the Black Box Model itself, but also on the surrogate model. This is not necessarily problematic if the primary objective is not to explain the behaviour of the model but to offer recourse to individuals affected by it. It may become problematic even in this context if the dependency turns into a vulnerability. To illustrate this point, we have used REVISE [5] with an underfitted VAE to generate the counterfactual in the right panel of Figure 2: in this case, the decoder step of the VAE fails to yield plausible values ($\{x' \leftarrow \mathcal{G}(z)\} \not\sim \mathcal{X}|y = t$) and hence the counterfactual search in the learned latent space is doomed.

3 A Framework for Conformal Counterfactual Explanations

In Section 2 we explained that Counterfactual Explanations work directly with Black Box Model, so fidelity is not a concern. This may explain why research has primarily focused on other desiderata, most notably plausibility (Definition 2.1). Enquiring about the plausibility of a counterfactual essentially boils down to the following question: 'Is this counterfactual consistent with the underlying data'? To introduce this section, we posit a related, slightly more nuanced question: 'Is this counterfactual consistent with what the model has learned about the underlying data'? We will argue that fidelity is not a sufficient evaluation measure to answer this question and propose a novel way to assess if explanations conform with model behaviour. Finally, we will introduce a framework for Conformal Counterfactual Explanations, that reconciles the notions of plausibility and model conformity.

3.1 From Fidelity to Model Conformity

The word *fidelity* stems from the Latin word 'fidelis', which means 'faithful, loyal, trustworthy' [9]. As we explained in Section 2, model explanations are considered faithful if their corresponding predictions coincide with the predictions made by the model itself. Since this definition of faithfulness is not useful in the context of Counterfactual Explanations, we propose an adapted version:

Definition 3.1 (Conformal Counterfactuals)**.** *Let $\mathcal{X}_\theta|t = p_\theta(x|y = t)$ denote the conditional distribution of x in the target class t, where $theta$ denotes the parameters of model M_θ. Then for x' to be considered a conformal counterfactual, we need: $x' \sim \mathcal{X}_\theta|t$.*

In words, conformal counterfactuals conform with what the predictive model has learned about the input data x. Since this definition works with distributional properties, it explicitly accounts for the multiplicity of explanations we discussed earlier. Except for the posterior conditional distribution $p_\theta(x|y = t)$, we already have access to all the ingredients in Definition 3.1.

How can we quantify $p_\theta(\mathbf{x}|y = t)$? After all, the predictive model M_θ was trained to discriminate outputs conditional on inputs, which is a different conditional distribution: $p_\theta(y|x)$. Learning the distribution over inputs $p_\theta(\mathbf{x}|y = t)$ is a generative task that M_θ was not explicitly trained for. In the context of Counterfactual Explanations, it is the task that existing approaches have reallocated from the model itself to a surrogate.

Fortunately, recent work by Grathwohl et al. [4] on Energy Based Models (EBM) has pointed out that there is a 'generative model hidden within every standard discriminative model'. The authors show that we can draw samples from the posterior conditional distribution $p_\theta(\mathbf{x}|y)$ using Stochastic Gradient Langevin Dynamics (SGLD). In doing so, it is possible to train classifiers jointly for the discriminative task using standard cross-entropy and the generative task using SGLD. They demonstrate empirically that among other things this improves predictive uncertainty quantification for discriminative models.

To see how their proposed conditional sampling strategy can be applied in our context, note that if we fix y to our target value t, we can sample from $p_\theta(\mathbf{x}|y = t)$ using SGLD as follows,

$$\mathbf{x}_{j+1} \leftarrow \mathbf{x}_j - \frac{\epsilon^2}{2}\mathcal{E}(\mathbf{x}_j|y = t) + \epsilon\mathbf{r}_j, \quad j = 1, ..., J \tag{2}$$

where $\mathbf{r}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is the stochastic term and the step-size ϵ is typically polynomially decayed. The term $\mathcal{E}(\mathbf{x}_j|y = t)$ denotes the energy function. Following Grathwohl et al. [4] we use $\mathcal{E}(\mathbf{x}_j|y = t) = -M_\theta(x)[t]$, that is the negative logit corresponding to the target class label t.

While $\mathbf{x}_K$ is only guaranteed to distribute as $p_\theta(\mathbf{x}|y = t)$ if $\epsilon \to 0$ and $J \to \infty$, the bias introduced for a small finite ϵ is negligible in practice [12, 4]. While Grathwohl et al. [4] use Equation 2 during training, we are interested in applying the conditional sampling procedure in a post hoc fashion to any standard discriminative model. Generating multiple samples in this manner yields an empirical distribution $\hat{\mathcal{X}}_\theta|t$, which we can use to assess if a given counterfactual x' conforms with the model M_θ (Definition 3.1).

TBD

- What exact sampler do we use? ImproperSGLD as in Grathwohl et al. [4] seems to work best.
- How exactly do we plan to quantify plausibility and conformity? Elaborate on measures.

3.2 Conformal Training meets Counterfactual Explanations

Now that we have a way of evaluating Counterfactual Explanations in terms of their plausibility and conformity, we are interested in finding a way to generate counterfactuals that are as plausible and conformal as possible. We hypothesize that a narrow focus on plausibility may come at the cost of reduced conformity. Using a surrogate model for the generative task, for example, may improve plausibility but inadvertently yield counterfactuals that are more consistent with the surrogate than the Black Box Model itself.

One way to ensure model conformity is to rely strictly on the model itself. Schut et al. [16] demonstrate that this restriction need not impede plausibility, since we can rely on predictive uncertainty estimates to guide our counterfactual search. By avoiding counterfactual paths that are associated with high predictive uncertainty, we end up generating counterfactuals for which the model M_θ predicts the target label t with high confidence. Provided the model is well-calibrated, these counterfactuals are plausible.

Interestingly, Schut et al. [16] point to this connection between the generative task and predictive uncertainty quantification

4 Experiments

References

[1] Patrick Altmeyer, Giovan Angela, Aleksander Buszydlik, Karol Dobiczek, Arie van Deursen, and Cynthia Liem. Endogenous Macrodynamics in Algorithmic Recourse. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.

[2] Ann-Kathrin Dombrowski, Jan E Gerken, and Pan Kessel. Diffeomorphic explanations with normalizing flows. In *ICML Workshop on Invertible Neural Networks, Normalizing Flows, and Explicit Likelihood Models*, 2021.

[3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014.

[4] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. March 2020. URL `https://openreview.net/forum?id=Hkxzx0NtDB`.

[5] Shalmali Joshi, Oluwasanmi Koyejo, Warut Vijitbenjaronk, Been Kim, and Joydeep Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. 2019.

[6] Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. 2020.

[7] Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

[8] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4768–4777, 2017.

[9] Merriam-Webster. "fidelity". URL `https://www.merriam-webster.com/dictionary/fidelity`.

[10] Christoph Molnar. *Interpretable Machine Learning*. Lulu. com, 2020.

[11] Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

[12] Kevin P. Murphy. *Probabilistic machine learning: Advanced topics*. MIT Press.

[13] Martin Pawelczyk, Teresa Datta, Johannes van-den Heuvel, Gjergji Kasneci, and Himabindu Lakkaraju. Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse. *arXiv preprint arXiv:2203.06768*, 2022.

[14] Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

[15] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016.

[16] Lisa Schut, Oscar Key, Rory Mc Grath, Luca Costabello, Bogdan Sacaleanu, Yarin Gal, et al. Generating Interpretable Counterfactual Explanations By Implicit Minimisation of Epistemic and Aleatoric Uncertainties. In *International Conference on Artificial Intelligence and Statistics*, pages 1756–1764. PMLR, 2021.

[17] Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards Robust and Reliable Algorithmic Recourse. 2021.

[18] Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 10–19, 2019.

[19] Sahil Verma, John Dickerson, and Keegan Hines. Counterfactual explanations for machine learning: A review. 2020.

[20] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

[21] Andrew Gordon Wilson. The case for Bayesian deep learning. 2020.

5 Submission of papers to NeurIPS 2022

Please read the instructions below carefully and follow them faithfully.

5.1 Style

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

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

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

5.2 Retrieval of style files

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

`http://www.neurips.cc/`

The file `neurips_2022.pdf` contains these instructions and illustrates the various formatting requirements your NeurIPS paper must satisfy.

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

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

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

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

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

The formatting instructions contained in these style files are summarized in Sections 6, 7, and 8 below.

6 General formatting instructions

The text must be confined within a rectangle 5.5 inches (33 picas) wide and 9 inches (54 picas) long. The left margin is 1.5 inch (9 picas). Use 10 point type with a vertical spacing (leading) of 11 points. Times New Roman is the preferred typeface throughout, and will be selected for you by default. Paragraphs are separated by ½ line space (5.5 points), with no indentation.

The paper title should be 17 point, initial caps/lower case, bold, centered between two horizontal rules. The top rule should be 4 points thick and the bottom rule should be 1 point thick. Allow ¼ inch space above and below the title to rules. All pages should start at 1 inch (6 picas) from the top of the page.

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

Please pay special attention to the instructions in Section 8 regarding figures, tables, acknowledgments, and references.

7 Headings: first level

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

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

7.1 Headings: second level

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

7.1.1 Headings: third level

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

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

8 Citations, figures, tables, references

These instructions apply to everyone.

8.1 Citations within the text

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

The documentation for `natbib` may be found at

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

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

```
\citet{hasselmo} investigated\dots
```

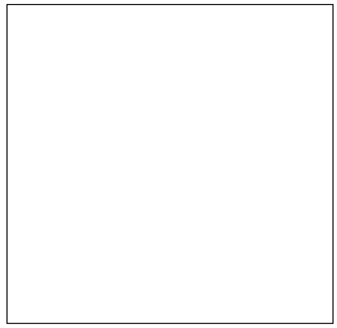

Figure 3: Sample figure caption.

produces

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

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

```
\PassOptionsToPackage{options}{natbib}
```

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

```
\usepackage[nonatbib]{neurips_2022}
```

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

8.2 Footnotes

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

Note that footnotes are properly typeset *after* punctuation marks.[3]

8.3 Figures

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

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

8.4 Tables

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

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

---

[2] Sample of the first footnote.
[3] As in this example.

Table 1: Sample table title

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

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

`https://www.ctan.org/pkg/booktabs`

This package was used to typeset Table 1.

9 Final instructions

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

10 Preparing PDF files

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

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

- You should directly generate PDF files using `pdflatex`.
- You can check which fonts a PDF files uses. In Acrobat Reader, select the menu Files>Document Properties>Fonts and select Show All Fonts. You can also use the program `pdffonts` which comes with `xpdf` and is available out-of-the-box on most Linux machines.
- The IEEE has recommendations for generating PDF files whose fonts are also acceptable for NeurIPS. Please see `http://www.emfield.org/icuwb2010/downloads/IEEE-PDF-SpecV32.pdf`
- `xfig` "patterned" shapes are implemented with bitmap fonts. Use "solid" shapes instead.
- The `\bbold` package almost always uses bitmap fonts. You should use the equivalent AMS Fonts:

  ```
  \usepackage{amsfonts}
  ```

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

  ```
  \newcommand{\RR}{I\!\!R} %real numbers
  \newcommand{\Nat}{I\!\!N} %natural numbers
  \newcommand{\CC}{I\!\!\!\!C} %complex numbers
  ```

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

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

10.1 Margins in LaTeX

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

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

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

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

Acknowledgments and Disclosure of Funding

Use unnumbered first level headings for the acknowledgments. All acknowledgments go at the end of the paper before the list of references. Moreover, you are required to declare funding (financial activities supporting the submitted work) and competing interests (related financial activities outside the submitted work). More information about this disclosure can be found at: `https://neurips.cc/Conferences/2022/PaperInformation/FundingDisclosure`.

Do **not** include this section in the anonymized submission, only in the final paper. You can use the `ack` environment provided in the style file to autmoatically hide this section in the anonymized submission.

References

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

Checklist

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

- Did you include the license to the code and datasets? [Yes] See Section 6.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

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

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? **[TODO]**
   (b) Did you describe the limitations of your work? **[TODO]**
   (c) Did you discuss any potential negative societal impacts of your work? **[TODO]**
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? **[TODO]**
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? **[TODO]**
   (b) Did you include complete proofs of all theoretical results? **[TODO]**
3. If you ran experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? **[TODO]**
(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? **[TODO]**
(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? **[TODO]**
(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? **[TODO]**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? **[TODO]**
   (b) Did you mention the license of the assets? **[TODO]**
   (c) Did you include any new assets either in the supplemental material or as a URL? **[TODO]**
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? **[TODO]**
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? **[TODO]**
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? **[TODO]**
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? **[TODO]**
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? **[TODO]**

A Appendix

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