The Capra Lab manuscript template

State a result with the title if possible. Aim for fewer than 10 words.

Your Name Here[1] and John A. Capra[2,3,*]

[1]Place, City, XX

[2]Bakar Computational Health Sciences Institute, University of California, San Francisco, CA

[3]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA

[*]*Correspondence to* labPI@email.org

Abstract

< 250 words. Shorter is better. Check journal limits, structured or unstructured, and so on.

Make a compelling elevator pitch for the paper that includes a brief intro, methods, results, conclusions. It can be split into sections or not. Always start with sections to provide a framework, but remove if needed. Revisit the abstract as you write the paper (even if you have written abstracts for conferences previously), because you will find better ways to summarize the paper as it evolves!

First sentence should make a broad and general statement that sets up to the importance of the topic. Avoid definitions in the first sentence. For example, don't say *TADs are 3D structures.* It is boring and does not convey importance or a gap. The first sentence can be the same as the first sentence of the Intro but often is more specific/less broad. Overall, it should establish the gap in knowledge as briefly as possible and avoid too much background. Put your main question/gap early (2nd or 3rd sentence). End the abstract by making the significance and implications for the field (and maybe next steps) clear. Be as broad as possible. No citations and minimize explicit references to other work in the abstract (meta-discourse like *Previous work that investigated X showed Y*); instead, just talk about Y and the gap that you will fill.

Background (No more than three sentences.) Motivate your work by stating the problem you addressed or the hypothesis you tested. The last sentence should clearly state the gap in knowledge.

Methods (One or two optional sentences; this can be combined with results if it is not a methods focused paper. *To address this gap, we applied X to Y.* or *we synthesized V and W to reveal Z.*

Results (Most of the rest of the abstract) *Here, we show...* What did you do and what did you find? This should summarize the main analyses and results of your work. Keep it high level; it is ok if some results are not mentioned.

Conclusions (No more than two sentences.) What did you learn from your studies? How does this change the field? What do your results say about the problem/hypothesis identified in the Background section?

1   Introduction

First paragraph(s) What is the problem? Why do we care? First sentence can be a general truth or general limitation that your audience accepts. E.g., *While differences between archaic and modern humans are well described, there is a poor understanding of the genetic mechanisms underlying such phenotypic differences.* or *Most aspects of archaic hominin biology cannot be directly studied due to their lack of preservation in fossils.* Don't waste the first sentence with a definition.

   Middle paragraph(s) What have others found about this specific problem? What are the current hypotheses and support for each? and/or what data have been previously analyzed? What are the gaps? E.g., *One such mechanism may be alternative splicing, but we cannot obtain such data from extinct taxa.* This part of the Intro may require a few paragraphs, but you do not need to fully review the history of work on the question. Focus on the current state of the field.

   Last paragraph Overview of the major questions addressed here and why they are important. *Here, we investigated X using dataset Y and dataset Z. We tested the A hypothesis. We also considered this other obscure hypothesis. We predicted x, y, and z.* E.g., *Here, we leverage a new deep neural network that can predict splice altering variants from sequence alone to examine such variants in archaic hominins.*

2 Results

Optional paragraph 0 Concise overview of what the reader needs to know about the methods or context (but don't take too long to get to the real results). *The goal of this work is to demonstrate "X".* If you use a lot of complicated or easily confused terminology, you can define it here to set the stage.

2.1 Declarative result statement

Context Set up the context and logic for the experiment/analysis. Clearly state the question and hypothesis. Sometimes this will already have been set up by the previous section and can be skipped.

Approach What was the experiment/analysis, briefly (cite methods section if needed)? Describe key factors of the approach, especially details that apply specifically to this result but not to others using the same method (i.e. specific simulation parameters that vary across results sections).

Result Clearly state the MAIN result of the experiment/analysis and any relevant statistics. Give a quantification of the effect size, and if you performed a significance test, give the quantification of significance (e.g., p-value or q-value) and test used in parentheses. E.g., *The mean expected 3D divergence is 78% higher than the observed 3D divergence ($P = 1.8 \times 10^{-48}$, t-test).* Reference relevant figure panels.

Details Elaborate on any results that modulate or demonstrate the robustness of your main result. These usually reference supplemental figures that support your main result.

Conclusion Conservative conclusions drawn from the result. What do these data say about your hypothesis? *This suggests. . .*

Example *Given X, we hypothesized that Tony is great. To evaluate/quantify Tony's greatness, we did X. We found Y (big picture overview of result) (Main text figure reference). We controlled for Z. Tony's greatness was robust even when considering W (Supplemental figure(s) reference). This suggests that Tony's greatness knows no bounds.*

2.2 Example results section to demonstrate figure references and placement

You should insert main text figures and tables into the document as soon as possible after the first mention in the text. We suggest you do this for initial submission even if the journal

3

requests that you put your figures and captions at the end in a separate section. Your reviewers will thank you! Here is an example of how to reference a full figure in-text (**Figure 1**) or just a sub-panel (**Figure 1A**). You might also want to reference two figures at the same time like this (**Figures 1**, **S1**). Sometimes you might reference a range of figures like these pair of *Capra* shown in **Figures S1**–**S2**. You can also reference a range of subpanels like **Figures 1A**–**1B**. You can do the same thing for tables (**Table S1**).

A

- **Title**. The figure title should tell a succinct result and lead the reader to the figure's conclusion (when possible).
- **Visually clear**: Remove all "*chartjunk*": unnecessary or confusing visual elements found in a figure that do not improve the message (in the best case) or add confusion (in the worst case). For example, chartjunk may include the use of too many colors, too many labels, gratuitously colored backgrounds, useless grid lines, etc.
- **Text style**. Figure and caption should be the same font as the main text. The font should be Helvetica or Arial.
- **Axis labels**. Always include axis labels.
- **Abbreviations**. All abbreviations defined in the caption.
- **Symbols/colors**. All symbols or colors defined - consistent throughout paper when possible
- **Dataset**: Specify data set used if not clear.
- **Sample size** (N = X). Some places require n to be lowercase for sample (rather than population) sizes
- **Raw data points**. Show whenever possible (especially if n < 10)
  - For scatter plots, point to one dot and have a small description (e.g., "one neuron")
- **Confidence intervals or SE bars**. Clearly labeled, for example: "Data are presented as mean values +/- SEM"
  - If using box plots, tell how they were made ("Boxes represent the IQR and outliers are more extreme than 1.5xIQR.")
- **Statistical significance** (P = X or q = X). Include details about the test used, one-sided vs two-sided (maybe degrees of freedom too)
- **Tables**. Avoid tables or do not overuse

B

- Generally, a single column width measures 85-88 mm and a double column width measures 170-180 mm.
- Max height around 225mm (but that is with caption) (A pdf page is 210 x 276 mm)
- Font size should be 5-14pt
  - Axis labels 8-9pt
  - Axis numbering 7-8pt
  - Subpanel labels = 13pt bold
- Axis lines should be 0.75 pt (ungroup anything that looks weird)
- Axis ticks should be 1mm (short ticks 0.85, long ticks 1.35) (1.1, 0.6)
- All lines should be wider than 0.25pt (should probably be 0.5+)
- If things are weird, embed font
- Check spelling
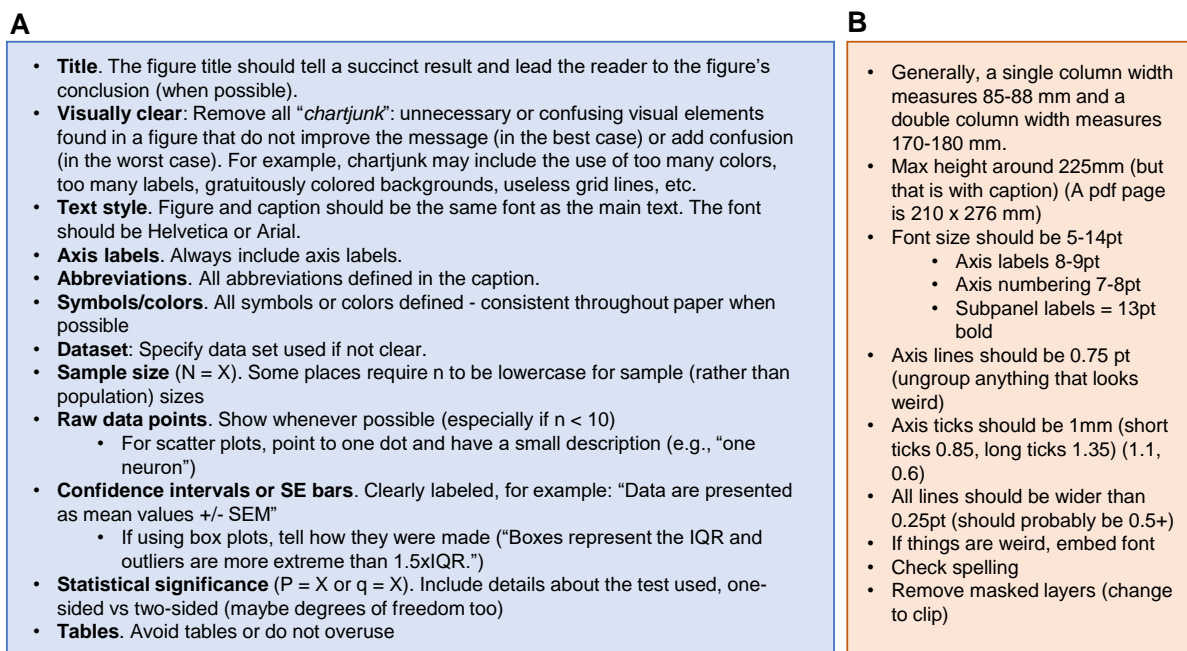- Remove masked layers (change to clip)

Figure 1: Making good figures requires attention to detail.
(**A**) General principals of making and labelling a good figure. These should help facilitate understanding and reproducibility. (**B**) Some more detailed technical specifics about making figures and their sizes. Inkscape is a great free vector figure editor and alternative to Adobe Illustrator.

Tips for captions and figures: All figures and tables should be inserted into the document as soon as possible after the first mention in the text. Make sure to number them sequentially as they are referenced in the text and provide a succinct descriptive caption. If there are multiple parts to a figure, give clear capital letter labels (A, B, C, etc.) to each panel. Make sure each figure and sub-figure are referenced somewhere in the text! The images should be saved in vector formats (such as PDF or EPS) to maintain high resolution. The exceptions to this are: 1) if you are including a photograph or 2) if you are plotting a file with thousands of overlapping individual data points that would each be represented in the vector file (use a bitmap format here). The title of the figure caption should state the overarching result. (This will often be similar to a results section title.) Captions should interpret the results briefly. The figures and captions should be able to be "stand alone" from the main text with enough context and methods for interpretation. (Many readers will only look at the figures and captions.) Subplots should contribute to one major overarching scientific point. If you can't come up with one conclusion/title for a figure, consider if it should be two figures. If two figures show the same conclusion, one should probably go to the supplement (e.g., demonstrating that a results replicates across cell types or with a different control).

2.3 Example results section to demonstrate references

Use `citep` to cite things parenthetically like this (Einstein, 1905; Kent et al., 2002). Or use `citet` to cite things in-text like this: Kent et al. (2002) did so many useful things!

3 Discussion

Approx 1 page

The discussion should be a broad overview of the significance of your findings to the community you are addressing. You can mirror the results section, but you don't have to. You should start with a brief summary of the main findings and then a paragraph for the accompanying context and interpretation of each of the big picture conclusions or contributions. Then you should address high-level limitations of the study (not just small limitations that you might address in your results/methods) Finally, end with future directions on a positive note. How can people use your model/framework results? This positive-critical-positive is a compliment sandwich.

Everything discussed should be within the frame of reference of your paper's major conclusions or contributions. You will likely want to address a shortlist of the key papers your paper speaks to, but this is not a literature review. Even if you don't talk extensively about a paper, you can still cite it if it is relevant to your conclusions. (There is little cost to being free with your citations!)

Contrary to popular belief, you can bring up a new result in the discussion, especially if it is preliminary or a response to reviewer critiques. Save this for findings that are in service to the discussion, rather than just a main finding of a paper. You can also use sub-sections if it is helpful in framing but this is somewhat non-standard.

Optional: Consider proposing a model or framework in your discussion that integrates or contextualizes your results. This can be accompanied by a schematic or "model" figure. This is a great way to emphasize the contribution of your study and guide future work.

First paragraph Brief summary of the paper that sets up the discussion of the major contributions. You can start by restating the major gap that you addressed (*Despite X, we previously had a limited understanding of Y. Here, we show/demonstrate Z. Together, these findings highlight/provide...*) This can parallel the last paragraph of your intro (where you outline the big picture questions and their context for the field) or parallel the results, but this is not required.

Middle Paragraph(s) (1-4 paragraphs) Major contribution/finding/conclusion and its context. What does your result/contribution mean? Who does it speak to and why do they care? What are the implications? Were your hypotheses correct, surprising, contradictory? If the results are not conclusive, why not? How would you perform further work to get to a more conclusive result?

Penultimate paragraph(s) (1-2 paragraphs) Limitations. Identify the broad categories/ themes of limitations and don't get too particular about the details of each limitation. (i.e., need more samples/ data, model misspecification or over-simplification). Keep it succinct, don't apologize, but show that you have thought about caveats. Should segue into the future directions. *Although our approach provides many novel benefits, it also has limitations that we hope future work will address.*

Last paragraph Future directions and final big picture conclusion. This could be broken out into a separate conclusions section. *As X happens, we anticipate that Y will facilitate understanding of Z which allows for W.* Why is this exciting for your audience? Where are the next steps? Note that this is could possibly result in scooped, but even if you are thinking of working on a next step, consider putting it out there, because science for all!! *We encourage future studies to address X.* Be positive about specific future opportunities inspired by your work. It leaves an impression).

4 Methods

<u>Goal</u>: Communicate the data and methods used to produce the results. Start writing this as you are doing your analyses. <u>Objective</u>: (1) To write clearly how you produced all parts of the results and (2) promote reproducibility of the findings. <u>Style</u>: Overall, the methods section should be clearly and thoroughly written. In theory, any knowledgeable reader (e.g., a graduate student working in your area) would be able to reproduce any of the findings just by reading your methods. <u>Organization</u>: Write subsections with clear subheadings for each method. These should be descriptive. Ideally, subsections will be structured so that the reader can "look up" the details for any subsection of the results. <u>Ordering subsections</u>: Subsection order should generally match the order of the results. However, if you repeat the same type of analyses in two separate parts of the paper, you can describe this once and refer to the sections those methods address. <u>Voice</u>: Write in the active past tense (E.g. "We computed..., we intersected..., we quantified..."). <u>Note on publicly available datasets</u>: Include a citation or url link to datasets used, along with the last download date. <u>Citations</u>: if you use it, cite it. There is no ambiguity here.

In some journals, the methods section precedes the results section, while in others it follows the Discussion section. If it comes before, then the text should provide some orientation and motivation. If it comes after, then it can be more of a list. Unless you know it will come first, write the methods section as if it were following the discussion section and then add orienting text later if necessary.

<u>Other tips</u>: Think of the methods section is a cookbook, and each subsection is its own chapter. Some chapters will discuss where to get the finest ingredients, some will discuss specific cooking techniques, and the rest will discuss the recipes that combine ingredients and technique. For each recipe (subsection), you must be clear on which ingredients are needed and the order that the ingredients are assembled.

You do not need to divulge every quotidian analytical detail (e.g., *I first loaded .tsv into a data frame using pandas (v14.0) function pd.read_csv(filename) and pivoted it into a table*). But you should provide every analytical detail relevant to reproducing the final data analyzed (e.g. *We removed all loci mapping to sex chromosomes using the subtract function from BEDTools (v2.2.1; Quinlan and Hall 2010)*.)

4.1 Some overarching part of your results

Data source 1

Methods about Data source 1 go here.

- For publicly available data: Dataset name, genome build, url/citation, last-download date, sample size, etc.
- For private data: Describe samples, sample collection, IRBs, names of kits and reagents used to process samples, sample size.
- Include any details about the dataset critical for analyses and generation of results (I.e. controls, selection criteria, covariates, ancestry, age, sex, status, cell type, tissue source, assay details).
- If applicable, explain how data were processed (i.e. excluded sex chromosomes, removed centromeric and repeat elements, etc.)
- Ask someone else in the lab to look over the draft to make sure you have not left out any essential details.

Analysis 1: descriptive title

Methods about Analysis 1 go here.

- State the data inputs.
- State your controls.
- State sample sizes.
- State any software package, its version, and arguments used to run the analysis.
- Explain any processing steps, the order, and the rationale of those steps relevant for producing the result figure.

4.2 Next methods section

Next sub-methods section

Next sub-methods section

4.3 Data availability

The publicly available data used for analysis are available in the following repositories:

The custom datasets we generated are available in the repository "[name here]" available here [link][citation][accession number].

4.4 Code availability

The publicly available code for analysis are available in the following repositories:

The custom code/software we generated are available in the repository "[name here]" available here [link][citation][accession number].

4.5 Acknowledgements

Acknowledge any help you received from people not in your group. List funding sources. Ask collaborators for information they want to include here.

This work was supported by the National Institutes of Health (NIH) General Medical Sciences award [award number] to [initials].

Acknowledge your computing resources (check their resources): "This work was conducted in part using the resources of the Advanced Computing Center for Research and Education (ACCRE) at Vanderbilt University, Nashville, TN."

4.6 Author Contributions

Summarize the contributions of each author to the project. This section must be approved by all group members. Use the categories, criteria, and style outlined here: http://www.cell.com/pb/assets/raw/shared/guidelines/CRediT-taxonomy.pdf

4.7 Competing interests

The authors declare no competing interests.

References

Einstein, Albert (1905). "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]". In: *Annalen der Physik* 322.10, pp. 891–921. DOI: http://dx.doi.org/10.1002/andp.19053221004.

Kent, W. James et al. (June 2002). "The Human Genome Browser at UCSC". In: *Genome Research* 12.6, pp. 996–1006. ISSN: 1088-9051. DOI: 10.1101/gr.229102.

Supplementary Information

Supplemental Text

Some supplemental text that doesn't fit in the main text or provides extra detail can go here.
We will use this space to outline some general tips.

Article tips, tricks, and nitpicks

- Formatting/scientific corrections
  - The first time you report a p-value, give the test used.
  - Report p values as $(P = 0.\#)$, be careful with $(P = 0)$ (use $< \#$ when P is very small)
  - Change "mutation" to "variant"...maybe change "SNP" or "SNV" to "variant(s)" if specificity is not important (i.e., don't forget that indels and CNVs/SVs are also types of genetic variants)
  - For genomic coordinates/distances, the format is # kb or # Mb (space between number and unit, not # Kb or # mb)
  - Use American spelling conventions.
- Organization
  - Put answers to the "why" question in the discussion
  - Bring important findings to the first sentence of a paragraph/section
  - With section headers, figure titles, and paper titles, try to state a result if possible (rather than a method, e.g. "Neanderthals were bald" rather than "Investigating hair patterns of Neanderthals")
- Wordsmithing
  - Where you can avoid being vague, be specific: instead of "establishes the relationship" say the relationship that was established. Instead of "Is incompletely understood" say what is the specific thing that is incompletely understood
  - Get rid of "metadiscourse". E.g., Change "Our findings XY and Z support previous findings, including studies by Smith et al and Johnson et al, which reported that Neanderthals might have been bald at one point" to "XY and Z support findings that Neanderthals may have been bald (citation)".
  - Don't claim you're the first ever (or save that for the cover later)
  - Generally speaking, avoid passive voice. If you can add "by zombies" after the verb and it still makes sense...rethink your phrasing
- Word choice
  - Change utilize to use
  - Change Impacted to influenced
  - Delete uses of "those/these" in reference to previous sentence (make sure the subject of the sentence is clear)
  - "Characterize" connotes that you don't really have a hypothesis (try quantify?)
  - Some favorite transition words: thus, nonetheless, furthermore, indeed, therefore
  - Fewer (discrete quantities) vs less (continuous quantities)
  - (e.g.,...) and (i.e.,...) can be helpful
  - "Data" is the plural of "datum".
  - Avoid personifying inanimate objects ("the gene wants to...")
  - Splitting infinitives is ok.
  - Contractions are not acceptable in academic writing.
  - Remove: Interestingly, surprisingly
- Punctuation
  - Parenthetical punctuation: clauses and sentences
  - Use an Oxford comma to separate the last and second-to-last elements in a list.
  - Beware of consistency between Hyphens(-), n-dash(–), m-dash(—)
  - With reporting intervals be consistent in how you use dashes and spaces (1-2 vs 1 - 2 vs $1 - 2$, etc)
  - Use a comma between independent clauses joined by a conjunction.
  - Use "that" for restrictive clauses (no commas). Use "which" for non-restrictive clauses (use commas).
  - Pluralization rules for abbreviations.

Supplemental Figures

Figure S1: This is a picture of Tony Capra. Nobody puts Tony in a supplement. . .
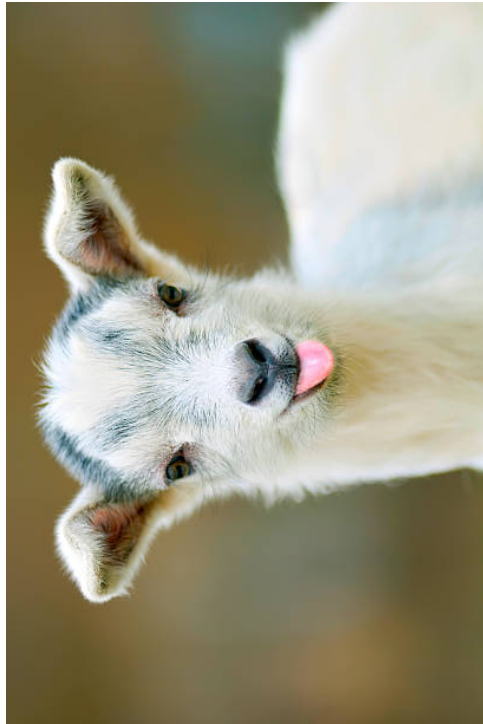

Figure S2: This cute sideways goat wishes you good luck with your manuscript.
This is how you rotate a figure but not the caption. To rotate the caption with the figure, use the function(s) `generateSidewaysFig` or `generateSidewaysFigSubpanels`.

Supplemental Tables

Models	Metric 1			Metric 2
	precision	recall	F-score	R@10
model 1	0.67	0.8	0.729	0.75
model 2	0.8	0.9	0.847	0.85

Table S1: Example table. I use `Excel2LaTeX.xla` to create easy tables. There are also other online websites. Look for one that is booktabs enabled. Here's good advice about making simple readable tables: https://users.umiacs.umd.edu/~jbg/images/tables.gif