Arc Science
Technical report

Molecullage

Danil Kotelnikov

Abstract. Molecullage makes figures of protein structures, sequences, contact maps, pipelines and plots for papers. A figure is written as an HTML page in millimetres, and the build prints it with a headless browser to a one-page vector PDF with live text and a 600 dpi PNG. A measurement script then reads the built figure and reports ink, empty space, type size, word counts, fonts and width against limits at three levels of strictness. A palette tool samples colours from a source figure and checks that they stay apart under four kinds of colour vision. A project setting decides whether one or two independent readers review a figure. The toolkit is open source and contains no learned model. Its limits are close to values measured on a set of published antibody-design figures.

In short

Molecullage is an open-source toolkit that builds publication figures from HTML pages laid out in millimetres and measures each built figure against fixed limits. It contains no learned model.

1. Pages

A figure is an HTML page whose dimensions are millimetres. The page script draws with a small kit: blocks, data labels, arrows, panel letters and mathematics set in Arial. Two widths are allowed, 178 mm and 89 mm. A project settings file fixes the type ladder in points: panel letter 9, title 7.5, label 7, smallest text 6.5, mathematics 8.4. It also fixes the word limits: 130 words in a schematic, 90 in a plate, 6 per box, 3 per arrow and 14 in the footnote. Eight template pages serve as starting points.

Each page can end with a claim boundary, one line at the foot of the figure that says what is real data and what is illustration. Plots are made with matplotlib and placed in the page at exact millimetre positions, in the same colours as the rest of the figure. The drawing rules are documented one by one, and each rule states what goes wrong when it is broken.

2. Build

The build prints a page with headless Chrome, Edge or Chromium. It writes a one-page PDF in which the text stays live, a PNG at 600 dpi and a copy of the page, and it reports the size in millimetres and the text sizes that it found. A token file with the palette can be passed to the build, which then recolours the page by replacing each family colour with the same family of the new file. The printer turns heat maps into pixels, and pixels cannot be recoloured, so colour keys are built as vector steps.

The tools are also exposed to a coding assistant through a skill and a tool server with twenty tools. They run the build, the measurements, the palette scripts and the structure renderer on the local machine, and only the fetching of a structure entry uses the network. A film engine in the same repository makes short narrated films from the same palette and renders.

3. Structure renders

Structures are rendered from Protein Data Bank entries in ChimeraX in a flat style. A render group has several transparent layers, for example the whole complex, the antigen alone, the antibody alone or a close-up. All layers of a group are cropped with one shared box, so they still stack on the page. Anchors are stored with each render: points in pixels that sit on a real residue or a real centre, such as a loop tip or the middle of an epitope, where the page ends its leader lines. The same module lists typed contacts.

4. Measurements

The measurement script reads the built figure. Ink is the share of pixels that are not near white. For empty cells the page is cut into a grid of 20 columns and 12 rows, and a cell is empty when less than 1% of its pixels are ink. The band is the tallest horizontal strip across the whole width with no ink, as a share of the height. The blank rectangle is the largest empty rectangle anywhere on the page, as a share of the area. The script also measures type size, picture width, words, fonts, pages and figure width.

At the normal level a figure passes when ink is at least 9% and empty cells are at most 25%. The longest empty band is at most 4% of the height and the largest blank rectangle at most 6% of the area. The smallest text is at least 6.5 pt and every picture at least 20 mm wide. The strict level raises the limits to 20% ink, 20% empty cells, a 2.5% band and a 4% blank rectangle, and it also fails a figure for too many words, a font other than Arial or a wrong width. The loose level reports failures as warnings. Each measurement is printed with its result and the reason for any failure.

5. Palette

The palette comes from a figure that may be reused. The sampler renders a PDF page at 300 dpi, groups nearby pixel colours and returns the exact colour that occurs most often in each group. Mixed colours at anti-aliased edges are rare and are dropped by a minimum share of 0.05% of the non-white pixels. An average such as a cluster centre would blend the edges into the fill and return a colour that is absent from the page. On a synthetic image with five planted fills, noise and an anti-aliased seam, the sampler returned the planted colours exactly.

The check measures the CIEDE2000 distance between colours after simulating normal vision, deuteranopia, protanopia and tritanopia. A pair counts as only as distinct as its worst case. The script tries every assignment of the sampled colours to the free roles and ranks the assignments by their smallest distance. The six core families of the shipped palette reach 11.5. This value is a computed colour distance. Dark tones for text are derived by lowering lightness in the OKLab colour space, and white text goes only on fills where it reaches a contrast of 4.5 to 1.

6. Independent reads

A project setting decides which independent reads a figure needs: none, a blind read, or a blind read and an overclaim read. The blind reader sees only the picture and answers three questions: what it shows, what could not be interpreted and what looks wrong. A second reader checks every number, label and claim against the paper and the data, or asks what a reviewer could say the picture proves that the work has not shown. The builder then fixes the findings. The gallery holds twelve figures from open-access papers, each reviewed by two readers and redrawn from the papers' own data and from structure entries. Where a redrawn figure says computed, the number comes from the data, and where a paper and its data disagree the figure says so.

7. Limits

The limits are close to values measured on a set of published antibody-design figures, and the documentation says to treat them as a guide. The measurements cover ink, empty space, type, words, fonts and colour distance. Numbers, labels and claims are checked by the readers when a project asks for them. The toolkit contains no learned model.

The toolkit was tested on Windows 11 with Node 24 and Python 3.11; commands for Linux and macOS are given. New structure renders need ChimeraX, which is not included. The code is released under the MIT licence, and the documentation and example figures under CC BY 4.0. Gallery figures redraw papers published under CC BY 4.0, and the originals are not copied. This page cites the repository documentation of the toolkit.

System design

Path of one figure from its sources to the measured outputs; every tool shown is part of the open-source release.

Implemented components

- Source figure.
- Structure entry.
- Plot data.
- Palette check.
- Flat render.
- Plot.
- Page in millimetres.
- Browser print.
- Vector PDF.
- PNG image.
- Measurements.
- Independent reads.

Connections

- Source figure → Palette check.
- Palette check → Page in millimetres.
- Structure entry → Flat render.
- Flat render → Page in millimetres.
- Plot data → Plot.
- Plot → Page in millimetres.
- Page in millimetres → Browser print.
- Browser print → Vector PDF.
- Browser print → PNG image.
- PNG image → Measurements.
- Vector PDF → Measurements.
- PNG image → Independent reads.
- Independent reads → Page in millimetres.
- Measurements → Page in millimetres.

Source

This report is the printed form of the page https://arcscience.app/molecullage/. The page is the source of record for every statement in it.