

AMP Atlas: A Browser-Only Workbench for Antimicrobial Peptide Exploration and Interactive Helical-Wheel Design

Truong-Son Hy

Department of Computer Science, University of Alabama at Birmingham, United States

thy@uab.edu

18 October 2026

Abstract

Accessible research software can connect published antimicrobial peptide (AMP) records to sequence analysis, molecular visualization, and interactive design without requiring a dedicated application server. We present AMP Atlas, an open-source, browser-only workbench that combines a DRAMP-derived database with local physicochemical calculations, published-neighbor retrieval, and sequence-guided structural comparison. The reported snapshot contains 12,784 source records, 10,579 canonical records, 5,183 source SMILES annotations, and 433 PDB chain structures. An inverse helical-wheel editor lets users specify peptide length, fill residues gradually, and immediately retrieve exact-length database matches under positional and property constraints. Selected-position residue frequencies support further editing while preserving the remaining constraints. A Web Worker performs searches and numerical calculations; a WebGL viewer provides interactive molecular representations. The static distribution occupies 138.8 MB and supports GitHub Pages hosting. Verification includes 23 Python tests, 30 browser-core/worker checks, a designer DOM integration suite, and agreement against fixed numerical references. This report describes the implemented algorithms and assumptions; it does not establish predictive accuracy for antimicrobial activity or experimentally validate designed peptides. The website is available at <https://hysonlab.github.io/amp-atlas/>, and the source code and database are available at <https://github.com/HySonLab/amp-atlas>.

1 Introduction

Research on antimicrobial peptides brings together sequence records, chemical modifications, physicochemical descriptors, literature annotations, and experimentally determined structures. For an individual researcher, moving between these representations can require several tools and repeated transfers of the same sequence. A lightweight, inspectable workbench can make published results easier to explore, compare, and communicate.

DRAMP 5.0 motivates this integration by enriching AMP records with chemical information and introducing FASTS, which places a SMILES representation in a FASTA-style annotation line [1]. AMP Atlas uses a downloaded general-record collection and associated source SMILES annotations; it does not reproduce every DRAMP collection or replace its curation. The distinction matters because a canonical sequence alone may omit chemical features that affect a peptide’s properties.

Helical-wheel projections provide another bridge between sequence and design. The classical representation places successive residues around an ideal helical axis [2]. Hydrophobic moments quantify the asymmetry of the resulting hydrophobicity distribution [3], and tools such as HELIQUEST support analysis of sequences with specified helical properties [4]. AMP Atlas adds an inverse interaction: users fill numbered wheel positions, inspect the corresponding linear sequence, and retrieve published records compatible with a partially

specified design.

The system is intended for exploratory research and preparation of hypotheses. Its contribution is the integration of a portable data snapshot, transparent calculations, molecular views, and incremental constraint-based search in a static deployment. This report documents that implementation [5], rather than proposing a new biochemical descriptor, alignment algorithm, or learned AMP predictor. In particular, the current software contains no trained model for activity, toxicity, minimum inhibitory concentration (MIC), helicity, or three-dimensional structure prediction.

2 System overview and scope

The public edition is a static website. HTML, CSS, JavaScript, JSON, PDB coordinates, and downloadable data files are served by GitHub Pages. Scientific calculations and searches execute on the visitor’s device. There is no application server, account system, remote inference endpoint, or API key required for visitors. An optional Python implementation remains available as a local reference and local API, but it is not required by the hosted edition.

Figure 1 separates the offline data preparation workflow from visitor-side operation. The interface exposes five connected workflows: browsing source records, analyzing an input sequence, comparing structures, editing a helical wheel, and downloading data. A completed wheel design can enter the

Table 1: Dataset composition, evaluated on 18 October 2026. Sizes use decimal MB (10^6 bytes).

Quantity	Value
DRAMP source records	12,784
Distinct source sequence strings	11,718
Canonical records, length 1–500	10,579
Distinct canonical sequences	10,056
Canonical records passing study heuristic	10,489
Records with source SMILES	5,183
Distinct source PubMed identifier strings	3,452
Stored PDB chain structures	433
Distinct PDB entries represented	378
Records flagged with a linked structure	315
Annotation shards	145
Search index / structure index	9.33 / 0.95 MB
SQLite database / complete static site	42.55 / 138.81 MB

same analysis workbench used for typed sequences; published results can return to the explorer. This reuse keeps literature evidence and design inputs connected.

Local processing avoids transmitting input sequences and uploaded PDB contents through the application’s scientific operations. This is a statement about the implemented data flow, not a formal privacy audit: hosting providers can log ordinary page and public-file requests, and following an external publication link contacts that provider. The current site requires HTTP(S) for workers and file fetches; it is not a guaranteed fully offline application.

3 Data collection and representation

3.1 Data sources and collection

The database was assembled from official DRAMP general-record and general-SMILES spreadsheets downloaded on 17 October 2026, together with referenced PDB files obtained through RCSB PDB [1, 6]. Source provenance and collection records are retained in the supplementary materials.

Table 1 summarizes the dataset used in this study. Counts refer to imported records or stored chains, as specified, and should not be interpreted as counts of unique chemical entities. For example, 11,718 distinct source sequence strings include noncanonical strings, whereas only 10,056 distinct sequences occur among canonical records. A PubMed identifier count is obtained from source annotation text; it is not an independently verified bibliography.

3.2 Validation and chemical annotations

Source sequences are retained as provided. A separate canonical flag requires a sequence of 1–500 residues over the 20-letter alphabet

$$\mathcal{A} = \{A, C, D, E, F, G, H, I, K, L, \\ M, N, P, Q, R, S, T, V, W, Y\}.$$

Noncanonical records remain browsable and downloadable; they are not silently converted to canonical peptides for calculation. Console input is uppercased and stripped of whitespace, then validated. FASTA headers and nonstandard residue symbols are rejected by the sequence console.

C-terminal annotations are mapped to amidated, free, or unknown status. Precomputed descriptors use the annotated amidation state when recognized; unknown status is calculated with free termini as an explicit approximation. Metadata retains source modification text and source SMILES. The software does not reconstruct covalent chemistry from SMILES, validate stereochemistry, or derive chemical identity from sequence equality. One imported row lacks a DRAMP identifier and is retained under the source-derived string None in an unassigned shard, rather than inventing an accession.

3.3 Structure acquisition and linking

The build report records 591 attempted referenced PDB downloads, including two failed downloads. The collection is therefore a bounded, source-referenced snapshot, not a complete structure survey. The parser reads the first PDB model, retains canonical ATOM C α positions, accepts blank or A alternate-location identifiers, and requires finite coordinates. Canonical chains with 5–150 observed residues are eligible for inclusion.

A candidate chain must match a referenced peptide’s canonical sequence exactly or match a contiguous observed fragment covering at least 80% of that peptide. A representative source record is selected for each retained chain, and relevant first-model atom records are saved for visualization. Display associations additionally use exact observed sequence equality across records. These rules produce 433 stored chains and 315 source records flagged with a local structure. They do not establish matching terminal chemistry, modifications, stereochemistry, or assay conditions. Missing coverage can reflect parser and collection rules as well as absent experimental structures.

3.4 Storage and distribution

SQLite stores peptide records, calculated properties, structure records, and provenance, with indexes on names, sequence, length, and structure associations. The static build emits a searchable JSON summary index, a structure index, and 145 annotation shards. Full annotations are loaded on demand, with at most 12 shard promises retained in the worker’s insertion-ordered cache. PDB atom files are fetched when required for viewing. SQLite and bulk exports are downloaded only when requested.

The complete static distribution occupies approximately 138.8 MB, including downloadable exports. The approximately 9.33 MB uncompressed search index and 0.95 MB structure index load separately from detailed records and atom files, reducing the amount of data required for initial exploration. Transfer compression depends on the host.

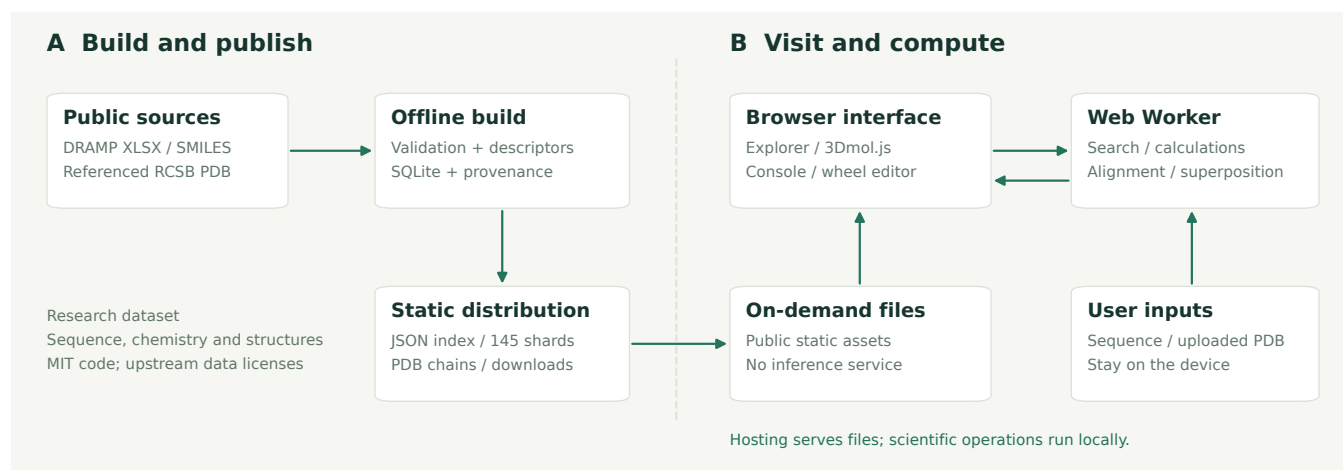


Figure 1: System architecture. Data preparation occurs offline; hosting distributes a fixed set of public files. The main browser thread manages the interface and molecular viewer, while a Web Worker owns searchable data and executes numerical operations. User-provided sequences and PDB contents are processed locally. Arrows indicate data flow, not measured network traffic.

4 Interface and research workflows

4.1 Record exploration and molecular visualization

The explorer searches record identifiers, names, sequences, and source-organism text, with pagination and filters for short peptides and linked structures. Selecting a record shows its sequence, calculated descriptors, original activity-related annotations, references, modification text, and available source SMILES. Links preserve a route back to DRAMP, PubMed, and structure providers.

The viewer embeds 3Dmol.js, a WebGL molecular visualization library [7]. Controls expose cartoon, sticks, spheres, and line representations, optional surfaces, residue/atom identification, labels, rotation, and image or coordinate export. The rendered molecular geometry comes from stored PDB atoms, not from the ideal wheel. Associated structures are labeled as sequence-matched references to avoid implying experimental determination of a user’s new design.

4.2 Sequence workbench

Users enter a canonical sequence, choose pH and terminal amidation, and request either sequence-based or descriptor-based neighbors. Output includes descriptor cards, amino-acid composition, a residue hydropathy profile, a charge–pH curve, an ideal-helical projection, alignments, and source-study links. Up to 15 distinct-sequence neighbors are returned. These are nearest records according to the selected numerical rule, not predicted successful analogues.

4.3 Inverse helical-wheel designer

The designer accepts lengths from 1 to 72 residues. Each numbered slot contains one canonical residue or an empty value. Users can click slots, choose from a residue palette, type residue letters, navigate with arrow keys, or clear a position

with Delete. Filling advances in sequence order. Imports, clears, fills, and resizing participate in an undo history of up to 40 edits. Resizing preserves the shared sequence prefix.

The linear output reads positions in numbered N-to-C order. It never sorts residues by their plotted polar angle. Empty positions appear as question marks in the displayed pattern; they are search wildcards, not actual residues. Copy, FASTA export, and transfer to full-sequence analysis require every slot to be filled. SVG export is available for incomplete wheels as well.

Live search is enabled by default. After a 250 ms pause in editing, the current length, specified residues, selected position, and property filters are sent to the worker. A count summary appears beside the editor, and up to 100 matching source records are displayed. Users may pause automatic search and submit manually. Filters support inclusive lower/upper bounds for charge, GRAVY, and hydrophobic percentage, plus a linked-structure requirement. Invalid bounds produce a validation message.

The selected-position panel lists observed residue counts among records compatible with all other positions and filters. Clicking a residue applies it to the design. Counts are explicitly identified as database frequencies: duplicates, source selection, and publication bias can influence them. They are neither model-generated recommendations nor probabilities of activity.

During partial entry, the interface reports how many assigned residues are basic (K/R/H), acidic (D/E), or hydrophobic (A/V/I/L/M/F/W/Y). It does not extrapolate a full mass, charge, or hydrophobic moment from a partially specified sequence. Once complete, the wheel panel displays mass, pH-7 charge, GRAVY, hydrophobic percentage, estimated pI, and ideal-helical moment under free-terminus assumptions. The workbench provides further analysis with its own pH and amidation controls.

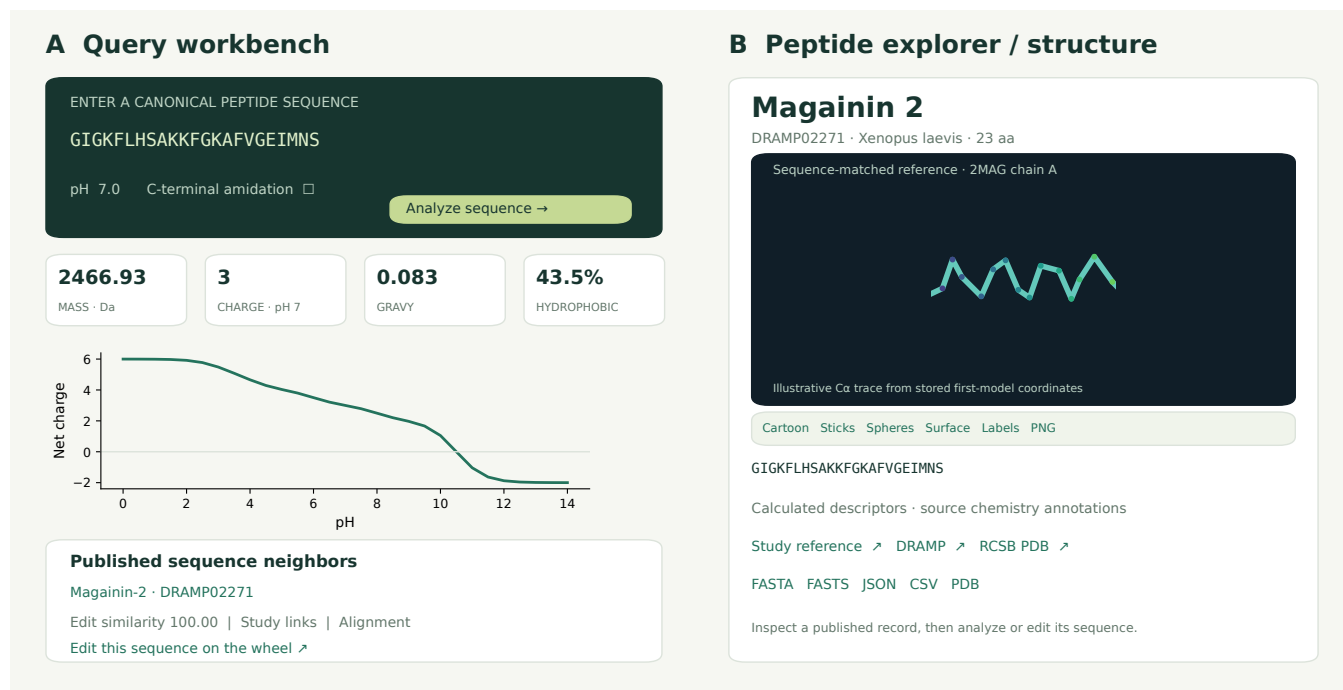


Figure 2: Source-derived interface illustrations for the sequence workbench (A) and peptide explorer (B), using the actual magainin 2 sequence, descriptor outputs, and stored PDB 2MAG chain A data [8]. These are vector reconstructions of selected interface components, not browser screenshots. The molecular panel shows an illustrative projection of stored C α coordinates; the live site uses the all-atom 3Dmol.js viewer [7]. The illustrated sequence is 23 canonical residues; this does not imply that every deposited residue or chemical group is represented in the descriptor model.

5 Algorithms

5.1 Sequence-derived descriptors

Let $s = (a_1, \dots, a_L)$ be a canonical sequence, n_a its residue counts, and I_{am} indicate C-terminal amidation. The reported average molecular mass in daltons is

$$M(s) = 18.0153 + \sum_{i=1}^L m(a_i) - 0.9848 I_{\text{am}}, \quad (1)$$

where $m(a)$ is the residue mass listed in Appendix A. This canonical linear-chain model omits disulfide oxidation, cyclization, nonstandard residues, N-terminal modifications, and adducts.

Charge uses independent Henderson–Hasselbalch terms. With positive groups $\mathcal{P} = \{\text{N terminus}, \text{K}, \text{R}, \text{H}\}$ and negative groups $\mathcal{N} = \{\text{C terminus}, \text{D}, \text{E}, \text{C}, \text{Y}\}$,

$$Q(p) = \sum_{g \in \mathcal{P}} \frac{n_g}{1 + 10^{p-pK_g}} - \sum_{g \in \mathcal{N}} \frac{n_g}{1 + 10^{pK_g-p}}. \quad (2)$$

Terminal group counts are one, except the C-terminal contribution is omitted for amidation. The fixed pK values are implementation parameters (Appendix A), not environment-dependent estimates. The charge curve samples pH 0–14 in 0.5-unit steps. Estimated pI uses 45 bisection iterations over $[0, 14]$; when no zero is bracketed, the calculation approaches an interval boundary and should not be interpreted as a measured pI.

For Kyte–Doolittle hydrophathy values $h(a)$ [9], mean hydrophathy (GRAVY) is

$$H(s) = \frac{1}{L} \sum_{i=1}^L h(a_i). \quad (3)$$

The ideal-helical hydrophobic moment follows the vector-sum construction [3], using the same, unnormalized Kyte–Doolittle values:

$$\mu_H(s) = \frac{1}{L} \sqrt{\left[\sum_i h(a_i) \cos \theta_i \right]^2 + \left[\sum_i h(a_i) \sin \theta_i \right]^2}, \quad (4)$$

where $\theta_i = (i-1)100^\circ$. Its numerical scale depends on the chosen hydrophathy table; it is not interchangeable with moments calculated using another scale or normalization. Neither this descriptor nor a wheel predicts helicity.

Hydrophobic fraction is $100 \sum_{a \in \{\text{A}, \text{V}, \text{I}, \text{L}, \text{M}, \text{F}, \text{W}, \text{Y}\}} n_a / L$; aromaticity uses F/W/Y. The reduced-cysteine extinction estimate is $5500n_W + 1490n_Y$ in $\text{M}^{-1}\text{cm}^{-1}$, corresponding to the aromatic terms of the standard sequence-based coefficient model [10]. Cysteine terms are not included. Composition and the unsmoothed per-residue hydrophathy profile are also returned.

5.2 Published-neighbor eligibility

Similarity and wheel searches require a canonical record with a source-study annotation. The implementation accepts either

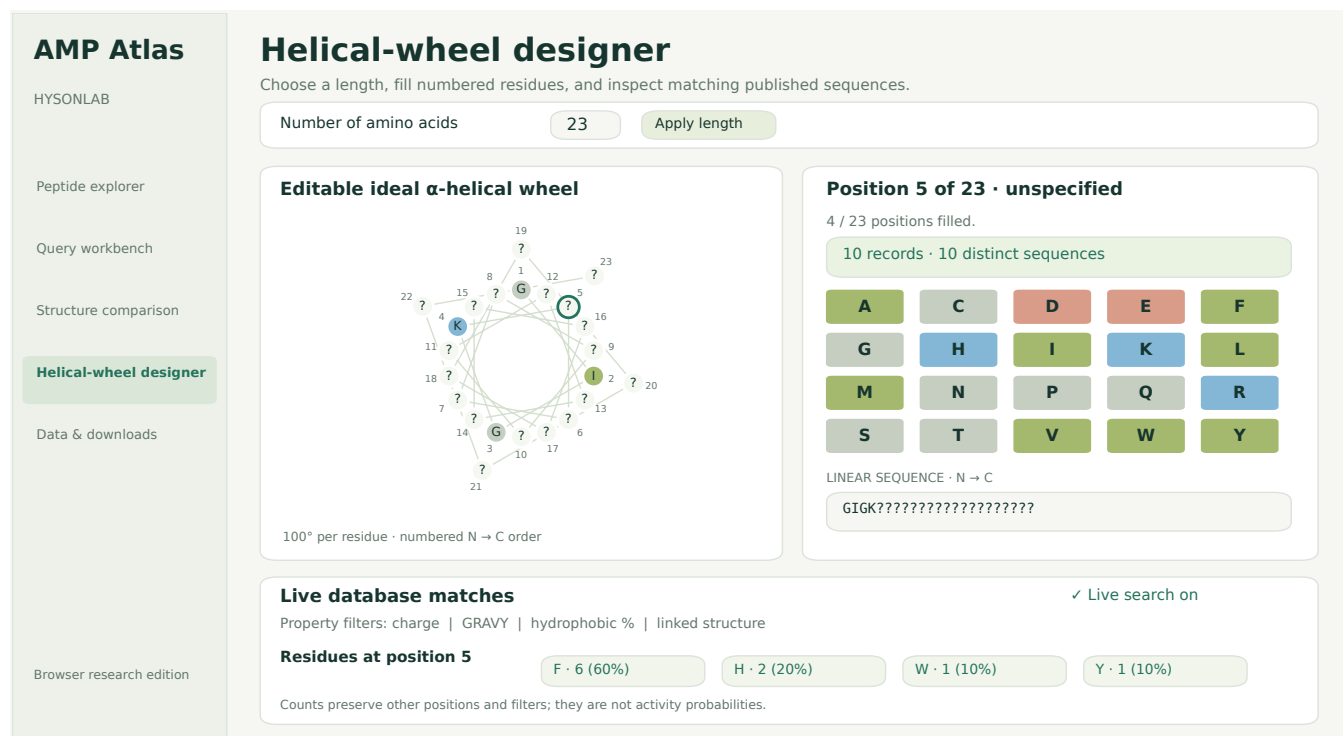


Figure 3: Source-derived reconstruction of the live helical-wheel designer with length 23 and the first four residues set to GIGK. The actual worker returns 10 matching source records. Position 5 is selected; compatible residues are F (6), H (2), W (1), and Y (1). Geometry, colors, position numbering, pattern, and query outputs come from the current implementation. The report rearranges selected components for print; this is not a browser screenshot. Extra rings prevent slot overlap after residue 18 and do not represent a cyclic peptide.

a six-to-nine-digit token in the PubMed field, or a title longer than 12 characters together with a reference longer than six characters, excluding the placeholder titles “not included yet” and “not found.” This is a pragmatic annotation heuristic, not manual verification of each publication or its relevance. It admits 10,489 canonical records in the reported snapshot.

The explorer retains all source records. Neighbor lists deduplicate by exact sequence after ranking and return the first representative, with identifier ordering resolving score ties. Identical sequences can correspond to different modifications, organisms, or experiments; representative selection does not merge their chemistry or evidence.

5.3 Global sequence similarity

Sequence retrieval uses unit-cost Levenshtein distance. For sequences s and t , the dynamic program is

$$D_{i,0} = i, \quad D_{0,j} = j, \quad (5)$$

$$D_{i,j} = \min \begin{cases} D_{i-1,j} + 1, \\ D_{i,j-1} + 1, \\ D_{i-1,j-1} + \mathbf{1}[a_i \neq b_j]. \end{cases} \quad (6)$$

The browser evaluates distances with rolling unsigned-16-bit rows: time is $O(|s||t|)$ per candidate and working memory is $O(|t|)$. Full alignments retain the traceback table and prefer a diagonal step on ties. This is a global minimum-edit alignment [11]; it is not a substitution-matrix or affine-gap protein

alignment. The optional Python distance implementation uses a bit-vector method [12] but computes the same unit-cost distance.

The displayed score is

$$S_{\text{seq}}(s, t) = 100 \left(1 - \frac{D_{|s|,|t|}}{\max(|s|, |t|)} \right). \quad (7)$$

Alignment identity is the proportion of identical residue pairs over all alignment columns, including gaps. The workbench’s sequence coverage is aligned residue-pair count divided by query length. These quantities have different denominators and are reported separately.

5.4 Physicochemical neighbor score

Descriptor retrieval uses the five-vector

$$z(s) = (L, M, Q(7), H, \mu_H)$$

and scale vector $\sigma = (20, 2000, 3, 1, 1)$. The score is

$$S_{\text{prop}}(s, t) = \frac{100}{1 + \sqrt{\frac{1}{5} \sum_{k=1}^5 \left(\frac{z_k(s) - z_k(t)}{\sigma_k} \right)^2}}. \quad (8)$$

These scales are fixed engineering choices, not learned weights or dataset standard deviations. Similarity is not calibrated to functional equivalence. Query descriptors for retrieval always

use pH 7, even when the workbench displays charge at another pH. Stored candidate descriptors use their snapshot terminal assumptions. Wheel property filters likewise operate on stored values, including their existing numerical rounding; they do not refit chemical models during filtering.

5.5 Wheel geometry and inverse mapping

The editor retains slots in sequence order. For zero-based index i , its displayed angle and radius are

$$\phi_i = (100i - 90)^\circ, \quad (9)$$

$$r_i = 105 + 58\lfloor i/18\rfloor, \quad (10)$$

$$(x_i, y_i) = (c + r_i \cos \phi_i, c + r_i \sin \phi_i), \quad (11)$$

where c is half the SVG view-box size. Successive rings avoid collisions because the ideal projection repeats every 18 residues. Pixel radii and rings are layout parameters; they are not physical helix dimensions or peptide cyclization constraints. Linearization is simply the indexed concatenation $(a_1, \dots, a_L)$ after every slot is filled, so the inverse mapping is deterministic.

5.6 Incremental constraint-based retrieval

Let $w \in (\mathcal{A} \cup \{\star\})^L$ be a partial wheel, with $\star$ an unspecified position. Let $\mathcal{D}_L$ contain eligible database records of length L , and let $F(r)$ indicate that record r passes all active descriptor bounds and the optional structure flag. The matching set is

$$\mathcal{M}(w, F) = \{r \in \mathcal{D}_L : F(r) \wedge \forall i, w_i = \star \vee s_i(r) = w_i\}. \quad (12)$$

At least one residue must be specified. The result contains both source-record and distinct-sequence counts, as well as the positional-match count before property filters. Thus, a zero after filtering can be distinguished from a zero caused by the sequence pattern. Returning the first 100 records limits display size, not the underlying match count.

For selected position j , form $w^{(-j)}$ by replacing w_j with $\star$. The residue count and displayed percentage are

$$c_j(a) = \sum_{r \in \mathcal{M}(w^{(-j)}, F)} \mathbf{1}[s_j(r) = a], \quad (13)$$

$$f_j(a) = 100c_j(a)/|\mathcal{M}(w^{(-j)}, F)|. \quad (14)$$

When the denominator is zero, no frequencies are shown. Otherwise, nonzero residues are sorted by descending count, then alphabetically. The current implementation scans records and accumulates counts, taking $O(N_L L)$ time in the worst case. There is no combinatorial enumeration of all 20^L possible designs.

A monotonically increasing generation counter prevents an older response or error from overwriting a newer state (Algorithm 1). Debouncing reduces redundant calls during rapid editing. It does not terminate an already running worker computation: an older request may still consume time, but its response is ignored. Manual submission executes immediately. DOM and controller tests check pausing, cancellation of timers, late responses, and workbench handoff.

Algorithm 1 Live wheel search with stale-response protection

```

1: Initialize generation  $g \leftarrow 0$  and no pending timer
2: procedure EDIT( $w, j, F$ )
3:    $g \leftarrow g + 1$ ; cancel pending timer
4:   Clear results belonging to the previous state
5:   if  $w$  is empty or bounds are invalid then
6:     Show an input/status message; return
7:   end if
8:   if automatic search is paused then
9:     Wait for manual submission; return
10:  end if
11:  Set a 250 ms timer carrying  $(w, j, F, g)$ 
12: end procedure
13: procedure EXECUTE( $w, j, F, v$ )
14:  Request worker matches and selected-slot counts
15:  Await the result or error
16:  if  $v = g$  then
17:    Render the returned result or error
18:  end if
19: end procedure

```

5.7 Sequence-guided structure comparison

A query is either a stored chain or a locally uploaded PDB file, limited to 2 MB. For uploads, the parser selects a requested chain or the longest eligible canonical chain in the first model. Structural queries require 8–150 observed canonical residues. Each target is globally aligned to the query sequence; all aligned residue pairs, including substitutions, define correspondences. Targets require at least eight paired positions and coverage

$$C = \frac{K}{\max(L_q, L_t)} \geq 0.8, \quad (15)$$

where K is the number of aligned residue pairs. There is no additional identity threshold.

For paired C α coordinates q_i, t_i , rigid fitting minimizes

$$\min_{R \in SO(3), u \in \mathbb{R}^3} \frac{1}{K} \sum_{i=1}^K \|q_i - (Rt_i + u)\|_2^2. \quad (16)$$

The browser constructs Horn’s symmetric 4×4 quaternion matrix and obtains the eigenvector of its largest algebraic eigenvalue using Jacobi rotations [13]. The Python reference uses an SVD-based Kabsch proper-rotation fit [14]. Both prohibit reflection and scaling. RMSD is the square root of the minimized objective and is reported in angstroms. A fitted transform is applied to the full observed target trace for overlay.

The selected stored query chain is excluded as an exact identifier; targets are deduplicated by PDB identifier and observed sequence. Results are ordered by increasing RMSD, then decreasing coverage, then identifier, with at most 15 displayed. C α radius of gyration is the root mean squared distance to the coordinate centroid; end-to-end distance uses the first and last observed C α atoms. These are unweighted trace descriptors. The procedure is sequence-guided rigid comparison, not sequence-independent fold search, flexible superposition, or a test of shared biological function.

6 Reproducible examples and verification

6.1 Progressive search example

The running example is the 23-residue sequence

GIGKFLHSAKKFGKAFVGEIMNS,

associated with magainin 2 in record DRAMP02271. PDB 2MAG provides a structural reference from an NMR study in membrane-mimicking conditions [8]; AMP Atlas retains only its first-model observed canonical trace for numerical comparison. The NMR environment and retained-model choice limit interpretation of any overlay.

At fixed length 23 and with no property filters, constraining the first 1, 2, 4, 8, 12, and 23 positions yields 65, 22, 10, 6, 1, and 1 source records, respectively (Figure 4). Distinct-sequence counts are 64, 21, 10, 6, 1, and 1. After the prefix GIGK is entered, selected-position counts for residue 5 are F:6, H:2, W:1, and Y:1. These values are exact outputs of the deployed worker against the bundled snapshot, rather than simulated interface values. A separate pattern specifying only G at position 1 and K at position 4 yields two records when restricted to linked structures and nonnegative stored charge.

Under free-terminus, pH-7 assumptions, the example’s calculated mass is 2466.93 Da, charge is 3.000, GRAVY is 0.083, hydrophobic fraction is 43.5%, pI estimate is 10.501, and ideal-helical Kyte–Doolittle moment is 1.456. These outputs illustrate the implemented model, not measurements. The closest returned structural target for stored query 2MAG_41 is 2FLY_41, with RMSD 1.1335 Å and 87.0% coverage under the stated correspondence rule. This nearest result is not evidence of identical chemistry or activity.

6.2 Software and numerical verification

The application has 23 Python tests and 30 browser-core/worker checks, supplemented by a designer DOM integration suite. Coverage includes validation, descriptors, global alignment, structure fitting, reflection rejection, static subdirectory paths, shard retrieval, pagination, exports, neighbor modes, PDB uploads, wheel geometry for all lengths 1–72, wildcard search, bounds validation, frequency counts, and stale-response handling. The DOM suite exercises actual designer controls and transfer into the existing analysis workbench using a DOM emulator. It does not validate real-browser rendering or WebGL behavior.

Across 30 fixed descriptor fixtures, the stored rounded Python and JavaScript descriptor outputs agree exactly for mass, charge, pI, GRAVY, hydrophobic fraction, moment, aromaticity, and reduced extinction coefficient. Across 20 fixed fitting fixtures, rounded RMSD outputs agree exactly and the maximum absolute fitted-coordinate discrepancy is 3.72×10^{-14} Å. The tests additionally reject a reflected configuration as a zero-error proper rotation. These checks support numerical consistency between implementations; they do not establish external biochemical accuracy.

Table 2: Illustrative warm worker execution times in milliseconds. n is the timed repetition count; ranges are minimum–maximum, not confidence intervals. These are Node.js measurements, not browser interaction or network latencies.

Operation	n	Median	Range
Partial wheel	7	0.84	0.84–1.14
Filtered wheel	7	0.81	0.80–0.83
Sequence neighbors	3	878.66	875.44–886.72
Descriptor neighbors	3	20.08	19.38–20.45
Structure neighbors	3	60.43	60.33–64.39

Static integrity checks validate all 145 record shards, structure paths, database-copy equality, and asset paths. The application was also checked by the repository’s verification workflow, and the corresponding published application assets were compared against the tested generated files. For this report, source-derived figures and algorithmic tests were used because automated browser capture was unavailable; no new visual cross-browser or WebGL compatibility study is claimed.

6.3 Illustrative execution times

Table 2 reports warm execution times for the unmodified deployed worker loaded into a Node.js VM against local files. The machine used Linux x64, Node.js v24.19.0, and an Intel Core i7-9750H CPU identified at 2.60 GHz. One untimed warm-up preceded each task’s repeated measurements. Wheel timings use the length-23 pattern G at position 1 and K at position 4; the filtered case additionally requires a linked structure and nonnegative charge. Neighbor timings use the full example sequence; structure timing uses 2MAG_41.

The measurements exclude initial JSON parsing/loading, DOM updates, molecular rendering, network transfers, and the deliberate 250 ms debounce. They use a small number of repetitions on one machine; no throughput, mobile-performance, or hardware-independent responsiveness claim follows. In particular, exhaustive global sequence comparison is substantially slower than wheel filtering at this scale, motivating future indexing work.

7 Limitations and future work

Chemical and biological scope. Canonical linear-chain descriptors cannot fully represent modified, cyclic, branched, or stereochemically altered peptides. Source SMILES improves evidence preservation, but it is not interpreted by the present calculator. Assay annotations remain source text; the software does not normalize MIC units, experimental organisms, media, toxicity protocols, or uncertainty. A matching record does not validate a designed peptide, and a missing match does not establish novelty or inactivity.

Structural interpretation. The corpus covers a small fraction of source records and is selected through available DRAMP references, download limits, and sequence-linking criteria.

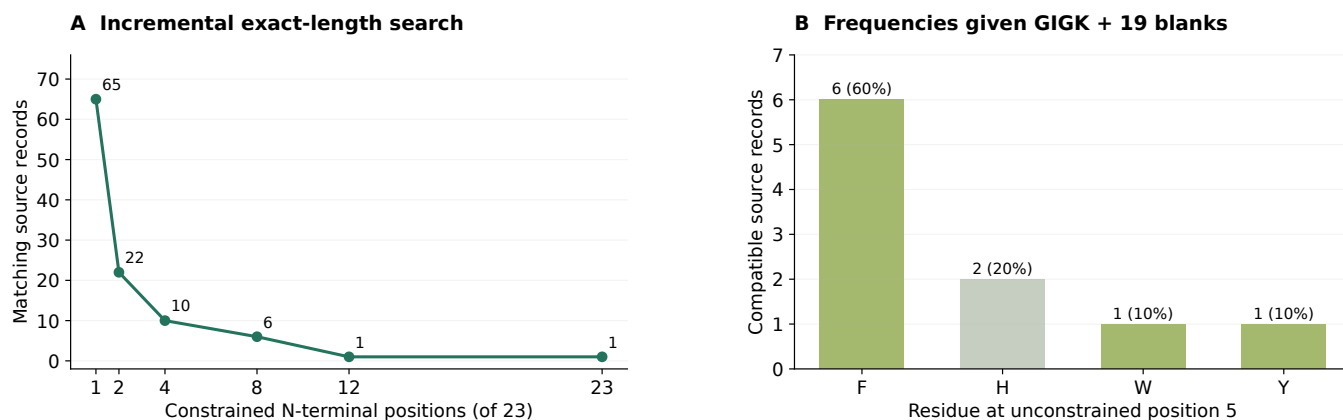


Figure 4: Reproducible query results from the bundled data. (A) Match counts as an increasing N-terminal prefix of the 23-residue example is specified. (B) Position-5 residue counts with GIGK fixed and the remaining positions blank. Each source record contributes once, so percentages reflect this particular collection and its redundancy. No activity labels or predictive probabilities enter these calculations.

First-model observed traces omit ensemble variability and unobserved residues. Rigid sequence-guided RMSD is sensitive to correspondence, length, and coverage, and unrelated short helices can superpose well. No fold-recognition or binding-mechanism benchmark has been performed.

Evidence and evaluation. Annotation heuristics can admit incomplete or erroneous study references. Duplicate source records influence selected-position frequencies, while sequence-deduplicated neighbor lists can hide chemically distinct examples. Current evaluation establishes software consistency and reproducibility for a fixed snapshot. It includes no prospective peptide synthesis, biological assay, user study, retrieval-relevance judgment, or comparison of discovery outcomes against another platform. The report’s interface figures are explicitly labeled source-derived reconstructions.

Scale and deployment. The static architecture reduces operational requirements for this collection, but the search index is loaded into each visitor’s worker memory. Sequence retrieval scans all eligible records, and full alignments require quadratic tables for selected candidates. For larger collections, length/position inverted indexes, compact binary data, additional sharding, or optional WebAssembly kernels could improve execution and transfer cost. A future larger deployment should measure browser memory, actual interaction latency, cache behavior, and device variability. The current 138.8 MB deployment alone does not establish scalability to much larger data.

AI extensions. A future learned model would require explicitly licensed training data, normalization of labels and chemistry, sequence-aware train/test separation, calibrated uncertainty, and independent evaluation. Such models should remain separate from deterministic descriptors and source frequencies so that users can distinguish calculations, observations, and predictions. These are development directions, not implemented results.

8 Availability and reproducibility

AMP Atlas is freely accessible through its public website, with source code and data distributed through GitHub [5]; both links are provided in the abstract. The application can also be run locally using a static HTTP server. SQLite, JSON, CSV, FASTA, and FASTS downloads support downstream analysis and reuse. FASTS exports retain source SMILES and source FASTS sequence text when present; this does not independently validate the encoded chemistry.

Supplementary materials provide the data provenance, exact query outputs, numerical evaluation results, and scripts for reproducing the reported analyses and figures. The repository includes the application build tools and verification suites. The evaluation harness executes the same scientific worker used by the website against locally stored data.

Application source is MIT licensed. DRAMP-derived material is attributed to its curators and distributed subject to its upstream CC BY 4.0 terms [1]; PDB archive data follow CC0, with citations to structure depositions and associated studies [6, 8]. Bundled visualization code retains its own license and notices. These upstream licenses remain distinct from the application license. AMP Atlas is an independent research workbench and is not an official DRAMP release.

9 Conclusion

AMP Atlas connects source peptide records, transparent sequence calculations, molecular inspection, and inverse helical-wheel editing in a browser-only application. The live designer treats partial sequences as explicit constraints and exposes the source records and selected-position counts behind its feedback. The available data, transparent algorithms, and numerical verification support reproducible exploration. The current contribution is a portable exploratory research tool; experimental assessment, chemical generalization, and predictive modeling remain future work.

References

- [1] Tianyue Ma, Yihan Huang, Hongyu Liu, Wentao Wu, Linxin Ren, Fahad Akhtar, Maryam Nawaz, Zichun Hua, Heng Zheng, and Xingzhen Lao. DRAMP 5.0: an update integrating sequence and structure in a novel FASTS format for AI-driven design of antimicrobial peptides. *Nucleic Acids Research*, 2026. doi: 10.1093/nar/gkag989. Advance article gkag989, published 15 October 2026.
- [2] Marianne Schiffer and Allen B. Edmundson. Use of Helical Wheels to Represent the Structures of Proteins and to Identify Segments with Helical Potential. *Biophysical Journal*, 7(2):121–135, 1967. doi: 10.1016/s0006-3495(67)86579-2.
- [3] David Eisenberg, Robert M. Weiss, and Thomas C. Terwilliger. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature*, 299(5881): 371–374, 1982. doi: 10.1038/299371a0.
- [4] Romain Gautier, Dominique Douguet, Bruno Antonny, and Guillaume Drin. HELIQUEST: a web server to screen sequences with specific α -helical properties. *Bioinformatics*, 24(18):2101–2102, 2008. doi: 10.1093/bioinformatics/btn392.
- [5] Truong-Son Hy. AMP Atlas: source code and research data. GitHub repository, 2026. URL <https://github.com/HySonLab/amp-atlas>. Accessed 18 October 2026.
- [6] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1): 235–242, 2000. doi: 10.1093/nar/28.1.235.
- [7] Nicholas Rego and David Koes. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics*, 31(8):1322–1324, 2015. doi: 10.1093/bioinformatics/btu829.
- [8] Jennifer Gesell, Michael Zasloff, and Stanley J. Opella. Two-dimensional ¹H NMR experiments show that the 23-residue magainin antibiotic peptide is an α -helix in dodecylphosphocholine micelles, sodium dodecylsulfate micelles, and trifluoroethanol/water solution. *Journal of Biomolecular NMR*, 9(2):127–135, 1997. doi: 10.1023/a: 1018698002314.
- [9] Jack Kyte and Russell F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982. doi: 10.1016/0022-2836(82)90515-0.
- [10] C. Nick Pace, Felix Vajdos, Lanette Fee, Gerald Grimsley, and Theronica Gray. How to measure and predict the molar absorption coefficient of a protein. *Protein Science*, 4(11):2411–2423, 1995. doi: 10.1002/pro.5560041120.
- [11] Robert A. Wagner and Michael J. Fischer. The String-to-String Correction Problem. *Journal of the ACM*, 21(1): 168–173, 1974. doi: 10.1145/321796.321811.
- [12] Gene Myers. A fast bit-vector algorithm for approximate string matching based on dynamic programming. *Journal of the ACM*, 46(3):395–415, 1999. doi: 10.1145/316542. 316550.
- [13] Berthold K. P. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America A*, 4(4):629–642, 1987. doi: 10.1364/ josaa.4.000629.
- [14] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5): 922–923, 1976. doi: 10.1107/s0567739476001873.

A Calculation parameters

Table 3 lists the complete canonical residue parameter tables used by the implementation. Mass values are average residue masses, not free amino-acid or monoisotopic masses. The hydropathy column reproduces the Kyte–Doolittle scale [9]. Other assumptions and units are defined in Section 5.

Table 3: Canonical residue parameters. m is average residue mass in Da; h is the Kyte–Doolittle hydropathy value.

Residue	m	h	Residue	m	h
A	71.0788	1.8	M	131.1926	1.9
C	103.1388	2.5	N	114.1038	−3.5
D	115.0886	−3.5	P	97.1167	−1.6
E	129.1155	−3.5	Q	128.1307	−3.5
F	147.1766	2.8	R	156.1875	−4.5
G	57.0519	−0.4	S	87.0782	−0.8
H	137.1411	−3.2	T	101.1051	−0.7
I	113.1594	4.5	V	99.1326	4.2
K	128.1741	−3.9	W	186.2132	−0.9
L	113.1594	3.8	Y	163.1760	−1.3

Charge calculations use pK values 8.0 (N terminus), 3.1 (C terminus), 10.5 (K), 12.5 (R), 6.0 (H), 3.9 (D), 4.1 (E), 8.3 (C), and 10.1 (Y). These are the fixed values in the reported implementation; no experimental fit was performed for this report. Candidate and query comparison assumptions should be considered when interpreting relative rankings.

B Figure provenance

The report’s figures are generated from the current interface geometry and actual worker output. The original wheel SVGs exported from the designer DOM are included as supplementary figure sources. The example molecular trace is derived from local PDB 2MAG chain A coordinates, with the original study cited in the caption. No source-paper figure is reproduced.