Suppl. 2. Model implementation details.

1. Preprocessing and Metadata Extraction

We extracted PMIDs, author names, publication years, and raw affiliation strings from PubMed XML to prepare a standardized dataset for the entity linking model.

1.1. Metadata Extraction and Structuring

We parsed each <PubmedArticle> node sequentially and flattened the hierarchical XML into a row-based tabular format. We specifically extracted PMIDs and publication years, excluding records with missing year information. To preserve affiliation-relevant contexts and maintain its structural representation, we normalized nested XML elements into key-value fields by concatenating multiple instances of the same tag within a single column using semicolons (;).

For author–affiliation matching, we first stored author names and affiliation strings as indexed concatenated fields (e.g., "1_Name; 2_Name" and "1_Affiliation A; 2_Affiliation B"). We then expanded these fields into a row-level representation, generating one unique row for each author–affiliation pair. We excluded articles lacking any affiliation information to prevent unverifiable assignments.

1.2. Author–Affiliation Reconciliation

We reconciled author–affiliation mappings to address heterogeneous name formats and block affiliation assignments. Since PubMed XML parsing often incorrectly simplifies composite initials (e.g., reducing T.M.L.-M. to L.), we expanded the initial-generation mechanism to produce multiple candidate variants, including traditional (L.), hyphenated (L.-M., L.M.), and full composite (T.M.L.-M.) forms.
To resolve cases where affiliation blocks were incompletely assigned to a subset of authors, we implemented a structured repair strategy:

1. Candidate Generation: We generated all possible initial candidates from each author name.
2. Pattern Detection: We identified parenthesized initials and full-name patterns embedded within the affiliation texts.
3. Hierarchical Mapping: We filled missing mappings by prioritizing initial-based matching followed by full-name matching. In cases of ambiguity, we determined priority using string length and frequency criteria.

1.3. Affiliation Text Cleaning and Normalization

We applied a pattern-based cleaning procedure to the raw affiliation strings using Python's *re* module. This process transformed raw strings into standardized inputs for the entity-linking model through three operations:

1. **Noise Removal**: We removed non-institutional elements including emails (identified by @), electronic address phrases (e.g., text following "electronic address:" strings), phone/fax patterns (e.g., Tel, Fax, +82-, 000-0000-0000), postal codes (US 5-digit and 5+4 formats), and address keywords (e.g., PO Box, Zip Code).

2. **Artifact Cleansing:** We eliminated leading punctuation (commas, periods, asterisks, hyphens, daggers †, brackets, pilcrows ¶) and removed all content enclosed within parentheses () and square brackets []. We removed standalone start-of-numerics/string tokens at the beginning but explicitly preserved ordinal expressions (e.g., 1st, 2nd, 3rd).
3. **Standardization:** We lowercased all text, collapsed repeated commas and whitespace into single instances, and trimmed leading or trailing whitespace and special symbols.

2. Stage 1: Exact Matching with Geographic Validation

We configured spaCy's EntityRuler to perform rule-based exact matching of institutional and geographic entities. This stage normalizes name variants and validates extracted countries/regions against ROR metadata to filter homonymous matches.

2.1. Geographic Knowledge Base and Pattern Generation

To enable precise disambiguation, we constructed a geographic dictionary aggregated from World Bank Open Data, United Nations M49 standards, and ISO 3166-1 alpha-2 codes. This dictionary normalizes geographic tokens through three primary mappings:

1. Country_synonyms: Maps variants (e.g., "America", "S.Korea") to canonical forms (e.g., "usa", "south korea").
2. Us_states: Resolves U.S. state abbreviations (e.g., "CA", "NY") to full names to ensure consistency with ROR.
3. City_state_mapping: Associates major cities with parent administrative regions or states.

We initialized a blank spaCy pipeline with English and two distinct EntityRuler components:

1. Organization Ruler: Matches institution names labeled as ROR. We sorted patterns by length in descending order (longest-first priority) to prevent partial overlaps. We explicitly excluded common geographic terms (e.g., "Washington", "Oxford") from organizational patterns to prevent spurious matches.
2. Location Ruler: Matches geographic entities labeled as COUNTRY or REGION, using the dictionary mappings to normalize geographic expressions to canonical forms.

2.2. Dual-Entity Extraction and Candidate Retrieval

We started the exact matching process with two parallel extraction tasks:

1. Institution Extraction: The Organization Ruler identifies substrings matching canonical or variant institution names. Since names may be non-unique (e.g., generic university names), a single match may link to multiple ROR records. We retrieve all associated ROR IDs as provisional candidates, which will be disambiguated in the subsequent validation.
2. Geographic Extraction: Concurrently, the Location Ruler extracts country and region entities from the affiliation text to serve as contextual constraints in the subsequent validation.

2.3. Geographic Validation and Candidate Classification

To resolve ambiguities arising from identical institutional names, we validated provisional candidates by

intersecting their ROR geographic metadata (country and region fields) with geographic entities extracted from the affiliation text. We classified candidate institutions into three outcome categories:

- Strict Match: The candidate's country and region metadata both intersect with extracted geographic cues. This represents the highest-confidence alignment.
- Partial Match: The candidate's country metadata intersects, but region-level alignment cannot be established (due to missing text or non-overlapping regions). These are retained but flagged as ambiguous.
- Rejected: The candidate's metadata conflicts with extracted country information or lacks intersection.

To control false positives, we accepted "Strict Matches" as valid links at Stage 1. Affiliation strings with no surviving candidates ("Rejected") or only ambiguous "Partial Matches" were forwarded to Stage 2 (Selective Fuzzy Matching) for recall recovery.

3. Stage 2: Selective Fuzzy Matching on Residuals

Our purpose was to increase recall while maintaining precision. We applied selective fuzzy matching to affiliation strings unresolved by Stage 1 ("residuals"). This stage addresses the residuals: orthographic variation, abbreviations, partial mentions, and typographical deviations not captured by exact matching. Such residuals cannot be addressed by a single similarity measure sufficiently. Thus, we employed a *multi-similarity weighted scoring method* that integrates similarity measures, rare-token features, and geographic adjustments.

3.1. Normalization and Geographic Blocking

Prior to scoring, we applied standardized preprocessing, including NFKD-based Unicode normalization, lowercasing, whitespace compression, email and special characters removal, abbreviation normalization, and the attenuation of non-discriminative organizational terms. We retrieved candidate organizations from the ROR corpus and applied *geography-based blocking*. We intersected the extracted 1–3-gram tokens with region and country lexicons (REGION_LEX and COUNTRY_LEX). To reduce the search space and prevent spurious matches, we eliminated candidates lacking geographic correspondence.

3.2. Multi-Similarity Integration and Base Score

We computed four similarity indicators for each candidate:

- **Token Set Ratio:** Measures set-based similarity insensitive to word order and repetition; the ratio captures order-insensitive organizational word-set overlap.
- **WRatio:** Measures Levenshtein-based global similarity, rewarding close spelling and length alignment.
- **Partial Ratio:** Detects cases where the core organization name is embedded in a longer affiliation string.
- **Rare Token Overlap:** Emphasizes distinct morphological tokens (e.g., long or alphanumeric strings).

We combined these metrics into a weighted base score (*BaseScore*):

$$BaseScore = w_{ts}R_{ts} + w_wR_W + w_pR_p + w_{rto}R_{rto}$$

We discarded candidates with a partial ratio below a certain threshold (MIN_PARTIAL) to suppress

spurious matches driven by short or coincidental substrings.

3.3. Geography-Based Adjustment and Final Scoring

We applied a geography-based adjustment (*GeoAdjust*) by comparing inferred geographic cues in the affiliation string with ROR metadata. We assigned a bonus score when cues matched and a penalty when they diverged. If geographic information was missing on either side, the adjustment was set to zero.
We computed the *FinalScore* by capping the adjusted base score:
$FinalScore = min(CAP, BaseScore + GeoAdjust)$, CAP=120
We retained candidates with FinalScore above a certain threshold (MIN_SCORE). The two highest-scoring candidates were returned as the primary and secondary predictions.

3.4. Parameter Optimization

We optimized parameters using the Gold Standard dataset split into train:validation:test sets (6:2:2). We partitioned the data based on the unique (PMID, num) key to prevent data leakage across splits.
We performed a comprehensive grid sweep over MIN_SCORE, MIN_PARTIAL, geography-adjustment strengths, geographic blocking activation, and component weights. Model selection prioritized a precision-weighted utility function $TP - \lambda \times FP$ where $\lambda = 2$. We resolved ties by sequentially comparing Precision, Recall, and F1-scores.

The final parameter configuration used in our analysis was:

- Thresholds: MIN_SCORE = 88, MIN_PARTIAL = 90
- Optimal Weights**:** w_{ts} = 0.35, w_w= 0.35, w_p = 0.10, w_{rto} = 0.20