

Corpus scale reassembly of fragmented bamboo slips using excavation context and latent slip properties

Jie Yang

Hubei University of Economics

Weishan Yang

yangweishan@hbeu.edu.cn

Hubei Engineering University

Article

Keywords: bamboo slips, fragment reassembly, excavation context, Shuihudi, computational heritage science, constrained optimisation

Posted Date: August 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10543021/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at npj Heritage Science on September 17th, 2026. See the published version at <https://doi.org/10.1038/s40494-026-02993-8>.

Corpus scale reassembly of fragmented bamboo slips using excavation context and latent slip properties

Jie Yang¹ and Weishan Yang^{2*}

¹College of Art and Design, Hubei University of Economics, Wuhan, China.

²School of Physics and Electronic Information Engineering, Hubei Engineering University, Xiaogan, China.

*Corresponding author(s). E-mail(s): yangweishan@hbeu.edu.cn;
Contributing authors: jieyang@hbue.edu.cn;

Abstract

Excavated bamboo slips are a primary written record of early China, but reach conservators as thousands of fragments whose grouping into slips is lost, and rejoining them must precede reading, dating or display. Computational matchers return ranked candidate lists that a specialist still resolves by hand. This study treats reassembly as one constrained optimisation over a whole corpus, carrying slip length, width and scribal hand as latent properties of the slip rather than as pairwise features. Applied to a corpus scaled to the Shuihudi Qin find of 3253 fragments, the framework recovered 33.4% of multi fragment slips intact where maximum weight bipartite matching recovered 0.5%, and reduced joins that fuse two different slips from 2122 to 380. Results reveal that excavation context is the most valuable of five evidence sources tested, yet no published system uses it. The approach enables corpus scale reassembly from evidence excavations already record.

Keywords: bamboo slips, fragment reassembly, excavation context, Shuihudi, computational heritage science, constrained optimisation

1 Introduction

Bamboo and wooden slips carried the administration and the law of early China, and the corpora recovered from tombs are primary evidence for the period. The Qin legal manuscripts from Tomb 11 at Shuihudi in Yunmeng, Hubei, excavated in 1975, are the best known example [1]. That find produced 1155 slips together with 80 loose fragments. The slips are 23.1 to 27.8 cm long and 0.5 to 0.8 cm wide, and each carried two or three cords that bound it into a roll. Bamboo of this kind survives only where burial keeps it waterlogged and anoxic, so the material reaches a conservation laboratory soft, dimensionally unstable and already broken [2, 3].

Rejoining the pieces is the gateway task for any find of this kind. Until fragments are grouped into slips and ordered within each slip, the text cannot be read, the document cannot be dated by its content, and neither display nor facsimile publication can proceed. Rejoining also governs conservation planning, because the number and shape of the reconstructed objects determines how they are mounted and stored. It is slow work. For a few thousand fragments the number of possible pairs runs into the millions, and each candidate has to be judged by a specialist against the break surface, the writing and the physical form of the piece.

Computational help has arrived for one half of the problem. Matching networks now score pairs of fragments well. WisePanda generates its training pairs from the physics of bamboo fracture and deterioration, and raises Top-50 matching accuracy from 36% to 52% among more than one thousand candidate fragments, with archaeologists reporting about twenty times faster work [4]. Work on Dunhuang manuscripts predicts the most likely reassembly candidates using a neural combinatorial layer and a multimodal language model [5], and patch level handwriting style has been used to group manuscript fragments by scribe [6]. Rejoining is one of several tasks on this material that machine learning has taken up. Damaged characters on Qin and Han slips are now restored by learned image models [7, 8], oracle bone fragments are rejoined from their images [9, 10] and their characters deciphered [11], and ancient texts are restored and attributed at corpus scale [12–14]. Reassembly is the step all of these depend on, since a character cannot be restored on a slip that has not yet been rebuilt.

The same pattern holds for other fragmented heritage material. Systems built for the Thera wall paintings acquire and match fresco fragments at volume, first by the shape of the break and then by colour and material, and later learn the match function from examples [15–17]. Thin artefacts of unknown geometry have been rejoined from their break surfaces alone [18], three dimensional fragments from their fracture regions [19, 20], flat irregular fragments from learned descriptors [21, 22], and papyri and ostraca from siamese and metric learning networks [23–25]. Methods of this kind are now common in heritage practice [26–28].

What all of these produce is a ranked list of candidate partners for each fragment. A list is not a reconstruction. Candidates conflict, because several fragments are often proposed as the partner of the same fragment, and only one of those claims can be right. A list also carries no model of the object being rebuilt, so a set of individually plausible joins can imply a slip longer than any slip ever was, or one whose width changes halfway down, or one whose binding notches fall where no cord ran. Resolving this concerns the corpus as a whole, and at present it is left to the specialist.

Global reasoning is not itself new. Square jigsaw solvers enforce loop consistency across the assembled grid [29], fragment placement has been posed as a shortest path problem over a graph of candidates [30], and a benchmark of real fresco fragments now scores whole reassemblies rather than pairs [31]. What these share is a model of the puzzle, its grid or its bounding frame. None carries a model of the object being rebuilt, and for slip material the object is where most of the surviving evidence sits.

An excavation records much more about a fragment than its broken edge. It records the context the piece came from, its dimensions, the position of any binding notch, and often an attribution of the hand that wrote it [32]. None of this reaches a pairwise matcher, and some of it cannot, because a notch only becomes evidence once a fragment has been given a position within a slip. Two questions follow. How should such evidence be used, and which of it actually determines how much of a corpus can be put back together?

This study answers both by stating reassembly as a single constrained optimisation over an entire corpus, in which each source of evidence enters as a constraint and removing each in turn measures what it was worth. The formulation is the substance of the first question. Slip length, width and scribal hand describe a slip and not a pair of fragments, and encoding them pairwise, as our own first model did, wastes them. Carried instead as latent properties shared along each reconstructed slip, the same evidence moves the entire operating curve of the method. The answer to the second question is that excavation context, which every excavation records and no published system uses, is the most valuable evidence of all.

2 Methods

Fig. 1 summarises the procedure. Fragments are described at their fracture ends, a trained encoder scores every candidate join, and one constrained optimisation over the whole corpus turns those scores into reconstructed slips.

2.1 Simulated corpora

A reconstruction cannot be scored without knowing which fragments truly join, and that is exactly what a real corpus does not tell us. We therefore work on simulated corpora, which is also the practice of the matching work we build on [4].

Each instance is a corpus of synthetic slips with complete ground truth. A slip is drawn with length in 231 to 278 mm and width in 5 to 8 mm, the ranges reported for the Shuihudi Qin corpus, and carries three binding cords at nominal fractions 0.06, 0.50 and 0.94 of its length, jittered by 3 mm, each cut as a notch in one edge. The substrate is an anisotropic bamboo texture, with longitudinal fibre striations that have a short correlation length across the width and a long one along the length, plus a few darker vascular bundles running the full length. Text is rendered from a clerical script (*lishu*) stroke grammar, a stack of horizontals crossed by one or two verticals with optional hooks and a flared terminal. Stroke width, slant, glyph scale, ink density and line pitch jitter are all fixed per scribal hand, so that the hand is a stable and learnable property of a slip. Palaeographers attribute slips to hands in exactly this

way [32], and automatic writer identification recovers such attributions from ink trace measurements and patch level style [6, 33, 34].

141

142 2.2 Fracture and taphonomy

143

144 Bamboo is strongly anisotropic, a graded composite of stiff fibre bundles in a soft
145 parenchyma matrix [35]. A transverse crack repeatedly deflects into the weak interface
146 between fibre and matrix, so the fracture surface is a sequence of pulled out fibre
147 bundles rather than a clean cut [36]. We model each break as a per column offset,
148 taken as band limited noise whose correlation length is the bundle spacing, scaled by
149 a Gamma distributed pull out length with shape below 2. That gives the heavy tail
150 which produces occasional long tongues, and an explicit dominant tongue is added
151 with probability 0.45. Material is lost at a break with probability 0.35, and when it is,
152 the two pieces belong to the same slip but no longer physically join. Whole fragments
153 are lost with probability 0.12. The evaluation respects both distinctions throughout.

154

Recovered pieces receive edge erosion, staining, ink fading, a photographic tilt and
155 a gentle bow from the curvature of the culm, following the damage reported for buried
156 bamboo strips and for waterlogged organic material generally [2, 3]. Each of the two
157 faces of a break independently loses its own sliver of material and carries its own
158 staining. This matters more than any other choice in the generator. If the two faces are
159 treated as one curve recorded twice, a network learns to match them almost perfectly
160 and the problem disappears, so we draw the loss and the staining independently for the
161 two faces. What that destroys cannot be recovered by any matcher, however trained.

162

Three co varying physical quantities set the difficulty: the relief the break carries,
163 the material lost independently from each face, and the independent staining of each
164 face. Five states, P1 to P5, span the range. Every result is reported as a function of
165 position on this ladder, and wherever possible against the matcher’s own measured
166 Top- k accuracy, so that the findings can be read against any published matcher rather
167 than against our generator’s parameters.

168

169 2.3 Pairwise matching

170

Fragments are rectified on their two long machined edges by a trimmed least squares
171 fit, which is the cue a conservator uses. Each fracture end is then described by the
172 silhouette of the break and by the texture strip immediately inside it, both resampled
173 to a canonical width, with the column support measured at that end of the piece so
174 that residual tilt cannot smear the descriptor.

175

A shared encoder embeds each fracture end into a unit vector, and the compatibility
176 of two ends is the inner product of their embeddings. Scoring a corpus is then a single
177 matrix product, which is what makes corpora of five thousand fragments tractable,
178 since a cross encoder over all ordered pairs would be twenty five million forward passes.
179 The encoder takes two channels, the texture strip and the standardised silhouette
180 profile broadcast down the height, so that profile and texture are registered column
181 by column. It collapses height by strided convolution while keeping the width axis
182 until late, because the evidence for a join is column registered. Training uses in batch

183

184

InfoNCE [37] over true joining pairs with a learned temperature, and photometric and sub pixel augmentation are applied independently to the two sides of each pair.

Two choices protect the later analysis. The matcher never sees slip width, binding notch heights, scribal hand or excavation context, because those are the evidence tested in the ablation and withholding them here keeps that test free of leakage. Training corpora are drawn from a seed range disjoint from every corpus used in evaluation, with preservation randomised over the whole ladder, so that the matcher is not out of distribution on the poorly preserved corpora. Otherwise the global stage would be credited with beating a straw man.

2.4 Calibration

Raw scores from a discriminative network are not probabilities, and a decision rule built on them is only as good as the mapping used [38, 39]. Raw scores are therefore mapped to log odds by Platt scaling fitted on dedicated calibration corpora with known ground truth, at balanced class ratio, so that the result is the likelihood ratio of the score under match against non match and carries no corpus prior. The prior belongs in the structure rather than in the weight of a candidate join. The uniqueness requirement already encodes that an end has at most one partner, and the residual reluctance to claim any join is a single scalar threshold θ .

The threshold is selected on calibration corpora from a grid of posterior confidences from 0.30 to 0.9997, converted to logits. Posterior confidence rather than raw logits is used because the fitted calibration slope varies by almost an order of magnitude across the preservation ladder, since a well preserved corpus separates matches from non matches so sharply that the logits are large. A grid fixed in logit units would mean something different at each rung, and would silently truncate the search for whichever method prefers a high threshold. The threshold is tuned separately for every method, so each is reported at its own best operating point. Nothing is fitted on evaluation corpora.

2.5 The assembly model

Reassembly is stated as one optimisation over the whole corpus. Let x_{ij} be a binary decision that is one when the lower break of fragment i joins the upper break of fragment j , taken over a candidate set A of the top k partners per fracture end, and let w_{ij} be the calibrated log odds of that join less the threshold θ . We maximise the total weight of the accepted joins,

$$\max \sum_{(i,j) \in A} w_{ij} x_{ij}, \quad (1)$$

subject to the conditions that describe what a slip physically is.

Each fracture end joins at most one other end, since a break has two faces and no more. A slip is a chain and never a cycle or a branch. To impose this, let π_i be the distance from the top of the reconstructed slip to the top of fragment i , and h_i the

231 height of that fragment. A selected join sets

232

$$233 \quad x_{ij} = 1 \Rightarrow \pi_j = \pi_i + h_i. \quad (2)$$

234

235 Because an accepted join forces this position to increase by the height of the preceding
 236 fragment, a single family of variables enforces the linear order, forbids cycles and
 237 branches, and caps the assembled length, all at once.

238 Two further conditions are simple gates on which joins may be considered. Frag-
 239 ments whose observed widths differ by more than a tolerance cannot join, and
 240 fragments from incompatible excavation contexts cannot join. Attribution of the
 241 scribal hand is noisy, so it enters the objective rather than the constraints, as the
 242 log likelihood ratio for same slip membership given agreement or disagreement of the
 243 attribution.

244

245 2.6 Latent slip properties

246

247 Slip length, slip width and scribal hand describe a slip, not a pair of fragments. For
 248 each fragment i we therefore introduce the length Λ_i , the width Ω_i and the hand η_i
 249 of the slip it belongs to, and require an accepted join to carry them along the chain,

250

$$251 \quad x_{ij} = 1 \Rightarrow \Lambda_j = \Lambda_i, \quad \Omega_j = \Omega_i, \quad \eta_j = \eta_i. \quad (3)$$

252

253 Three consequences follow, and together they are the methodological contribution
 254 of this work. First, a reconstructed chain must fit inside *its own* slip, rather than
 255 merely inside the longest slip in the corpus. Second, a notch observed at local height
 256 ν on fragment i sits at absolute height $\pi_i + \nu$ within the slip, and can only lie where
 257 a cord ran, so

$$258 \quad |\pi_i + \nu - f_c \Lambda_i| \leq \tau_\nu \quad \text{for one cord fraction } f_c, \quad (4)$$

259

260 which is linear in the latent length and ties the notches on *different* fragments of one
 261 chain to a single shared value. That coupling is what anchoring notches to corpus
 262 wide windows discards. This is also absolute positional evidence, available only once
 263 a fragment has been given a position, which is why no pairwise scorer can use it.
 264 Third, the hand becomes a latent category of which the observed attributions are noisy
 265 readings, so each fragment contributes evidence when its observation agrees with the
 266 hand of its chain, instead of a bonus applied once per join, which double counts the
 same weak evidence along a chain.

267

268 The latent width also removes a drift that a pairwise gate permits. With a toler-
 269 ance applied between neighbours, a four fragment chain can drift by three times that
 tolerance from end to end, whereas one shared width cannot.

270

271 A notch that cannot be placed in any window is charged a penalty rather than
 272 rendering the instance infeasible. Cord jitter and the corpus's own length variation put
 273 the occasional genuine notch outside every window. Imposed as a hard requirement
 274 this makes the whole corpus unreconstructable, and in a sensitivity sweep over the
 275 tolerance it would misattribute that brittleness to the information content of the
 constraint.

276

The model is solved with the CP-SAT constraint programming solver [40], distributed in Google OR-Tools (<https://developers.google.com/optimization>), and we report whether the returned solution is proven optimal. Hyperparameters are listed in Supplementary Table 1. Only the threshold θ is tuned, and it is tuned on calibration corpora disjoint from every evaluation corpus.

2.7 Compared methods

Every method compared here is a constraint subset of the same model, solved by the same solver on the same candidate set, so differences follow from the evidence used and not from the implementation (Supplementary Table 2). The comparison of interest is maximum weight bipartite matching, which is the standard way to resolve conflicts between candidates and is the strongest alternative that does not model the slip. It is Eq. (1) under the uniqueness requirement alone. A dense Hungarian solver [41] would be equivalent but is cubic in the fragment count and cannot reach these corpus sizes. We also report taking the top ranked candidate for each fragment, which is the naive use of a ranker, the mutual best heuristic, and an oracle specialist who finds the correct partner whenever it appears in the top k of a list and never errs. That oracle is an upper bound on any workflow whose output is a ranked list of joins, and it is charged its true cost, the number of candidate pairs a specialist must inspect.

2.8 Evaluation

Three questions are asked, and they are not interchangeable. Whether individual joins are right, measured by precision, recall and F_1 over true joins. Whether the slips are right, measured by the adjusted Rand index [42] of the induced partition of fragments into slips. Whether the output is usable as it stands, measured by the fraction of true multi fragment slips recovered exactly end to end.

We report separately the rate of joins that link fragments of two different slips, because the two kinds of error differ in cost to a conservator. Linking two fragments of one slip that no longer physically join misplaces an adjacency but leaves the fragment in the right slip, and little is lost. Joining fragments of two different slips corrupts two reconstructions at once and must be found and undone.

Corpora contain 1151 fragments unless stated otherwise, each cell is repeated over five seeds, and we report means with standard deviations. Instances are deterministic given a seed, and training, calibration and evaluation corpora come from disjoint seed ranges. Nothing is fitted on the evaluation corpora.

2.9 Sensitivity analysis

Sobol total order indices [43, 44] were estimated over ten parameters, comprising eight properties of burial and observation and the two tolerances of the method, from 384 model evaluations on corpora of 120 slips, using SALib [45]. Confidence intervals are reported at 95%.

3 Results

3.1 The simulated corpora reproduce published difficulty

A benchmark is only useful if it is as hard as the material, so we first place the preservation ladder against a published result (Fig. 2).

A single matcher, trained once, reaches Top-1 accuracy of 0.930, 0.896, 0.764, 0.502 and 0.237 across the five states. At the worst state, on a corpus of 5131 fragments, it reaches Top-50 accuracy of 0.565. This is comparable to the 0.52 reported for WisePanda among more than one thousand candidates [4], on a corpus about five times smaller. States P4 and P5 therefore bracket the difficulty that current systems face, and we treat them as the reference regime. Supplementary Fig. 1 confirms that the scores are informative and correctly calibrated.

3.2 Reassembly under constraints recovers whole slips

Table 1 and Fig. 3 give the comparison across preservation. At P4 the constrained model reaches join F_1 of 0.678 against 0.356 for bipartite matching, a slip grouping index of 0.662 against 0.189, and 37.8% of multi fragment slips recovered intact against 4.6%. At P5 bipartite matching collapses to a grouping index of 0.001, which is no better than chance, while the constrained model holds at 0.454.

Errors also change in kind, and the distinction matters to a conservator. Taking the top ranked candidate produces joins that fuse two different slips at rates of 44.5% to 86.3%. Bipartite matching gives 12.9% to 86.8%. The constrained model gives 1.7% to 29.9%, a fivefold reduction at P4. Fig. 4 shows what these outcomes look like on individual reconstructions.

3.3 Excavation context is the most valuable evidence

Removing each source of evidence in turn gives a clear ordering (Fig. 5, Table 2).

Excavation context is the most valuable single source. Removing it costs 0.137 of grouping index at P4 and 0.156 at P3, and it more than triples the rate of joins that fuse two slips, from 0.104 to 0.368. No other source comes close.

Slip length and slip width follow. Each is worth about 0.02 of grouping index at P4, and considerably more in error rate. Scribal hand is worth 0.037 of grouping index, but only when modelled correctly, as the next section shows.

Binding notch position is the weakest source. Even when modelled well it adds 0.006 of grouping index, although it does reduce slip fusing joins by about 15%. Cords sit at fractions of the slip length, and those lengths are known only within the range of the corpus, which is 231 to 278 mm, so the admissible band around each cord is several millimetres wide and the constraint rarely binds.

3.4 Properties of the slip must be modelled as such

Slip length, width and scribal hand could instead be applied to pairs of fragments, which is the natural thing to do when the evidence arrives as pairwise scores. We tested both. As pairwise features, notch position added almost nothing and the scribal

hand actively lowered accuracy at P4, where removing it improved the grouping index from 0.547 to 0.575. Carried as properties of the slip, shared along the reconstructed chain, the same evidence raises the grouping index by 0.115 at P4, 0.077 at P2 and 0.045 at P3, and the hand term becomes useful rather than harmful.

The difference is consistent across seeds. The latent formulation wins on all five seeds at P2, P3 and P4, with paired t tests giving $p = 0.0004$, $p = 0.0009$ and $p < 0.0001$ (Supplementary Fig. 2). Compared over the full range of thresholds, it delivers the same grouping quality with about half the slip fusing errors (Fig. 6).

3.5 Reassembly of a Shuihudi scale corpus

We ran the whole procedure on a corpus configured to the Shuihudi find, with 1155 slips of the reported dimensions and three binding cords, at preservation state P4. This produced 3253 fragments and 1364 true joins (Fig. 7). The matcher reached Top-1 accuracy of 0.398.

Bipartite matching returned join F_1 of 0.286 and a grouping index of 0.082. It recovered 0.5% of multi fragment slips intact and made 2122 joins that fuse two different slips. The constrained model returned join F_1 of 0.621 and a grouping index of 0.576. It recovered 33.4% of slips intact and made 380 such joins, a reduction by a factor of 5.6 in the errors a specialist would have to find and undo. At this size the solver reaches a feasible but not proven optimal solution within its budget, so these values are a lower bound on what the model delivers.

3.6 Constraints substitute for matching accuracy

The gain from constraints is not constant. It grows as preservation falls (Fig. 8), which inverts the natural expectation that a global stage is a refinement to add once matching is good enough.

The effect can be priced. At P5 the constrained model, working from a Top-1 accuracy of 0.237, matches the grouping quality that bipartite matching would require a Top-1 accuracy of 0.806 to reach (Table 3). The constraints are worth 0.57 of matching accuracy. At P1 to P4 the question has no answer, because bipartite matching does not reach the constrained grouping index at any matching accuracy, including that of a perfect matcher.

The same comparison can be made in units of specialist time. At P5 the constrained model, with no human review at all, reaches the grouping quality that an oracle specialist working down a ranked list would reach only after inspecting about 39 000 candidate pairs (Supplementary Fig. 3).

3.7 Robustness, scale and cost

Results hold across corpus size. From 425 to 5060 fragments, bipartite matching falls from a grouping index of 0.305 to 0.065 and from 13.6% to 0.3% intact slips, while the constrained model falls from 0.732 to 0.491 and from 43.3% to 28.4% (Fig. 9 and Supplementary Table 3). The gap widens with size, because a larger corpus gives each fragment more chances to be captured by a plausible impostor, and structural constraints are what reject impostors that a score alone cannot.

415 A Sobol analysis over ten parameters shows that the two tolerances of the method
416 are the least influential inputs, with total order indices of 0.029 for the notch window
417 and 0.087 for the width gate (Supplementary Fig. 4). What dominates is the condition
418 of the corpus, led by the independent staining of the two fracture faces at 0.435. The
419 results are therefore driven by preservation and not by the tuning of the method.

420 The optimisation is solved to proven optimality at 429 and 1200 fragments, in 3.2
421 s and 69.7 s. At 2572 fragments the optimality gap is about 2% when a 240 s budget
422 expires (Supplementary Fig. 5).

423

424 4 Discussion

425

426 Of the five sources of evidence tested, the one that matters most is the context a frag-
427 ment was excavated from, and it is the one that no published reassembly system uses.
428 Matching networks are trained on images of breaks. Excavation context is recorded as
429 a matter of routine, costs nothing further to obtain, and in our corpora is worth more
430 than slip dimensions, notch position and scribal hand combined. It is being discarded
431 at exactly the step where it is most informative.

432 This has a direct implication for excavation and conservation practice. Where slips
433 are lifted in a block and separated in the laboratory, which is the usual treatment
434 for waterlogged bamboo, the position of each fragment within the block should be
435 recorded as it is separated and kept in a form that a reassembly system can read. A
436 coarse grid reference is enough, since the constraint only has to exclude fragments that
437 cannot have lain together. Where that information already exists in older excavation
438 archives, it is worth recovering before reassembly is attempted. How much it is worth
439 will depend on how finely a site is subdivided and how far the fragments of one slip
440 disperse, and both vary between sites. The general point, that provenance information
441 is informative at the step where it is currently dropped, is unlikely to be specific to
442 bamboo slips, since fresco, ceramic, papyrus and oracle bone reassembly systems are
443 built on the same pairwise foundation [10, 15, 16, 18, 20, 23], and the benchmarks
444 now being built for them score the assembled object rather than the pair [31].

445 The second finding is that constraints buy something a better matcher cannot. The
446 gain from the constrained model grows as preservation falls, and at P5 it is worth 0.57
447 of Top-1 matching accuracy. At better preservation the exchange has no rate at all,
448 because bipartite matching never reaches the constrained grouping index even with a
449 perfect matcher. That ceiling is structural rather than an artefact of extrapolation.
450 A pairwise matcher can only propose joins whose fracture surfaces still exist, and in
451 these corpora material is lost at about a third of all breaks. Width, notch position
452 and hand still place such a fragment in its slip. Better matching and better structural
453 knowledge are therefore not interchangeable, and past a point the second provides what
454 the first cannot. For a laboratory choosing where to invest, this argues for recording
455 the physical and contextual attributes of every fragment before investing further in
456 image based matching.

457 The third finding concerns how the evidence is encoded, and it is the part we expect
458 to transfer furthest. Slip length, width and scribal hand describe a slip. Expressed as
459 pairwise features, which is the natural move when the evidence arrives as a matrix
460

of pairwise scores, two of the three contributed almost nothing and the third made results worse. The evidence was not weak, the encoding was. Represented as latent properties shared along each reconstructed slip, the same three attributes moved the whole operating curve and halved the rate of the errors that fuse two slips. The general statement is that a property of the object being rebuilt belongs in the model of that object, not in the weight of an edge between two of its pieces. Manuscript leaves, ceramics and wall paintings each have such properties, a leaf size, a vessel profile, a plaster recipe, and identifying them is the substance of applying this approach elsewhere.

Three limits are visible in the results themselves. The binding notch constraint contributes least, because cord positions are fractions of a slip length that is known only to within the range of the corpus, so the admissible band is several millimetres wide and the constraint rarely binds. A corpus of uniform slip length, or a joint estimate of length across slips from one roll, would make it much stronger. At the worst preservation state the model still recovers only 22.6% of slips intact, because when the matcher provides almost no signal the constraints can reject impossible configurations but cannot supply the correct one. Above about 2500 fragments the solver returns feasible rather than proven optimal solutions within our budget, so the largest corpora are reported as lower bounds.

The evaluation is entirely simulated. This is deliberate, because a reconstruction metric needs complete ground truth and a real corpus is exactly where that is missing, and it follows the practice of the work we build on, which also trains from generated pairs [4]. A generator nonetheless encodes the assumptions of its author, and ours is calibrated against a single published accuracy figure. We therefore report all results as a function of the measured accuracy of the matcher rather than of our own parameters, and we randomise preservation during training so that no comparison method is weakened artificially. A real corpus would still test assumptions that no internal validation can reach. The constraints also assume a linear object of bounded length and constant cross section, so the approach transfers to other fragmented material but these particular constraints do not.

Two steps follow directly. The first is to apply the procedure to the output of a published matcher on a real corpus, which needs only a compatibility matrix and the excavation records, and which would test the assumptions of the generator against material. The second is to record and keep fragment level context in current excavations of slip material, since our results indicate that it is worth more than any other single source we tested.

Data availability. All corpora used in this study are simulated and are produced deterministically by the released generator from the seeds recorded in the experiment scripts, so any corpus can be regenerated exactly. Where the solver is given a wall clock budget and returns a feasible rather than a proven optimal solution, which is the case above roughly 2500 fragments and for the Shuihudi scale corpus, the selected solution depends on the parallel search and repeated runs differ by about one point in the reported measures. Solutions proven optimal do not vary. The result tables underlying all figures are available at <https://github.com/Yeung8023/bamboo-slip-reassembly>.

Code availability. The benchmark generator, the matching network, the assembly model and all analysis scripts are available at <https://github.com/Yeung8023/bamboo-slip-reassembly>.

Acknowledgements. This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors.

Author contributions. J.Y. conceived the study, designed and implemented the benchmark generator, the pairwise matching network and the constrained assembly model, ran all experiments and the statistical analysis, prepared Figures 1 to 9, Tables 1 to 3 and the Supplementary Information, and wrote the main manuscript text. W.Y. supervised the project, contributed to the design of the study and to the interpretation of the results, and revised the manuscript. All authors reviewed the manuscript.

Competing interests. The authors declare no competing financial or non-financial interests.

Supplementary information. This article is accompanied by Supplementary Information containing Supplementary Figures 1 to 5 and Supplementary Tables 1 to 3.

References

- [1] Editorial Group for the Shuihudi Qin Tomb Bamboo Slips. *Bamboo Slips from the Qin Tomb at Shuihudi [Shuihudi Qin mu zhujian, 睡虎地秦墓竹简]* (in Chinese) (Wenwu Press, Beijing, 1990). Tomb 11, Yunmeng, Hubei, excavated December 1975; 1155 slips and 80 fragments, 23.1 to 27.8 cm long and 0.5 to 0.8 cm wide.
- [2] Li, M.-Y. *et al.* Investigation into the deterioration process of archaeological bamboo strips of China from four different periods by chemical and anatomical analysis. *Polymer Degradation and Stability* **109**, 71–78 (2014).
- [3] Broda, M. & Hill, C. A. S. Conservation of waterlogged wood—past, present and future perspectives. *Forests* **12**, 1193 (2021).
- [4] Zhu, J. *et al.* Rejoining fragmented ancient bamboo slips with physics-driven deep learning. *Nature Communications* **17**, 3550 (2026).
- [5] Zhang, Y. *et al.* LLMCO4MR: LLMs-aided neural combinatorial optimization for ancient manuscript restoration from fragments with case studies on Dunhuang. In *Computer Vision: ECCV 2024*, Vol. 15133 of *Lecture Notes in Computer Science*, 253–269 (Springer, Cham, 2024).
- [6] Zheng, Y. *et al.* Dunhuang manuscript fragment reassembly based on patch-level handwriting style recognition. *npj Heritage Science* **13**, 507 (2025).
- [7] Cao, S., Pan, T., Wang, Y. & Song, T. Character restoration of Qin and Han bamboo slips based on improved conditional generative adversarial networks. *npj Heritage Science* **13**, 278 (2025).

- [8] Zhou, S. *et al.* Semi-supervised restoration of damaged characters in Liye Qin bamboo slips. *npj Heritage Science* (2026). Advance online publication, doi:10.1038/s40494-026-02785-0. 553
554
555
556
- [9] Zhang, Z., Guo, A. & Li, B. Internal similarity network for rejoining oracle bone fragment images. *Symmetry* **14**, 1464 (2022). 557
558
559
- [10] Zhang, Z. *et al.* Deep rejoining model and dataset of oracle bone fragment images. *npj Heritage Science* **13**, 66 (2025). 560
561
562
- [11] Guan, H. *et al.* Deciphering oracle bone language with diffusion models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15554–15567 (Association for Computational Linguistics, 2024). 563
564
565
566
- [12] Assael, Y., Sommerschild, T. & Prag, J. *Restoring ancient text using deep learning: a case study on Greek epigraphy.* In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 6367–6374 (Association for Computational Linguistics, 2019). 567
568
569
570
571
572
- [13] Assael, Y. *et al.* Restoring and attributing ancient texts using deep neural networks. *Nature* **603**, 280–283 (2022). 573
574
575
- [14] Sommerschild, T. *et al.* Machine learning for ancient languages: a survey. *Computational Linguistics* **49**, 703–747 (2023). 576
577
578
- [15] Brown, B. J. *et al.* A system for high-volume acquisition and matching of fresco fragments. *ACM Transactions on Graphics* **27**, 1–9 (2008). 579
580
581
- [16] Toler-Franklin, C., Brown, B., Weyrich, T., Funkhouser, T. & Rusinkiewicz, S. Multi-feature matching of fresco fragments. *ACM Transactions on Graphics* **29**, 1–12 (2010). 582
583
584
- [17] Funkhouser, T. *et al.* Learning how to match fresco fragments. *Journal on Computing and Cultural Heritage* **4**, 1–13 (2011). 585
586
587
- [18] Oxholm, G. & Nishino, K. A flexible approach to reassembling thin artifacts of unknown geometry. *Journal of Cultural Heritage* **14**, 51–61 (2013). 588
589
590
- [19] Zhang, K., Yu, W., Manhein, M., Waggenspack, W. & Li, X. *3D fragment reassembly using integrated template guidance and fracture-region matching.* In *2015 IEEE International Conference on Computer Vision (ICCV)*, 2138–2146 (IEEE, 2015). 591
592
593
594
595
- [20] Zhang, Y., Li, K., Chen, X., Zhang, S. & Geng, G. A multi feature fusion method for reassembly of 3D cultural heritage artifacts. *Journal of Cultural Heritage* **33**, 596
597
598

191–200 (2018).

[21] Derech, N., Tal, A. & Shimshoni, I. Solving archaeological puzzles. *Pattern Recognition* **119**, 108065 (2021).

[22] Cao, Y., Fang, Z., Tian, H. & Wei, R. 2D irregular fragment reassembly with deep learning assistance. *IEEE Access* **12**, 28554–28563 (2024).

[23] Abitbol, R., Shimshoni, I. & Ben-Dov, J. Machine learning based assembly of fragments of ancient papyrus. *Journal on Computing and Cultural Heritage* **14**, 1–21 (2021).

[24] Pirrone, A., Beurton-Aimar, M. & Journet, N. Self-supervised deep metric learning for ancient papyrus fragments retrieval. *International Journal on Document Analysis and Recognition* **24**, 219–234 (2021).

[25] Ostertag, C. & Beurton-Aimar, M. Matching ostraca fragments using a siamese neural network. *Pattern Recognition Letters* **131**, 336–340 (2020).

[26] Pintus, R. *et al.* A survey of geometric analysis in cultural heritage. *Computer Graphics Forum* **35**, 4–31 (2016).

[27] Casillo, M., Colace, F., Gaeta, R., Lorusso, A. & Pellegrino, M. Artificial intelligence in archaeological site conservation: trends, challenges, and future directions. *Journal of Computer Applications in Archaeology* **8**, 224–241 (2025).

[28] Harmanda, T. T., Hidayanto, A. N., Aji, R. F. & Ilin, I. Artificial intelligence adoption in the cultural heritage sector: a systematic review of trends, application domains, and challenges. *Computers in Human Behavior Reports* **23**, 101147 (2026).

[29] Son, K., Hays, J. & Cooper, D. B. Solving square jigsaw puzzle by hierarchical loop constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**, 2222–2235 (2019).

[30] Paumard, M.-M., Picard, D. & Tabia, H. Deepzzle: solving visual jigsaw puzzles with deep learning and shortest path optimization. *IEEE Transactions on Image Processing* **29**, 3569–3581 (2020).

[31] Tsesmelis, T. *et al.* *Re-assembling the past: the RePAIR dataset and benchmark for real world 2D and 3D puzzle solving.* In *Advances in Neural Information Processing Systems 37*, 30076–30105 (2024).

[32] Li, S. A study of the handwriting of the Tsinghua bamboo-slip manuscripts *Shifa and *Zi Chan. *Bamboo and Silk* **7**, 75–113 (2024).

[33] Brink, A. A., Smit, J., Bulacu, M. L. & Schomaker, L. R. B. Writer identification using directional ink-trace width measurements. *Pattern Recognition* **45**, 162–171

- (2012). 645
- [34] Popović, M., Dhali, M. A. & Schomaker, L. Artificial intelligence based writer 646
identification generates new evidence for the unknown scribes of the Dead Sea 647
Scrolls exemplified by the Great Isaiah Scroll (1QIsaa). *PLOS ONE* **16**, e0249769 648
(2021). 649
- [35] Amada, S., Ichikawa, Y., Munekata, T., Nagase, Y. & Shimizu, H. Fiber texture 650
and mechanical graded structure of bamboo. *Composites Part B: Engineering* 651
28, 13–20 (1997). 652
- [36] Habibi, M. K. & Lu, Y. Crack propagation in bamboo’s hierarchical cellular 653
structure. *Scientific Reports* **4**, 5598 (2014). 654
- [37] van den Oord, A., Li, Y. & Vinyals, O. Representation learning with contrastive 655
predictive coding. *arXiv* (2018). Preprint at arXiv:1807.03748. 656
- [38] Niculescu-Mizil, A. & Caruana, R. *Predicting good probabilities with supervised* 657
learning. In *Proceedings of the 22nd International Conference on Machine* 658
Learning (ICML), 625–632 (ACM, 2005). 659
- [39] Guo, C., Pleiss, G., Sun, Y. & Weinberger, K. Q. On calibration of modern neural 660
networks. *arXiv* (2017). Preprint at arXiv:1706.04599. 661
- [40] Perron, L., Didier, F. & Gay, S. *The CP-SAT-LP solver*. In *29th International* 662
Conference on Principles and Practice of Constraint Programming (CP 2023), 663
Vol. 280 of *Leibniz International Proceedings in Informatics (LIPIcs)*, 3:1–3:2 664
(Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2023). 665
- [41] Kuhn, H. W. The Hungarian method for the assignment problem. *Naval Research* 666
Logistics Quarterly **2**, 83–97 (1955). 667
- [42] Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 668
193–218 (1985). 669
- [43] Sobol’, I. M. Global sensitivity indices for nonlinear mathematical models and 670
their Monte Carlo estimates. *Mathematics and Computers in Simulation* **55**, 671
271–280 (2001). 672
- [44] Saltelli, A. *et al.* Variance based sensitivity analysis of model output. design and 673
estimator for the total sensitivity index. *Computer Physics Communications* **181**, 674
259–270 (2010). 675
- [45] Herman, J. & Usher, W. SALib: an open-source Python library for sensitivity 676
analysis. *Journal of Open Source Software* **2**, 97 (2017). 677

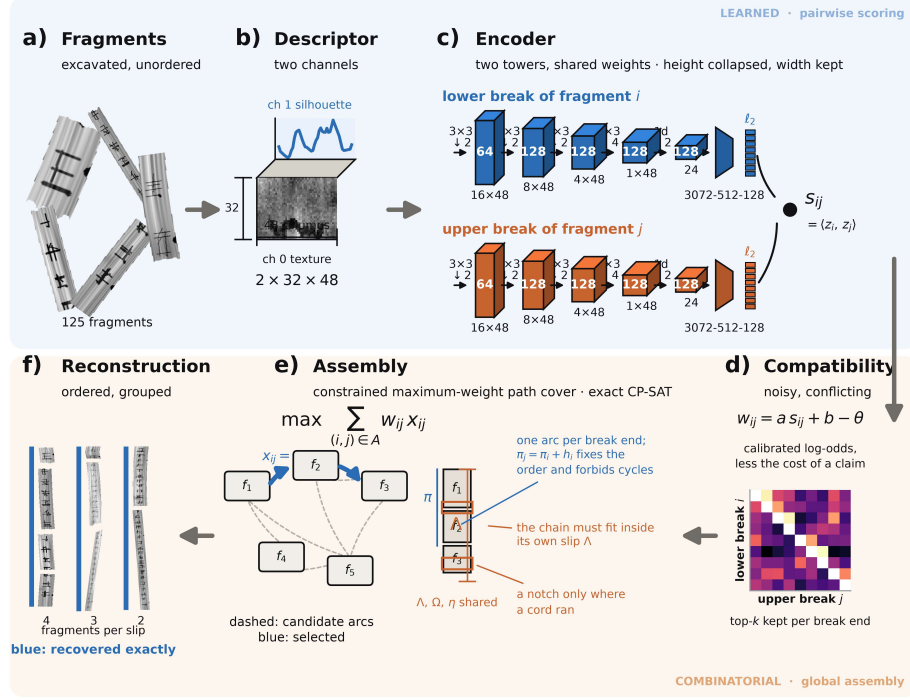


Fig. 1 The reassembly procedure. a) Fragments arrive unordered. b) Each fracture end becomes a two channel descriptor, the silhouette of the break and the texture just inside it. c) A two tower encoder embeds the two ends. d) Inner products give a compatibility matrix that is noisy and self contradictory. e) Reassembly is a constrained optimisation over the whole corpus. One position variable per fragment fixes order, forbids cycles and bounds length, and slip length, width and hand are carried as properties of the slip. f) The output is a set of ordered chains. All content shown is real. The fragments come from the generator, the compatibility block is the output of the trained matcher, and the assembled slips are the solver solution on that corpus. In panel f slip width is exaggerated, since at true proportions a 23 cm slip of 6 mm width renders as a hairline, while proportions within each slip are exact.

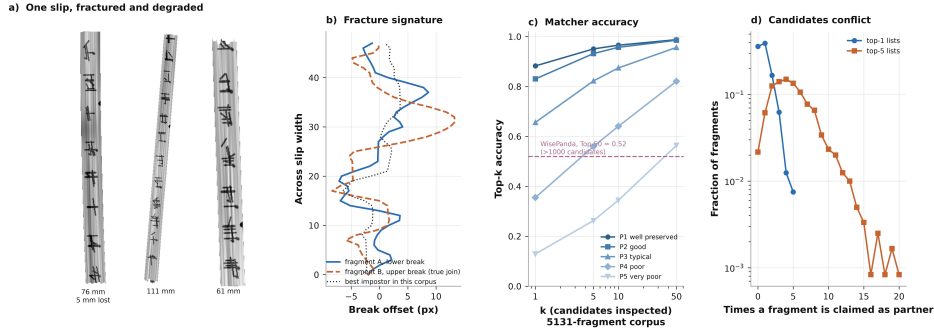


Fig. 2 The simulated corpora and the matcher. a) One slip after fracture and burial damage. Material lost at a break means that adjacent fragments of one slip need not physically join. b) The silhouette of a true join against the closest impostor in the same corpus. The blue solid line is the lower break of fragment A, the orange dashed line is the upper break of fragment B, which is the true join, and the grey dotted line is the best impostor. c) Top- k matching accuracy along the preservation ladder on 5131 fragments. The pink dashed line is the accuracy reported for WisePanda among more than one thousand candidates [4]. d) Candidates conflict. The distribution of the number of times each fragment is proposed as a partner by others.

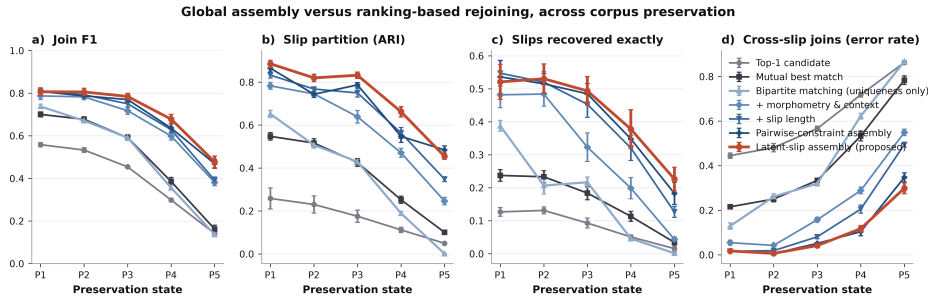


Fig. 3 Reassembly quality across the preservation ladder. Corpora of 1151 fragments, five seeds. Error bars are standard deviations over seeds. Each method is a constraint subset of the same model, solved by the same solver on the same candidate set. a) Join F_1 . b) Slip partition index. c) Fraction of multi fragment slips recovered exactly. d) Rate of joins that fuse two different slips, where lower is better.

What success and failure look like

Each column is one reconstructed chain. The true slip id of every fragment is given at its right. A correct chain has one id throughout; a fused chain does not.

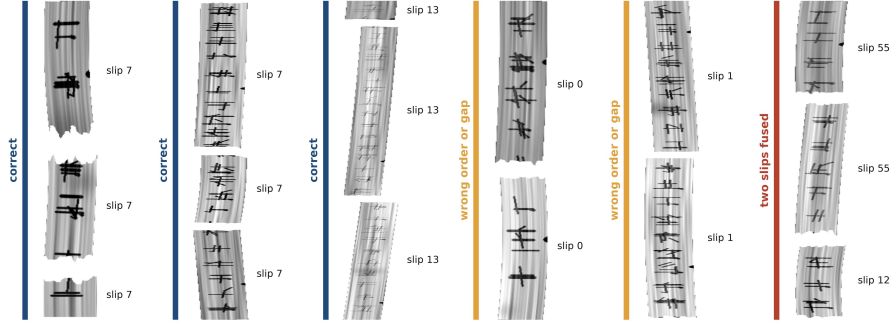


Fig. 4 What success and failure look like. Each column is one reconstructed chain from a corpus at preservation state P4, with the true slip identifier of every fragment shown at its right, and a coloured bar at its left giving the outcome. A correct chain carries one identifier throughout. The three columns marked in dark blue are recovered exactly. The two marked in amber keep one identifier but have the order or the gap wrong, which is a mild error. The two marked in red mix two identifiers, which is the damaging error.

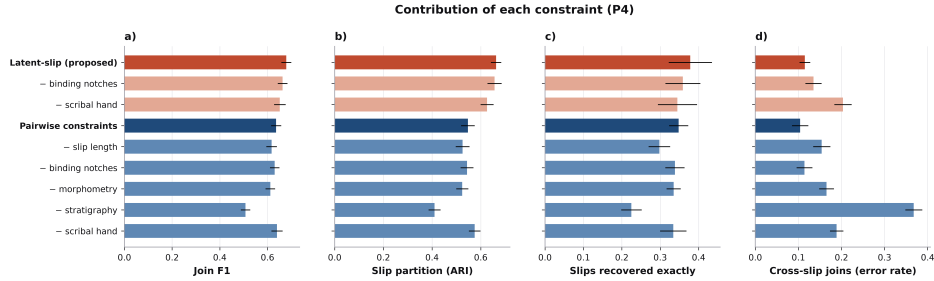


Fig. 5 What each source of evidence is worth, measured by removing it. Preservation state P4, five seeds, error bars are standard deviations. The upper group in orange carries slip properties as properties of the slip, the lower group in blue carries them as pairwise features. Excavation context, shown as stratigraphy, is the most valuable single source. Scribal hand helps only in the upper group, and in the lower group removing it improves the result. a) Join F_1 . b) Slip partition index. c) Fraction of slips recovered exactly. d) Rate of joins that fuse two slips, where lower is better.

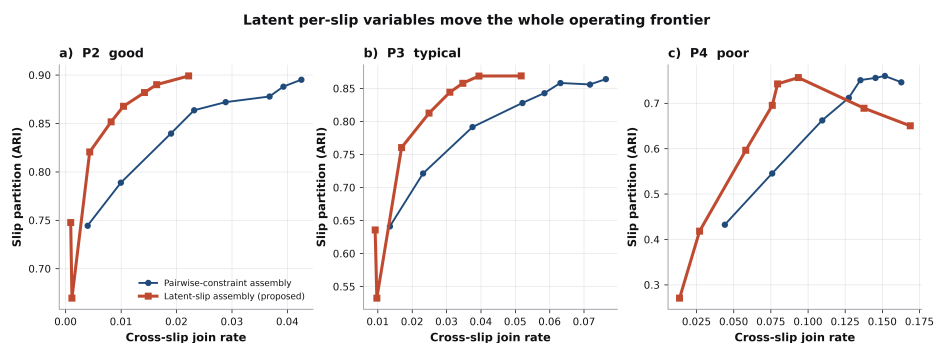


Fig. 6 Carrying slip properties as properties of the slip moves the whole operating curve. Each point is one value of the join threshold, and up and to the left is better. Panels a), b) and c) are preservation states P2, P3 and P4. At equal partition quality the latent formulation makes about half as many joins that fuse two slips.

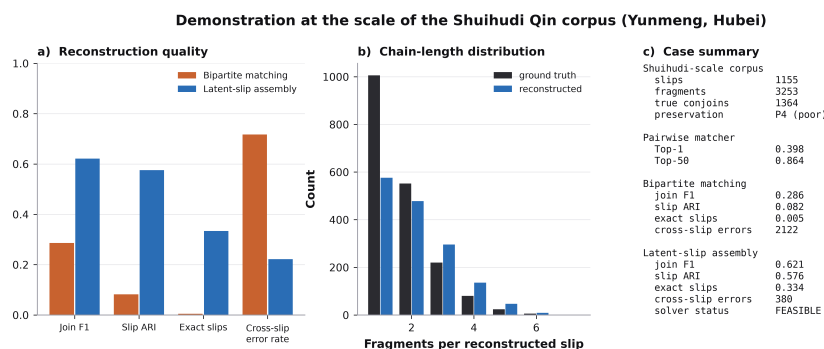


Fig. 7 Reassembly of a corpus scaled to the Shuihudi Qin find. The corpus has 1155 slips of 23.1 to 27.8 cm with three binding cords, at preservation state P4, giving 3253 fragments and 1364 true joins. a) Reconstruction quality. b) Distribution of reconstructed chain lengths against the truth. c) Summary of the case.

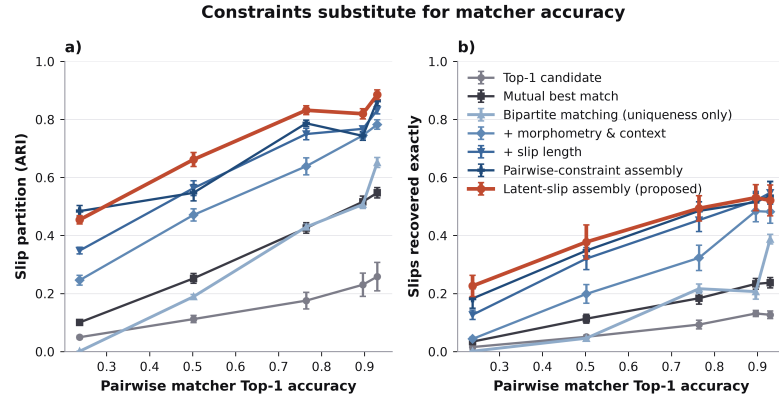


Fig. 8 Constraints substitute for matching accuracy. The results of Fig. 3 plotted against the measured Top-1 accuracy of the matcher, so that they can be read against any published matcher rather than against the parameters of our generator. a) Slip partition index. b) Fraction of multi fragment slips recovered exactly. The gap between the constrained model and bipartite matching widens as matching accuracy falls. Error bars are standard deviations over five seeds.

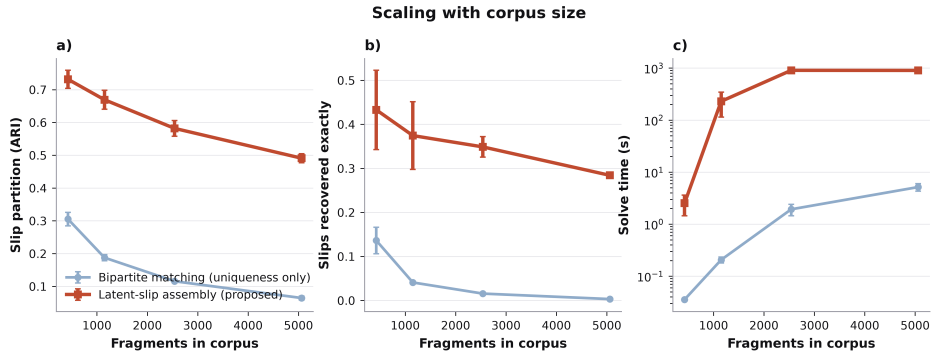


Fig. 9 Behaviour with corpus size at preservation state P4. Three seeds per size, and error bars are standard deviations over seeds. a) Slip partition index. b) Fraction of multi fragment slips recovered exactly. c) Solve time, on a logarithmic axis. The gap between the two methods widens with corpus size. At the two largest sizes the constrained model reaches a feasible but not proven optimal solution within a 900 s budget.

Table 1 Reconstruction quality across the preservation ladder. Corpora of 1,151 fragments; mean over five seeds (standard deviation in parentheses). Bipartite matching is the model under the uniqueness constraint alone. Expert review is an oracle specialist working down a ranked list. Cross-slip rate is the fraction of accepted joins that link fragments of two different slips.

State	Method	Join F_1	Slip ARI	Exact slips	Cross-slip
P1	Top-1 candidate	0.558 (0.010)	0.258 (0.049)	0.127 (0.013)	0.445 (0.012)
	Mutual best match	0.701 (0.012)	0.547 (0.018)	0.237 (0.018)	0.216 (0.009)
	Bipartite matching	0.738 (0.011)	0.652 (0.017)	0.388 (0.016)	0.129 (0.014)
	+ morphometry & context	0.787 (0.017)	0.783 (0.016)	0.482 (0.039)	0.054 (0.011)
	+ slip length	0.807 (0.017)	0.832 (0.013)	0.547 (0.039)	0.016 (0.003)
	Pairwise-constraint assembly	0.808 (0.017)	0.865 (0.017)	0.536 (0.049)	0.016 (0.004)
	Latent-slip assembly (ours)	0.807 (0.017)	0.885 (0.016)	0.521 (0.053)	0.017 (0.003)
	Expert review, top-10	0.993 (0.002)	0.661 (0.035)	0.977 (0.005)	0.000 (0.000)
	Expert review, top-50	0.999 (0.002)	0.668 (0.034)	0.995 (0.006)	0.000 (0.000)
P2	Top-1 candidate	0.533 (0.011)	0.230 (0.040)	0.131 (0.011)	0.481 (0.019)
	Mutual best match	0.677 (0.012)	0.516 (0.020)	0.233 (0.018)	0.250 (0.013)
	Bipartite matching	0.671 (0.010)	0.507 (0.014)	0.207 (0.026)	0.261 (0.013)
	+ morphometry & context	0.783 (0.013)	0.745 (0.010)	0.484 (0.037)	0.042 (0.004)
	+ slip length	0.792 (0.013)	0.768 (0.010)	0.520 (0.032)	0.019 (0.004)
	Pairwise-constraint assembly	0.787 (0.014)	0.743 (0.015)	0.515 (0.031)	0.006 (0.002)
	Latent-slip assembly (ours)	0.806 (0.017)	0.820 (0.017)	0.531 (0.045)	0.007 (0.002)
	Expert review, top-10	0.992 (0.003)	0.660 (0.032)	0.976 (0.011)	0.000 (0.000)
	Expert review, top-50	0.999 (0.001)	0.671 (0.035)	0.998 (0.002)	0.000 (0.000)
P3	Top-1 candidate	0.454 (0.007)	0.176 (0.029)	0.093 (0.015)	0.567 (0.011)
	Mutual best match	0.592 (0.013)	0.426 (0.018)	0.184 (0.020)	0.334 (0.009)
	Bipartite matching	0.591 (0.011)	0.429 (0.009)	0.217 (0.016)	0.319 (0.010)
	+ morphometry & context	0.718 (0.017)	0.639 (0.029)	0.323 (0.043)	0.157 (0.010)
	+ slip length	0.750 (0.017)	0.750 (0.020)	0.453 (0.040)	0.081 (0.009)
	Pairwise-constraint assembly	0.770 (0.015)	0.787 (0.011)	0.485 (0.031)	0.051 (0.005)
	Latent-slip assembly (ours)	0.785 (0.015)	0.832 (0.015)	0.494 (0.044)	0.041 (0.004)
	Expert review, top-10	0.973 (0.005)	0.633 (0.032)	0.919 (0.015)	0.000 (0.000)
	Expert review, top-50	0.995 (0.004)	0.666 (0.033)	0.985 (0.012)	0.000 (0.000)
P4	Top-1 candidate	0.298 (0.008)	0.112 (0.012)	0.051 (0.006)	0.718 (0.011)
	Mutual best match	0.385 (0.019)	0.252 (0.017)	0.113 (0.015)	0.533 (0.022)
	Bipartite matching	0.356 (0.011)	0.189 (0.007)	0.046 (0.008)	0.622 (0.013)
	+ morphometry & context	0.599 (0.019)	0.471 (0.021)	0.199 (0.032)	0.289 (0.015)
	+ slip length	0.629 (0.026)	0.563 (0.025)	0.320 (0.038)	0.205 (0.018)
	Pairwise-constraint assembly	0.636 (0.021)	0.547 (0.028)	0.348 (0.025)	0.104 (0.019)
	Latent-slip assembly (ours)	0.678 (0.022)	0.662 (0.024)	0.378 (0.059)	0.117 (0.014)
	Expert review, top-10	0.892 (0.013)	0.539 (0.025)	0.721 (0.025)	0.000 (0.000)
	Expert review, top-50	0.967 (0.004)	0.629 (0.032)	0.901 (0.015)	0.000 (0.000)
P5	Top-1 candidate	0.141 (0.013)	0.050 (0.006)	0.016 (0.008)	0.863 (0.008)
	Mutual best match	0.162 (0.018)	0.100 (0.010)	0.034 (0.016)	0.784 (0.019)
	Bipartite matching	0.136 (0.008)	0.001 (0.001)	0.001 (0.002)	0.868 (0.005)
	+ morphometry & context	0.381 (0.015)	0.246 (0.017)	0.044 (0.006)	0.549 (0.014)
	+ slip length	0.392 (0.015)	0.348 (0.012)	0.127 (0.016)	0.495 (0.011)
	Pairwise-constraint assembly	0.469 (0.022)	0.484 (0.019)	0.183 (0.033)	0.347 (0.021)
	Latent-slip assembly (ours)	0.477 (0.028)	0.454 (0.014)	0.226 (0.036)	0.299 (0.024)
	Expert review, top-10	0.693 (0.014)	0.354 (0.018)	0.403 (0.016)	0.000 (0.000)
	Expert review, top-50	0.877 (0.008)	0.522 (0.029)	0.691 (0.008)	0.000 (0.000)

967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012

Table 2 Leave-one-out over the constraint set at P4. Each row removes one constraint from the formulation named above it.

Formulation	Join F_1	Slip ARI	Exact slips	Cross-slip
Latent-slip (proposed)	0.679	0.663	0.378	0.115
– binding notches	0.663	0.657	0.359	0.135
– scribal hand	0.652	0.626	0.345	0.203
Pairwise constraints	0.636	0.547	0.348	0.104
– slip length	0.617	0.525	0.298	0.154
– binding notches	0.630	0.543	0.338	0.114
– morphometry	0.612	0.524	0.335	0.165
– stratigraphy	0.508	0.410	0.225	0.368
– scribal hand	0.640	0.575	0.334	0.189

Table 3 What the constraint set is worth, in units of matcher accuracy. For each state, the Top-1 accuracy that bipartite matching would require in order to reach the slip partition the proposed model attains at the accuracy it actually had.

State	Matcher Top-1	Proposed ARI	Equivalent Top-1
P5	0.237	0.454	0.806 (+0.569)
P4	0.502	0.662	<i>unreachable</i>
P3	0.764	0.832	<i>unreachable</i>
P2	0.896	0.820	<i>unreachable</i>
P1	0.930	0.885	<i>unreachable</i>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [02supplementary.zip](#)