WHAT DO COLLAPSE DIAGNOSTICS CERTIFY? TESTING ALIGNMENT OBJECTIVES FOR ORTHOGRAPHIC SHORTCUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Objectives that align representations without negatives are judged by label-free diagnostics of collapse, effective rank and uniformity, which also choose their hyperparameters. We ask what these diagnostics certify about what a representation has learned. In cross-lingual alignment from parallel text a model can match related languages by shared spelling rather than meaning; a substitution cipher on one side of every such pair removes the shared letters with the supervision held byte-identical, and the fraction of a model's score kept under the cipher is how much of its alignment was not spelling. The diagnostics certify spread, not content. (1) At the weights their effective rank selects, the two anti-collapse devices, VICReg and SimSiam, reach contrast's rank or beyond and keep 0.16 and 0.29 of their score against contrast's 0.65. (2) Tuning brings both within reach of contrast while the rank stays flat, and the spectrum every diagnostic reads does not move under the cipher. (3) Selecting by these diagnostics favours a shortcut solution, and on a standard pretrained encoder a run that learned nothing. The objectives that learn the lexicon share a target that cannot drift toward the shortcut, the batch in contrast or a frozen teacher in distillation: attraction to a frozen teacher learns the lexicon as contrast does, without negatives. Retrieval of the training pairs, which needs no labels, tracks held-out accuracy at $\rho = 0.96$ where rank selects the shortcut. Anti-collapse devices should be selected by a task-shaped criterion and tested with the shortcut removed, not certified by spread alone.

1 INTRODUCTION

Joint-embedding objectives that use no negatives must be kept from collapse, and whether they were is usually read off the representation itself. The alignment and uniformity of Wang & Isola (2020) and the effective rank (Roy & Vetterli, 2007; Garrido et al., 2023) need no labels, so they also serve to choose hyperparameters. These diagnostics say how far a representation spreads. They do not say along what.

We study a setting where that difference can be measured. In cross-lingual word alignment a model is given pairs of words that translate each other and must place them together. Related languages share spelling, so a model can satisfy part of the task by matching letters and a representation that spreads along spelling is not collapsed. Whether it has learned a lexicon is a separate question.

To separate the two we adapt the Fake-English construction of K et al. (2020). A substitution cipher rewrites one side of every Latin-script language pair, so that a word and its translation can no longer share letters. Everything else is held fixed, down to the bytes of the supervision files. A configuration trained on the *ciphered* corpus is compared with the same configuration trained on the *plain* one, and the ratio of the two scores says how much of its alignment was spelling (Figure 1).

The testbed is a byte-level encoder in which the alignment objective is the only pressure on a word's representation (Section 3). We train contrast and two *anti-collapse devices* that replace its negatives, VICReg (Bardes et al., 2022) and a SimSiam predictor (Chen & He, 2021), on identical translation links in 8 languages, on two corpora, and we repeat the comparison on a standard pretrained encoder at word level and at sentence level. Our contributions are the following.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

- *We show, by the Fake-English cipher of K et al. (2020), that the collapse diagnostics certify that a representation is spread, not what it has learned* (Sections 4 and 5).
- *We show that retrieval of the training links, a criterion shaped like the task, orders the settings where the diagnostics saturate and, with the shortcut removed, registers the failures they miss* (Section 6).
- *We show that the objectives that learn the lexicon share a target that cannot drift toward the shortcut, the batch in contrast or a frozen teacher in distillation, and that the shortcut is confined to objectives with neither* (Section 7).

2 Related work

The intervention. Rewriting a language's surface forms to remove lexical overlap is an established instrument. K et al. (2020) shift every character's code point to build a language that shares no characters with any other, and find that word-piece overlap matters little to multilingual BERT; Dufter & Schütze (2020) do the same at the level of token indices, Conneau et al. (2020b) remove identical strings with a language prefix, and Škrjanec et al. (2026) vary vocabulary sharing to study cognates in bilingual models. Those studies ask whether overlap is necessary for transfer in a pretrained model. We hold the supervision byte-identical, vary the training objective, and read the result as a quantity that separates configurations with the same score.

Surface form and lexicon induction. Language-specific units in multilingual models depend on orthography far more than on word order (Verma et al., 2026). Bilingual lexicon induction normally uses string similarity as free signal (Artetxe et al., 2017; Riley & Gildea, 2018; Kementchedjhieva et al., 2019); here spelling is the shortcut under test. The mapping frameworks of Artetxe et al. (2018); Lample et al. (2018) fit an orthogonal map between monolingual word vectors that never see a character, so the map is invariant under the cipher (Section 4); the supervised map fitted on the same links is the matched baseline every table carries, and the contrastive rows of Section 5.4 are the recipe of Li et al. (2022).

Shortcuts in self-supervised objectives. Networks prefer easy features (Geirhos et al., 2020). Chen et al. (2021) show that an easy feature added to images suppresses the others under a contrastive

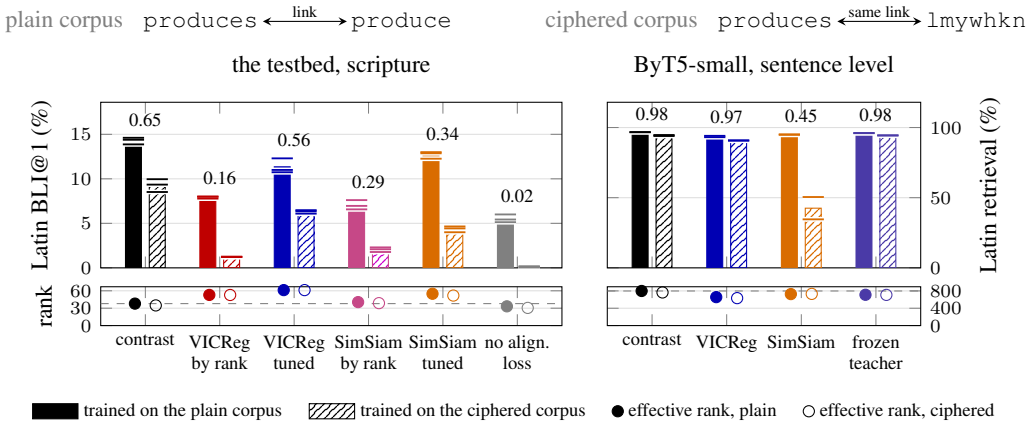


Figure 1: The test and the result. Top: one side of every Latin-script pair is rewritten by a fixed substitution of the alphabet; everything else is held. Below: the Latin-script score of a configuration trained on the plain corpus (solid) and on the ciphered corpus (hatched), the bar the mean over seeds and the ticks across it the seeds, with the mean kept fraction over each pair, for the testbed on scripture (BLI@1, 3 seeds) and for pretrained ByT5-small at sentence level (2 seeds, their published projection heads; its SimSiam keeps 0.36 and 0.53); beneath, the effective rank plain (filled) and ciphered (open), seed means, contrast's dashed. The rank-chosen devices reach contrast's rank and keep 0.16 and 0.29 against contrast's 0.65; no rank moves under the cipher by more than 0.08 of its value.

loss, and Robinson et al. (2021) that contrastive learning keeps suppressing features even when the negatives are made harder. Both interventions attenuate the shortcut feature and damage others with it; a substitution cipher removes spelling exactly and leaves every other property of the data in place.

The diagnostics. The alignment/uniformity decomposition (Wang & Isola, 2020) and effective rank (Roy & Vetterli, 2007; Garrido et al., 2023) are the standard label-free readings of a joint-embedding space, alongside the analysis of dimensional collapse (Jing et al., 2022) and of the predictor's dynamics (Tian et al., 2021). Their limits are known in part: Fang et al. (2024) show that the uniformity score cannot distinguish degrees of dimensional collapse and replace it with a Wasserstein distance to the uniform distribution, which we measure alongside the older readings, and Aldeneh et al. (2025) that rank does not identify the best layer of a speech model. Our claim is of a different kind: a function of the spectrum, however refined, certifies spread and not content, and the diagnostics can look right while the content is a shortcut.

3 SETTING

Testbed. A byte encoder maps a word's bytes to a real vector, a sequence model sees the sequence of those vectors, and a decoder must spell each word back. The encoder has no attention across words, so at inference a word's vector depends on that word alone. The vector is split into two halves of 64 coordinates: the alignment objective acts on the *aligned half* and the decoder reads only the *spelling half*, so nothing but the alignment objective acts on the aligned half, and every score below is cosine retrieval in that half. The model has 16.0M parameters and every configuration trains for 8,000 steps unless stated; Appendix A lists the architecture and every training setting, which all configurations share. A *configuration* is one training setup, and configurations differ only in the alignment loss.

Corpora and links. The first corpus, *scripture* below, is verse-aligned scripture in 8 languages spanning six scripts: 3 Latin-script languages (German, Spanish, French) against Russian, Japanese, Chinese, Hindi and Arabic, each paired with English. The Bible portion is a repack of the eBible corpus (Akerman et al., 2023) and the Qur'an portion the Tanzil translations (Tanzil Project, 2007) from OPUS (Tiedemann, 2012). The second corpus is TED2020 (Reimers & Gurevych, 2020), modern spoken-register subtitles in the same languages. A *record* is one aligned segment, a short run of verses or a subtitle line, with its two sides, the language's and the English. Word *links* are extracted by mutual-best Dice over co-occurrence within records (Och & Ney, 2003; Dyer et al., 2013). The alignment objective consumes the links of the training text, the *training links* or training pairs below.

Test. The test is bilingual lexicon induction, BLI@1, against a lexicon induced by the same procedure from held-out text, with every pair that also occurs among the training links removed (Kapoor & Narayanan, 2023). Queries, candidate pool and gold convention are fixed across methods, since retrieval depends on the pool through hubness and BLI scores are not comparable across query sets and gold conventions: queries and pool are intersected down to what every method can represent, gold is set-valued, and a hit means the argmax over the pool lies in the gold set. We call this shared test the *strict test*. A score is the per cent of queries hit, *pooled* over the languages unless a group is named; the *Latin score* is the score over the Latin-script languages alone, and *the other languages* are those the cipher of Section 4 never touches.

Measures and baseline. Beside BLI@1 we report the mean reciprocal rank of the gold (Glavaš et al., 2019) and paired McNemar tests (McNemar, 1947) on the per-query hits, a range being the minimum and maximum over seeds; Appendix C.2 repeats the main comparisons against an external dictionary. The *character-similarity baseline* predicts the candidate closest to the query in character bigrams. The gold lexicon is induced by co-occurrence and never by string similarity, so this baseline measures what spelling alone can do.

Objectives. Write u and v for the unit-norm aligned halves of the two words of a link and $\langle u, v\rangle$ for their cosine. Every objective below is a mean over the links of a batch. *Attraction* is

$$\mathcal{L}_{\mathrm{att}} = 1 - \langle u, v\rangle, \tag{1}$$

the alignment loss of Wang & Isola (2020) and, on unit vectors, VICReg's invariance term; it is satisfied by collapse. *Contrast* is a symmetric, temperature-scaled form of the InfoNCE loss of

van den Oord et al. (2018), applied in both directions over the linked pairs (u_i, v_i) of a batch at temperature $\tau = 0.07$,

$$\mathcal{L}_{\text{con}} = -\frac{1}{2}\left[\log \frac{e^{\langle u_i, v_i\rangle/\tau}}{\sum_j e^{\langle u_i, v_j\rangle/\tau}} + \log \frac{e^{\langle u_i, v_i\rangle/\tau}}{\sum_j e^{\langle u_j, v_i\rangle/\tau}}\right], \tag{2}$$

so a linked word must be nearer its partner than any other linked word in the batch. The two *anti-collapse devices* add to attraction, with no word-level negatives. VICReg (Bardes et al., 2022) adds its variance and covariance terms at weights w_v and w_c, here on unit-norm vectors, with correlations in place of covariances and the hinge's target the isotropic spread of the unit sphere. The SimSiam form (Chen & He, 2021), the *predictor objective* below, puts a predictor h on one side and a stop-gradient on the other, with h trained at a multiple of the encoder's learning rate; Appendix A writes both losses out. Every aligned configuration also carries the same small sentence-level contrastive term, $\mathcal{L}_{\text{con}}$ over the records of a batch on the mean of a record's aligned halves, at weight 0.1 against unity for the word-level loss and at temperature 0.07. Its gradient reaches the encoder, so "no negatives" means no word-level negatives; the sentence-level negatives are the only negatives the predictor objective has (Appendix B.3 removes the term). The *control* has no alignment loss and no sentence-level term, and its level is the *floor*.

Diagnostics. Let Z be the matrix of centred, unit-norm aligned halves of a set of words, with singular values s_i and $p_i = s_i / \sum_k s_k$. Effective rank (Roy & Vetterli, 2007) is

$$\text{erank}(Z) = \exp\Big(-\sum_i p_i \log p_i\Big), \tag{3}$$

which is d for an isotropic representation and unity at collapse. The tables read it on the *training probe*, the sample of held-out records on which the training loop reports its diagnostics, every word token counted. Uniformity and alignment are those of Wang & Isola (2020),

$$\mathcal{U} = \log \mathbb{E}_{x,y}\, e^{-2\|x-y\|^2}, \qquad \mathcal{A} = \mathbb{E}_{(x,y)\in\text{gold}}\, \|x-y\|^2, \tag{4}$$

the first over pairs of the test's word types, where lower is more uniform, and the second over its held-out gold pairs. The Wasserstein uniformity of Fang et al. (2024) fits the aligned halves with a Gaussian and takes its Wasserstein distance from the isotropic Gaussian that approximates the uniform distribution on the sphere (Appendix A); it vanishes for a uniform representation and for unit vectors is a function of the covariance spectrum alone, as effective rank is.

Standard encoder. Sections 5.4, 6.2 and 7 repeat the comparison on the encoder of ByT5-small (Xue et al., 2022), a standard byte-level transformer, from its pretrained weights and from random initialisation, with the alignment loss as the only loss, at word level and at sentence level (Appendix E); at sentence level the devices also run in their *published form*, their papers' projection heads and predictor in place of a linear head.

4 The cipher test

For each Latin-script language we draw a fixed derangement of the alphabet and apply it, preserving case, to that language's side of every record. English, the other languages, digits, punctuation and every non-ASCII byte are untouched. The permutation is applied in place to the prepared corpus, so the files that hold lengths, record boundaries, language labels and links are the plain corpus's own bytes. Every configuration therefore trains on identical supervision, the same links between the same word tokens. The only thing removed is that a word and its translation can share letters. The gold lexicon's target side passes through the same cipher, and the test is otherwise unchanged.

A substitution cipher is a bijection on word types, so type identity, frequency and co-occurrence are preserved, and the test has its own controls in the baseline rows of Table 1: a supervised linear map, orthogonal Procrustes over count-based embeddings (Levy & Goldberg, 2014) fitted on the same links and retrieved by nearest neighbour or with CSLS (Lample et al., 2018), scores the same on both corpora, 21.78% on plain and 21.78% on ciphered scripture; the character-similarity baseline falls from 9.17% to 0.09%; and the languages that were not ciphered should not move, which Section 5.3 checks per configuration.

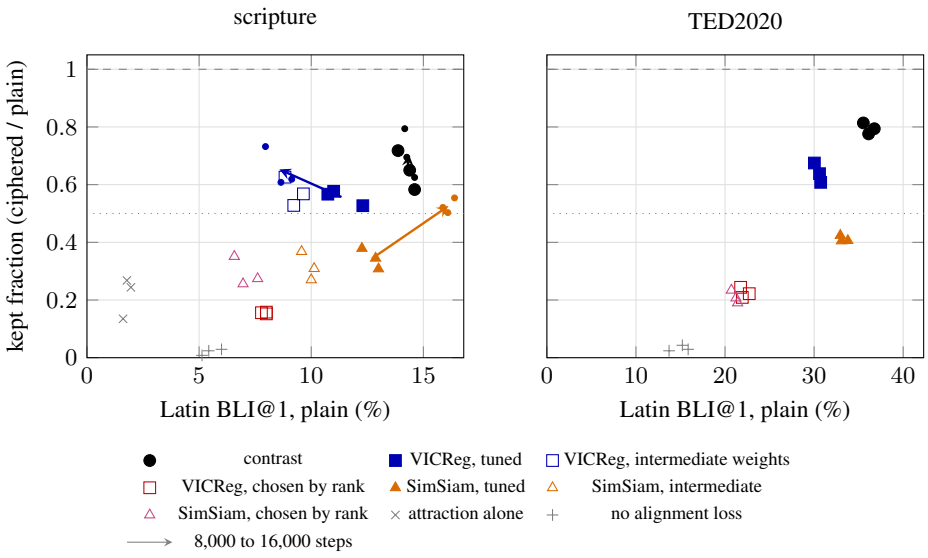


Figure 2: The cipher test. Each point is a trained model: its Latin-script score trained on the plain corpus, against the fraction of that score the same configuration and seed keeps when trained on the ciphered corpus; the dashed line is no loss, the dotted line equal parts kept and lost. Arrows run from each configuration's centroid at 8,000 steps to its centroid at 16,000 steps (Table 5), the small dots the seeds there.

Table 1: Scripture, strict test, means over 3 seeds: BLI@1 pooled and over the 3 Latin-script and 5 other languages, MRR, the Latin score of the same configuration trained on the ciphered corpus and the fraction kept; then effective rank on the training probe, uniformity and alignment (Wang & Isola, 2020) and the Wasserstein uniformity (Fang et al., 2024) on the test's word types, lower being more uniform. Baselines have no rank; their entries are Latin scores. Bold is best per column, subscripts are seed standard deviations, Δ is BLI@1 minus contrast's, and the kept column is shaded by value.

	scores						diagnostics			
	BLI@1	Δ	MRR	Latin	other	ciphered	kept	rank	unif.	W_2	align.
contrast	**10.09**$_{\pm 0.23}$		**13.65**	**14.29**	**6.97**	**9.28**	**0.65**$_{\pm 0.07}$	37.9	−3.32	0.77	1.57
VICReg, chosen by rank	4.04$_{\pm 0.07}$	-6.05	6.05	7.93	1.15	1.23	0.16$_{\pm 0.00}$	52.9	−3.19	0.68	1.63
VICReg, tuned	8.19$_{\pm 0.39}$	-1.90	11.73	11.35	5.85	6.30	0.56$_{\pm 0.03}$	61.2	−3.01	0.66	1.45
SimSiam, chosen by rank	4.78$_{\pm 0.32}$	-5.31	7.14	7.04	3.10	2.06	0.29$_{\pm 0.05}$	40.5	−2.53	0.87	1.32
SimSiam, tuned	8.52$_{\pm 0.18}$	-1.57	11.93	12.71	5.41	4.36	0.34$_{\pm 0.04}$	54.9	−3.35	0.56	1.59
attraction alone	0.89$_{\pm 0.05}$	-9.20	1.96	1.78	0.23	0.39	0.22$_{\pm 0.07}$	11.1	−0.15	1.34	0.11
no alignment loss	2.43$_{\pm 0.18}$	-7.66	3.74	5.52	0.14	0.12	0.02$_{\pm 0.01}$	33.2	−2.06	0.98	1.40
character-similarity baseline				9.17		0.09					
Procrustes, same links				21.78		21.78					
Procrustes, same links, CSLS				22.87		22.87					

For a configuration we report the *kept fraction*: its score on the Latin-script languages when trained on the ciphered corpus, divided by the same score when trained on the plain corpus, per seed. A model that learned translation and used cognates only as a convenience keeps most of its score. A model whose alignment in those languages was the cognates keeps little. The ciphered language is the situation the other scripts are already in, since Russian shares no letters with English either, so a kept fraction below unity is expected even for a model that learned a lexicon.

5 THE DIAGNOSTICS CERTIFY SPREAD, NOT CONTENT

This section reads the diagnostics beside the cipher, first at the weights the diagnostics themselves select, then across a sweep, then on a standard encoder.

5.1 HEALTHY DIAGNOSTICS SELECT A SHORTCUT SOLUTION

Rank-chosen weights. We first choose each device's weight the way the literature recommends when labels are unavailable, by the effective rank of its own training probe at a single seed (Garrido et al., 2023); the tables call these settings *rank-chosen*, and the best settings of the sweep in Section 5.2 *tuned*. Both devices then do what they are for (Table 1). Attraction alone ends at a rank of 11.1 and scores 0.89%, below the control's 2.43%. VICReg restores the rank to 52.9 and the SimSiam form to 40.5, at or above contrast's 37.9. On uniformity and alignment the VICReg configuration is close to the contrastive one, and by the Wasserstein reading it is the more uniform of the two: close on every label-free reading and ahead on two. It scores 4.04%, which is 0.40 of contrast, and the SimSiam form 4.78%, which is 0.47; the mean reciprocal rank orders the configurations as BLI@1 does.

Under the cipher. The cipher says what those scores were made of. On the Latin-script languages contrast falls from 14.29% to 9.28% and keeps 0.65. VICReg falls from 7.93% to 1.23% and keeps 0.16. The SimSiam form keeps 0.29, and the control, which never saw a link, 0.02. Both devices land at the level they reach in the scripts that share no letters with English. TED2020 gives the same picture (Table 4, Appendix B.2).

Three checks agree. The spectrum that rank, uniformity and the Wasserstein reading summarise does not move: for each configuration the plain and the ciphered ranks differ by under 4.8 at every seed (Figure 5, Appendix B.2). A ridge regression from character n-gram counts explains 0.66 of the variance of VICReg's aligned half, against 0.34 for contrast, so the variance VICReg restored is in large part the variance the byte encoder gives for free. The probe is no substitute for the cipher: it explains 0.35 of the SimSiam form's aligned half, no more than contrast's, and the two keep very different fractions. Contrast is not free of the shortcut either, keeping 0.65 (Section 8).

5.2 TUNING MOVES THE SCORE AND NOT THE DIAGNOSTICS

The sweep. Neither device was well served by the rank rule. We sweep VICReg's variance and covariance weights over 21 settings and SimSiam over 20 settings of its loss weight and of a multiplier on the predictor's learning rate (Table 3). A low variance weight with a high covariance weight reaches 8.19% over 3 seeds, 0.81 of contrast. SimSiam's loss weight alone never passes 4.87%; a predictor that learns faster than the encoder, as the analysis of Tian et al. (2021) suggests, reaches 8.52%, 0.84 of contrast.

The rank does not see any of this. Across the 11 VICReg settings whose rank lies between 59 and 62, the score runs from 4.23% to 8.62%. Contrast remains ahead of both tuned devices in every seed, by 1.2 to 2.4 points over VICReg ($p \leq 0.002$) and 1.1 to 1.9 over SimSiam ($p \leq 0.002$), and on TED2020 as well (Table 4). It reaches that score with no search at all.

5.3 THE SAME SCORE IS NOT THE SAME LEXICON

Tuning raised both devices to about the same place. On the Latin-script languages the tuned VICReg scores 11.35% and the tuned SimSiam 12.71%, a little ahead. The tuned SimSiam is also, by alignment and uniformity themselves, indistinguishable from contrast: a uniformity of -3.35 where contrast has -3.32, an alignment of 1.59 where contrast has 1.57, and a higher rank; by the Wasserstein reading it is the most uniform configuration in the table (Table 1).

Under the cipher. The two devices part. VICReg keeps 0.56 of its Latin score, close to contrast's 0.65. SimSiam keeps 0.34, a little above what it kept before tuning, and that fraction is the sentence-level term's: with the term removed it keeps 0.01, and a few points of its plain score go with it (Appendix B.3). Tuning moved SimSiam's score far more than its dependence on spelling; for VICReg the kept fraction rose with the first step of tuning and stayed there as the score rose further (Figure 2). TED2020 repeats the separation with the same settings carried over (Table 4). The

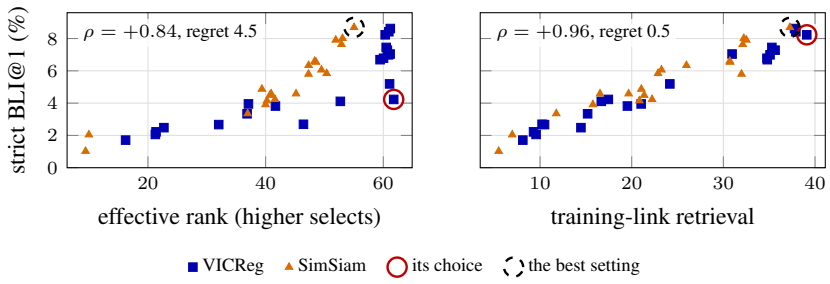


Figure 3: Selecting without labels. Each panel is one criterion against strict BLI@1 over the 41 non-contrastive settings (scripture, a single seed; Figure 7 in Appendix D.2 draws the remaining criteria). The red ring is the setting the criterion selects, the black ring the best setting; the corner gives the Spearman correlation and the regret of the selection in points. Rank saturates once the representation is spread and selects a shortcut solution; training-link retrieval tracks the test.

languages that were not ciphered barely move for contrast and VICReg; for SimSiam they do move (Appendix B.2).

A longer budget. At 16,000 steps (Table 5, Appendix B.2; the arrows of Figure 2) contrast rises to 10.66% and the tuned VICReg falls to 6.82%, each in every seed, and the tuned SimSiam passes contrast on plain text and stays below it under the cipher, keeping 0.53 against 0.70: the plain test would rank it first at this budget and the cipher test last.

5.4 The pattern holds on a standard encoder

Words. On ByT5-small from random initialisation, contrast and the tuned VICReg keep 0.77 and 0.74 of their Latin-script score in every seed; the other configurations sit too near the untrained encoder's character-overlap floor to keep a fraction of.

Sentences. Encoders of this kind are trained in practice on sentence pairs, and there the scores are far from the floor. With the devices in their published form at 2 seeds (Table 13, Appendix E), contrast keeps 0.97 to 0.98 and published VICReg 0.88 to 0.99. No SimSiam setting keeps more than 0.53, and its fraction rises with the predictor's rate.

Other encoders and corpora. Out of domain on Tatoeba (Artetxe & Schwenk, 2019) what the devices transfer is what the cipher removes, XLM-R base orders as ByT5 does, and on TED2020 every device but the slowest SimSiam predictor keeps most of its score, so the sentence-level separation is scripture's (Tables 15, 14 and 20).

6 Selecting without labels

If the diagnostics do not see what a device learned, they cannot choose its weights. This section asks what can, over the settings of the sweep and then on the standard encoder.

6.1 Training-link retrieval selects where the diagnostics do not

Where rank fails. Over the 41 non-contrastive settings, effective rank correlates with the test at $\rho = 0.84$ and uniformity at $\rho = 0.59$ (Figure 3; Appendix D.2). Both saturate. Once a representation is spread they no longer order the settings, and the setting each ranks first is the same shortcut solution, a *regret* of 4.5 points below the best setting. The Wasserstein uniformity does not repair this: over the VICReg settings it selects the setting rank selects, and over the VICReg and SimSiam settings together it lands on the best setting only because that setting is also the most uniform SimSiam.

A criterion closer to the task. We score each model on the links it was trained on, at the level of word types and over a large pool: for a sample of linked source types, is the nearest target type

a linked one? This is *training-link retrieval*, "train" in the tables. The test lexicon is disjoint from these links by construction, so the criterion never reads the test. It correlates with the test at $\rho = 0.96$ and selects a setting 0.5 points below the best, and it does as well against the languages in which no spelling shortcut exists, $\rho = 0.98$, where rank falls to $\rho = 0.75$. Each setting is a single run; a bootstrap over the settings (Table 10) puts the paired difference between this correlation and rank's at 0.03 to 0.28, above zero, while the intervals of rank and uniformity overlap. The criterion measures fit to the supervision, and a model could in principle fit the links without generalising; in this family of models, over the settings we scored, it does not. We regard the result as evidence that a task-shaped criterion is available at no labelling cost, not as a general recipe.

6.2 Rank selects a dead run on the standard encoder

Words. On ByT5-small, rank again selects 3.5 and 4.7 points below the best from random initialisation and from the pretrained weights, training-link retrieval within 0.1 and 0.1.

Sentences. In the sentence-level batch of Section 5.4 one configuration, VICReg 5/100 with a linear head, retrieves at most 0.4 per cent in every seed and on both corpora; it has the highest effective rank of the batch and is the most uniform, so rank and uniformity each select it, with a regret of 94.4 points, where retrieval of the training pairs selects with a regret of 0.0.

7 A fixed target learns the lexicon

An *anchor* is anything beside the pair loss that fixes where the representation may go: a frozen teacher, a decoder that must reconstruct, a language-model term. This section reads the first, and Appendix F the rest.

A frozen teacher. Multilingual distillation (Reimers & Gurevych, 2020), the negative-free recipe used in practice, is attraction to a frozen teacher that encodes the English side, which is never ciphered. On ByT5-small at 2 seeds (Figure 4; Table 18) the student keeps 0.98 at the learning rate selected on a *development half* of the test's queries and scored on the other (Appendix D.1), and the other languages move by at most 0.6 points. A fixed target without negatives learns the lexicon as contrast does, and the lexicon is the student's own: with an English-only sentence encoder as teacher, all-MiniLM-L6-v2 (Wang et al., 2020; Reimers & Gurevych, 2019; Sentence-Transformers, 2021), which never read the other languages, it keeps 0.93 to 0.96.

Invisible on plain text. With the teacher trained on this corpus, whether a given setting learned it is again invisible on plain text: a decade slower the student keeps 0.54 to 0.55, while across the rates the plain scores lie within 0.8 points and the rank between 708 and 749; under the cipher the slow student retrieves 81 per cent of its training pairs against 99, so the criterion of Section 6.1 registers what rank does not, on TED2020 too (Appendix F.6).

Sweeping the teacher's momentum alone, from frozen to stop-gradient, finds no silent regime: a fast target collapses where every diagnostic sees it (Figure 10; Appendix F).

8 Discussion and limitations

What the diagnostics certify. Effective rank, uniformity, its Wasserstein refinement and alignment are functions of the spectrum of the representation, and Figure 5 shows that spectrum unchanged under an intervention that removes most of what two configurations had learned. They certify that a representation is spread. In this setting they do not certify what it is spread along, and a selection rule built on them prefers the spread that the byte encoder gives for free. We expect the same of any diagnostic that reads the spectrum alone, however it is refined.

Negatives as a cost, not a prohibition. Contrast keeps 0.65 of its Latin-script score on scripture, and character n-grams explain part of its aligned half. Negatives make the spelling solution expensive, since it settles only the cognate links and the loss must still separate every other pair, but they do not rule it out, which is what Robinson et al. (2021) found on images. What the objectives that

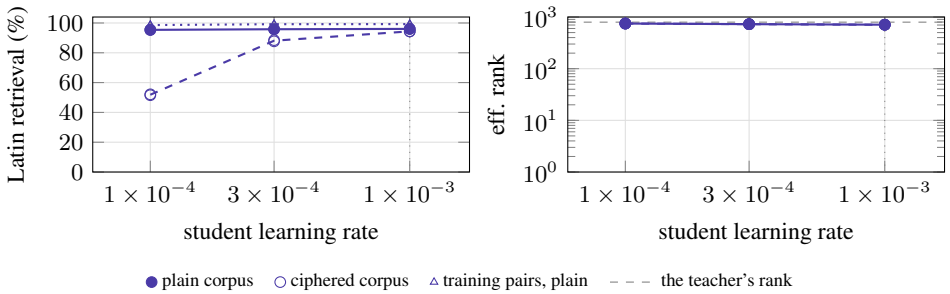


Figure 4: A fixed target. Distillation to a frozen contrast teacher on ByT5-small against the student's learning rate, 2 seeds (Table 18). Left: Latin hard-pool retrieval trained on the plain corpus (filled, solid) and on the ciphered corpus (open, dashed), with the plain runs' training-pair retrieval dotted and the selected rate marked. Right: effective rank at the encoder output, the teacher's dashed. Plain text and rank do not separate the rates; the cipher does.

learned the lexicon share is not the presence of negatives but a target that cannot drift toward the shortcut: the batch in contrast, the teacher in distillation. The predictor objective, whose target is its own moving copy, learned the cognates alone when its sentence-level negatives were removed. The anchors were run mostly on one corpus at one budget, the slow student's fit at the longer budget is unfinished (Appendix F.6), and on XLM-R the same teacher keeps at most 0.35 at this budget and 0.71 at a longer one (Appendix F.3), so the fixed anchor suffices on the byte-level encoder and not, at the budgets run, on the subword one. The claim about VICReg and SimSiam is that the tuned configurations we found differ in kind, not that no setting of SimSiam learns a lexicon: a search on the ciphered corpus finds one that keeps 0.96 (Table 16), at a single seed, which plain text does not distinguish from the settings that keep little.

The verdict is about rank one. On zero-shot natural language inference over frozen vectors (Appendix C.3) every device run on it transfers as well as contrast and keeps its margin under the cipher, the rank-chosen VICReg included; BLI orders the aligned configurations and the task does not. The nearest neighbour is what the cipher removes from the devices, and Appendix C.1 shows contrast's advantage over them concentrated at the first rank. Lexicon induction (Lample et al., 2018; Glavaš et al., 2019), cross-lingual retrieval and bitext mining (Artetxe & Schwenk, 2019) begin at the nearest neighbour, so a diagnostic blind to a nearest-neighbour failure cannot certify a representation for those uses.

The testbed. The tuned device settings were first chosen on the whole test; Appendix D.1 re-selects on a development half and restates every comparison on the other, where the orderings survive, but the split cannot undo choices made with the whole test in view. The model is a testbed and not a method: one small model in which the alignment objective is the only pressure on the aligned half, and a supervised linear map on the same links scores higher on one of the two corpora. The cipher removes shared letters and nothing else, so the kept fraction compares configurations and is not an absolute measure.

9 Conclusion

We asked whether the label-free diagnostics of collapse certify what an alignment objective has learned. In cross-lingual word and sentence alignment they do not: two anti-collapse devices chosen by rank look as healthy as contrast and lose most of their score when cognates are hidden, tuning raises both while the diagnostics register none of it, and the tuned devices keep different fractions at the same score and the same uniformity. Retrieval of the training links orders the settings where the diagnostics do not, in the testbed and on a standard encoder. The objectives that learn the lexicon share a target that cannot drift toward the shortcut, the batch in contrast or a frozen teacher in distillation, and whether an anchored setting learned it is again invisible on plain text. Whether the same holds under the linear probes of the image setting is the question we would ask next.

REPRODUCIBILITY STATEMENT

Every number, table and figure coordinate in this paper is generated from scored artefacts by one script, which also asserts each claim marked in the source and fails the build otherwise; a second script refuses any number typed into the prose. The experiments behind the new results are 126 single-GPU training runs; seed counts, corpora and budgets differ across the experiments and are listed with each table. The cipher is a short script applied to the prepared corpus, with its permutations recorded beside the data. Appendix A lists the architecture and every training hyperparameter, read from the configuration each checkpoint records and asserted equal across every run behind the tables. Code, configurations and checkpoints will be released.

ETHICS STATEMENT

This work involves no human subjects and no annotation; every corpus is a public resource used as released. The first corpus is two religious canons, chosen because they are massively parallel in these languages at verse granularity. That a word and its rendering receive nearby vectors is a claim about translation statistics in particular translations, not about doctrinal or cultural equivalence. The study covers well-resourced languages paired with English and should not be assumed to carry to under-served ones (Joshi et al., 2020).

AI DISCLOSURE

In this work, we used generative AI tools to implement methods, design or provide feedback on research methodology or experiments and interpret results. We have not used generative AI tools to propose or refine hypotheses and clean and reformat dataset. Generating synthetic data sets, helping to develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to create or modify scientific figures or images, suggest experimental parameters, create or edit software code, draft parts of a research paper, summarize or analyze existing literature, edit a research paper to improve readability, identify relevant literature and format references. We have reviewed all AI-assisted work: AI-written code was run by the authors and its outputs are the scored artefacts from which one script generates every number, table and figure and asserts every claim in the text, and AI-drafted prose was read by the authors against those results, with every reference checked against its source. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Taeho Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. The eBible corpus: Data and model benchmarks for Bible translation for low-resource languages. *arXiv preprint arXiv:2304.09919*, 2023.

Zakaria Aldeneh, Vimal Thilak, Takuya Higuchi, Barry-John Theobald, and Tatiana Likhomanenko. Towards automatic assessment of self-supervised speech models using rank. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2025, Hyderabad, India, April 6-11, 2025*, pp. 1–5. IEEE, 2025. doi: 10.1109/ICASSP49660.2025.10889651. URL https://doi.org/10.1109/ICASSP49660.2025.10889651.

Mikel Artetxe and Holger Schwenk. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610, 2019. doi: 10.1162/tacl_a_00288. URL https://aclanthology.org/Q19-1038/.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 451–462, 2017.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Ting Chen, Calvin Luo, and Lala Li. Intriguing properties of contrastive losses. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 11834–11845, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/628f16b29939d1b060af49f66ae0f7f8-Abstract.html.

Xinlei Chen and Kaiming He. Exploring simple Siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15750–15758, 2021.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating cross-lingual sentence representations. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6022–6034. Association for Computational Linguistics, 2020b. doi: 10.18653/v1/2020.acl-main.536. URL https://aclanthology.org/2020.acl-main.536/.

Philipp Dufter and Hinrich Schütze. Identifying elements essential for BERT's multilinguality. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4423–4437. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.emnlp-main.358. URL https://aclanthology.org/2020.emnlp-main.358/.

Chris Dyer, Victor Chahuneau, and Noah A. Smith. A simple, fast, and effective reparameterization of IBM model 2. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 644–648, 2013.

Xianghong Fang, Jian Li, Qiang Sun, and Benyou Wang. Rethinking the uniformity metric in self-supervised learning. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=3pf2hEdu8B.

Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann LeCun. RankMe: Assessing the downstream performance of pretrained self-supervised representations by their rank. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10929–10974. PMLR, 23–29 Jul 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020. doi: 10.1038/s42256-020-00257-z.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 710–721, Florence, Italy, 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1070.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, 2020.

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth. Cross-lingual ability of multilingual BERT: An empirical study. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=HJeT3yrtDr.

Sayash Kapoor and Arvind Narayanan. Leakage and the reproducibility crisis in machine-learning-based science. *Patterns*, 4(9):100804, 2023.

Yova Kementchedjhieva, Mareike Hartmann, and Anders Søgaard. Lost in evaluation: Misleading benchmarks for bilingual dictionary induction. 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)*, pp. 3336–3341, 2019.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*, 2014.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. Improving word translation via two-stage contrastive learning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4353–4374, Dublin, Ireland, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.299. URL https://aclanthology.org/2022.acl-long.299/.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525, 2020. doi: 10.18653/v1/2020.emnlp-main.365.

Parker Riley and Daniel Gildea. Orthographic features for bilingual lexicon induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 390–394, 2018.

Joshua Robinson, Li Sun, Ke Yu, Kayhan Batmanghelich, Stefanie Jegelka, and Suvrit Sra. Can contrastive learning avoid shortcut solutions? In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 4974–4986, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/27934a1f19d678a1377c257b9a780e80-Abstract.html.

Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610, 2007.

Sentence-Transformers. all-MiniLM-L6-v2. Model card, https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2, 2021. MiniLM-L6 fine-tuned as a sentence encoder on English sentence pairs.

Iza Škrjanec, Irene Elisabeth Winther, Vera Demberg, and Stefan L. Frank. Is cross-lingual transfer in bilingual models human-like? A study with overlapping word forms in Dutch and English. *arXiv preprint*, abs/2604.07067, 2026. URL https://arxiv.org/abs/2604.07067.

Tanzil Project. The Tanzil project: Qur'an navigator and translations. https://tanzil.net/trans/, 2007. Translations distributed for non-commercial use; obtained through the OPUS-Tanzil v1 release.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

Jörg Tiedemann. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pp. 2214–2218, 2012.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Aastha A K Verma, Anwoy Chatterjee, Mehak Gupta, and Tanmoy Chakraborty. Multilingual language models encode script over linguistic structure. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 33685–33719. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1554. URL https://aclanthology.org/2026.acl-long.1554/.

Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 2495–2504. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00252. URL https://openaccess.thecvf.com/content/CVPR2021/html/Wang_Understanding_the_Behaviour_of_Contrastive_Loss_CVPR_2021_paper.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122, New Orleans, Louisiana, 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 2022.

A MODEL AND TRAINING

Table 2 lists the model and the training settings shared by every configuration of the paper. The values are read from the configuration that each checkpoint records at the end of training, and the build asserts that the 176 checkpoints behind the main tables agree on every row. Three exceptions are asserted by name: one configuration trained alongside the others and reported nowhere in this paper, attraction with the decoder reading the aligned half too, differs in the decoder's wiring alone; the longer budget of Table 5 differs in its step count alone; and the 14 checkpoints of the sentence-term ablation (Appendix B.3) differ in that term's weight alone, at zero. The alignment loss is the only setting that changes between configurations, and its weights are those of Tables 1 and 3. The SimSiam predictor is a two-layer perceptron with a batch-normalised hidden layer, as in Chen & He (2021), on the aligned half; its learning rate is the base rate times the multiplier the tables report. Training runs in bfloat16 autocast on one GPU. A batch of 192 record pairs carries 5,063 linked word pairs on average on scripture and 4,394 on TED2020. Contrast ranks over at most 512 of them, deduplicated by word type; attraction and both devices act on all of them, so the batch over which VICReg's variance and covariance terms and SimSiam's batch normalisation are computed is every linked pair of the batch, not the 192 records.

The losses in full. In the notation of Section 3, VICReg's variance and covariance terms are computed on each side of the batch and averaged,

$$\mathcal{L}_{\mathrm{vic}} = \mathcal{L}_{\mathrm{att}} + w_v \, \frac{1}{d} \sum_j \max\!\left(0,\; 1 - \frac{\sigma_j}{\gamma/\sqrt{d}}\right) + w_c \, \frac{1}{d(d-1)} \sum_{j \neq k} \rho_{jk}^2, \tag{5}$$

where σ_j is the standard deviation of coordinate j over the batch, ρ_{jk} the correlation of coordinates j and k, $d = 64$, and γ is unity, so that the hinge's target is the isotropic spread of the unit sphere. The predictor objective, with predictor h and stop-gradient sg, is

$$\mathcal{L}_{\mathrm{sim}} = \frac{1}{2} \left[\big(1 - \langle h(u), \mathrm{sg}(v) \rangle\big) + \big(1 - \langle h(v), \mathrm{sg}(u) \rangle\big) \right]. \tag{6}$$

The Wasserstein uniformity of Fang et al. (2024) fits the unit-norm aligned halves with a Gaussian of mean μ and covariance Σ, and is

$$W_2 = \left(\|\mu\|^2 + 1 + \operatorname{tr}\Sigma - \tfrac{2}{\sqrt{d}} \operatorname{tr}\Sigma^{1/2} \right)^{1/2}, \tag{7}$$

its distance from the isotropic Gaussian that approximates the uniform distribution on the sphere.

B THE TESTBED IN FULL

This appendix backs Section 5 on the testbed: every setting of the sweep, the second corpus with the readings of the spectrum and the budget, and the ablation of the sentence-level term.

B.1 EVERY SETTING

Table 3 scores every non-contrastive setting of the sweep of Section 5.2 on scripture at a single seed, with the effective rank, the uniformity and the training-link criteria of Section 6.1 beside the test score. Section 5.2 reads it.

B.2 THE SECOND CORPUS, THE SPECTRUM AND THE BUDGET

TED2020. Table 4 repeats Table 1 on TED2020 with the tuned settings carried over from scripture; Sections 5.1 to 5.3 read it beside the scripture table. The controls of Section 4 hold here as well: the supervised linear map goes from 26.34% to 26.41% and, retrieved with CSLS, the hubness-corrected

Table 2: The model and the training settings shared by every configuration, read from the configuration each checkpoint records. The build asserts that the 176 checkpoints behind the tables agree on every row, with the exceptions Appendix A names: a configuration trained alongside the others and reported in no table, in which the decoder reads the aligned half too, differs in the decoder's wiring alone; the longer budget of Table 5 runs 16,000 steps with its cosine schedule stretched to match; and the sentence-term ablation of Table 6 sets that term's weight to zero. The learning rate warms up linearly and decays by cosine from the peak to the floor; the alignment losses are ramped in linearly over their own warmup. Linked word pairs per step are means over the logged steps of the training runs, one value per corpus; contrast ranks over the capped, deduplicated set, attraction and both devices act on all of them. Every aligned configuration adds the same sentence-level InfoNCE over the 192 record pairs of the batch; the control carries none.

	setting	value
model		
byte encoder	layers / heads / width	2 / 4 / 192
	bytes per word	20
word vector	halves $\times$ width	2×64
sequence model	layers / heads / width	6 / 6 / 384
	context, words	256
decoder	layers / heads / width	2 / 4 / 192
	context dropout	0.5
all	dropout	0
	bias terms	none
	parameters	16.0M
training		
optimiser	AdamW, β_1 / β_2	0.9 / 0.95
	weight decay	0.1
	gradient clip	1
	precision	bfloat16 autocast
schedule	peak / floor learning rate	6×10^{-4} / 6×10^{-5}
	warmup steps	500
	steps	8,000
batch	record pairs	192
	words per record, at most	128
	linked word pairs, scripture / TED2020	5,063 / 4,394
	pairs contrast ranks over	512
alignment	warmup steps	2,000
	sentence-level InfoNCE weight	0.1
	temperature, sentence / word	0.07 / 0.07

retrieval of the mapping frameworks with a neighbourhood of 10, from 27.06% to 27.12% (on scripture from 22.87% to 22.87%), and the character-similarity baseline falls from 21.81% to 0.26%. The languages that were not ciphered are the intervention's control, and on this corpus they move a little for every configuration: contrast loses 0.9 points of 19.80% and the tuned VICReg 1.6 of 13.91%, each keeping 0.95 and 0.88. For SimSiam they are not a control: the rank-chosen configuration keeps 0.59 of its score outside the Latin scripts when the Latin scripts are ciphered and the tuned one 0.57, as the tuned one keeps 0.73 on scripture. What SimSiam learns across scripts appears to lean on structure it acquires through the cognates. We report this as an observation and do not explain it.

Mean reciprocal rank. The MRR column of Tables 1 and 4 orders the configurations exactly as BLI@1 does, on both corpora: contrast's 13.65% against 10.09% at rank one, the rank-chosen VICReg's 6.05% against 4.04%. A miss is rarely a near miss for one objective and a distant one for another, so the top-one score is not hiding a reordering below it.

The spectrum. Figure 5 draws the normalised singular values that effective rank, uniformity and the Wasserstein reading summarise, for each configuration of Table 1 trained on the plain and on the ciphered corpus: the two curves lie on each other and the two ranks differ by under 4.8 at every seed, while the kept fractions of Figure 2 run from 0.02 to 0.65. The cipher moves the lexicon and leaves the spectrum.

Table 3: Every non-contrastive setting scored, scripture, a single seed. VICReg settings are variance / covariance weight; SimSiam settings are loss weight and the multiplier on the predictor's learning rate. Columns: strict BLI@1, its non-Latin part, the fraction of contrast's score at the same seed, the effective rank, the uniformity over English word types, and the two training-link criteria of Table 10.

	setting	BLI@1	non-Latin	frac.	rank	unif.	train	train, no shortcut
SimSiam	0.1, ×1	3.35	1.55	0.34	36.9	−2.68	11.8	4.4
SimSiam	0.3, ×1	3.91	2.07	0.40	40.0	−2.66	15.8	8.6
SimSiam	0.5, ×1	4.58	2.65	0.47	40.9	−2.55	19.7	12.6
SimSiam	1, ×1	4.87	3.33	0.50	39.4	−2.25	21.1	14.6
SimSiam	1.5, ×1	4.49	3.07	0.46	40.7	−1.87	21.4	15.6
SimSiam	2, ×1	4.23	2.62	0.43	41.6	−1.70	22.2	16.7
SimSiam	3, ×1	4.15	2.68	0.42	40.2	−1.43	20.8	15.3
SimSiam	5, ×1	1.02	0.16	0.10	9.4	−0.17	5.5	0.1
SimSiam	0.5, ×10	4.58	1.97	0.47	45.2	−2.81	16.5	7.6
SimSiam	1, ×10	6.34	3.52	0.65	47.3	−2.89	26.0	17.9
SimSiam	2, ×10	6.60	4.39	0.67	48.3	−2.94	30.7	23.8
SimSiam	3, ×10	6.54	4.46	0.67	48.6	−2.66	30.7	24.1
SimSiam	5, ×10	5.80	3.94	0.59	47.3	−2.45	32.0	26.1
SimSiam	25, ×1	2.04	0.13	0.21	10.0	−0.10	7.0	0.1
SimSiam	1, ×30	5.84	3.33	0.59	50.4	−3.03	22.9	14.3
SimSiam	2, ×30	7.64	4.49	0.78	52.9	−3.14	32.1	24.3
SimSiam	1, ×100	6.06	3.17	0.62	49.5	−3.00	23.3	14.7
SimSiam	2, ×100	8.01	5.11	0.82	53.1	−3.26	32.2	24.4
SimSiam	3, ×100	8.69	5.59	0.88	55.0	−3.27	37.3	29.6
SimSiam	2, ×300	7.91	4.81	0.81	51.9	−3.16	32.5	24.7
VICReg	1 / 1	2.48	1.78	0.25	22.7	−0.53	14.5	10.4
VICReg	25 / 1	1.71	0.48	0.17	16.2	−0.37	8.1	3.2
VICReg	5 / 5	3.95	2.81	0.40	37.1	−1.39	21.0	16.6
VICReg	25 / 25	4.11	1.23	0.42	52.7	−3.16	16.7	7.8
VICReg	1 / 25	6.78	5.30	0.69	60.0	−2.27	34.8	29.9
VICReg	2 / 25	8.23	6.11	0.84	60.3	−2.48	39.1	34.5
VICReg	5 / 25	6.69	5.11	0.68	59.4	−2.57	34.8	29.9
VICReg	100 / 25	2.67	0.19	0.27	32.0	−2.68	10.5	1.4
VICReg	1 / 50	7.28	5.72	0.74	60.7	−2.43	35.6	30.8
VICReg	2 / 50	8.42	6.43	0.86	60.9	−2.57	37.9	33.3
VICReg	5 / 50	7.45	5.20	0.76	60.5	−2.75	35.3	30.1
VICReg	1 / 100	3.82	3.00	0.39	41.7	−0.64	19.5	14.6
VICReg	2 / 100	6.99	4.94	0.71	60.9	−2.70	35.1	29.9
VICReg	5 / 100	8.62	5.88	0.88	61.2	−3.02	37.8	32.3
VICReg	25 / 100	4.23	1.23	0.43	61.8	−3.37	17.4	8.1
VICReg	100 / 100	2.69	0.32	0.27	46.4	−3.11	10.2	1.5
VICReg	1 / 250	2.06	0.36	0.21	21.3	−0.09	9.6	3.0
VICReg	2 / 250	3.34	2.13	0.34	36.9	−0.52	15.2	9.7
VICReg	5 / 250	7.04	4.52	0.72	61.2	−3.24	30.9	24.5
VICReg	5 / 500	5.19	2.75	0.53	61.1	−3.34	24.2	16.8
VICReg	5 / 1000	2.22	0.16	0.23	21.4	−0.40	9.3	1.8

The training budget. A device that is slower to converge might close the gap with a longer budget, more training steps. Table 5 holds contrast, the tuned VICReg and the tuned SimSiam at 16,000 steps on scripture, plain and ciphered, which Figure 2 draws as arrows from the shorter budget. Contrast rises to 10.66% and the tuned VICReg falls to 6.82%, each in every seed. The tuned SimSiam rises to 12.10% and passes contrast on plain text, 16.12% against 14.35% on the Latin-script languages; on the ciphered corpus it scores 8.48% against 10.10%, below contrast in every seed, keeping 0.53 against 0.70. The plain test would rank the tuned SimSiam first at this budget; the cipher test ranks it last. Its kept fraction did rise with the budget, from 0.34, so the longer run learned some lexicon as well as more spelling. Where the remaining gap sits is read at depth in Appendix C.1: contrast's advantage is concentrated at the first rank.

Table 4: TED2020, same protocol and columns as Table 1, on the corpus's own strict test. The tuned settings are the ones chosen on scripture; this corpus's development half confirms VICReg's (Appendix D.1), and SimSiam was not swept here. Bold is best per column, subscripts are seed standard deviations, Δ is BLI@1 minus contrast's, and the kept column is shaded by value.

	scores							diagnostics
	BLI@1	Δ	MRR	Latin	other	ciphered	kept	rank
contrast	**25.90**$_{\pm 0.36}$		**32.09**	**36.14**	**19.80**	**28.72**	**0.79**$_{\pm 0.02}$	40.0
VICReg, chosen by rank	10.21$_{\pm 0.18}$	-15.69	13.95	22.14	3.10	4.98	0.23$_{\pm 0.02}$	54.2
VICReg, tuned	20.09$_{\pm 0.25}$	-5.81	26.11	30.46	13.91	19.50	0.64$_{\pm 0.03}$	62.2
SimSiam, chosen by rank	11.51$_{\pm 0.16}$	-14.39	15.46	21.12	5.78	4.44	0.21$_{\pm 0.02}$	40.2
SimSiam, tuned	22.17$_{\pm 0.27}$	-3.73	27.96	33.27	15.55	13.70	0.41$_{\pm 0.01}$	56.6
no alignment loss	5.98$_{\pm 0.38}$	-19.92	7.85	14.93	0.64	0.47	0.03$_{\pm 0.01}$	31.4
character-similarity baseline				21.81		0.26		
Procrustes, same links				26.34		26.41		
Procrustes, same links, CSLS				27.06		27.12		

Table 5: Training budget, scripture, BLI@1 in per cent, means over 3 seeds with the standard deviation over them as a subscript; then, at the longer budget, the Latin-script score of the same configuration trained on the plain and on the ciphered corpus and the fraction kept, as in Table 1. Figure 2 draws the two budgets as arrows.

	8,000 steps	16,000 steps	Latin	ciphered	kept
contrast	10.09$_{\pm 0.23}$	10.66$_{\pm 0.15}$	14.35	10.10	0.70$_{\pm 0.09}$
VICReg, tuned	8.19$_{\pm 0.39}$	6.82$_{\pm 0.19}$	8.58	5.58	0.65$_{\pm 0.07}$
SimSiam, tuned	8.52$_{\pm 0.18}$	12.10$_{\pm 0.29}$	16.12	8.48	0.53$_{\pm 0.03}$

B.3 The sentence-level term

Every aligned configuration of the testbed carries a sentence-level InfoNCE term beside its word-level loss: a record's aligned halves are averaged over its words, and the record must be nearer its translation than the other records of the batch, at weight 0.1 against unity for the word-level loss. The term is held fixed across configurations, and it is not detached, so its gradient reaches the encoder through the record mean. A reviewer asked whether it acts as an anti-collapse anchor on the aligned half and so confounds the comparison of objectives without word-level negatives. Trained with the term alone, the aligned half scores 5.29% on the Latin-script languages of scripture and 16.09% on TED2020, beside 5.52% and 14.93% for the control, at effective ranks 31.3 and 35.8 where the control has 33.2 and 31.4: on its own it neither aligns words nor spreads them beyond what the control has. Table 6 removes it from the six scripture configurations of Table 1, at seed 1337, everything else unchanged.

Contrast keeps 0.72 with the term and 0.71 without, the rank-chosen VICReg 0.15 and 0.11, at the same plain scores and ranks. The tuned VICReg loses plain score, 12.30% to 8.87%, and keeps the same fraction, 0.53 with the term and 0.48 without, at the same rank. SimSiam is where the term mattered. Its ciphered score goes to the floor: the tuned setting keeps 0.01 without the term against 0.34 with it, and the rank-chosen setting 0.03 against 0.26. On plain text the tuned setting loses 3.7 points, 12.87% to 9.13%, at a rank that rises from 55.0 to 60.2; the rank-chosen setting's plain score does not move, 6.96% to 6.61%, and its rank falls to 14.5. The sentence-level negatives were the source of what the predictor objective had learned beyond the cognates. With no negatives anywhere it learns the cognates and nothing else, at a rank that for the tuned setting is higher than before, which is the claim of Section 5.1 in its purest form; it also means the SimSiam kept fractions of Tables 1 and 4 include what the sentence-level negatives supplied. Attraction without the term retrieves 0.57%, under the control, at effective rank 41.1, above the control's 33.2: the rank is that of the centred vectors, so a nearly constant representation with a residual spread scores high while retrieving nothing. This ablation is at a single seed. The encoders of Section 5.4 carry no such term.

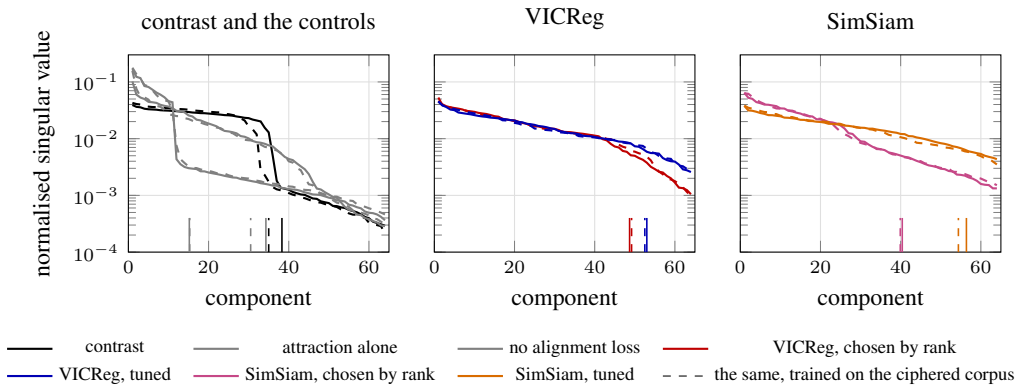


Figure 5: The spectrum the diagnostics read. Normalised singular values of the centred aligned half over the test's word types, scripture, seed 1337, for each configuration of Table 1 trained on the plain corpus (solid) and on the ciphered corpus (dashed); the short vertical mark at the foot of each curve is its effective rank. Effective rank, uniformity and the Wasserstein reading are functions of this curve. For every configuration the ciphered curve lies on the plain one and the two ranks differ by under 4.8 at every seed, while the kept fractions of Figure 2 run from 0.02 to 0.65: the cipher moves the lexicon and leaves the spectrum.

Table 6: The sentence-level term removed. The six scripture configurations of Table 1 at seed 1337, with the term (the paper's runs) and with its weight at zero, otherwise identical: Latin-script BLI@1 on the strict test, plain and ciphered, the fraction kept, and the effective rank of the aligned half, plain and ciphered. The control, which carries no term, scores 5.52% at rank 33.2.

					rank	
	term	Latin	ciphered	kept	plain	ciph.
contrast	with	13.87	9.96	0.72	38	35
	without	12.78	9.09	0.71	36	34
VICReg, chosen by rank	with	8.00	1.22	0.15	53	53
	without	7.35	0.83	0.11	51	51
VICReg, tuned	with	12.30	6.48	0.53	61	61
	without	8.87	4.26	0.48	61	61
SimSiam, chosen by rank	with	6.96	1.78	0.26	39	38
	without	6.61	0.17	0.03	15	8
SimSiam, tuned	with	12.87	4.43	0.34	55	52
	without	9.13	0.13	0.01	60	61
attraction alone	with	1.61	0.22	0.14	11	11
	without	0.57	0.00	0.00	41	36

C Other readings of the test

BLI@1 asks whether the gold partner is the nearest candidate, under a gold induced as the training links were. The readings below each change one thing: the depth of the ranking, the gold, and the task.

C.1 Depth of the gold partner

BLI@1 asks whether the partner is the nearest candidate, which is what a contrastive loss trains. Figure 6 reads the same test at greater depth.

The tuned VICReg has 0.81 of contrast's hits at the first rank, 0.90 within the first ten and 0.97 within the first hundred; the tuned SimSiam has 0.84, 0.90 and 0.96. The advantage of contrast over a tuned device is therefore concentrated at the top of the ranking. The rank-chosen devices do not close the gap: at the hundredth candidate the rank-chosen VICReg holds 0.64 of contrast's hits and the

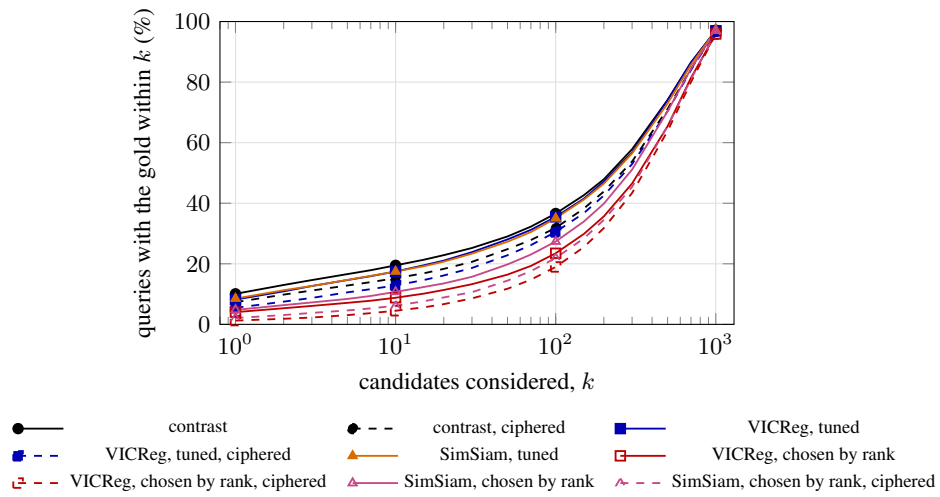


Figure 6: Where the gold partner sits. For each configuration, the share of strict queries whose best gold target is within the first k candidates of its ranking of the strict pool, scripture, pooled over the 3 seeds and the languages; at the first candidate this is BLI@1. Solid: trained on the plain corpus; dashed: on the ciphered corpus, for the configurations whose ciphered ranking was run. The gap between contrast and each tuned device narrows with depth. Under the cipher the rank-chosen devices lose most of their hits at the first rank and little of their share at the hundredth: what the cognates bought was the top of the ranking.

rank-chosen SimSiam 0.75, from 0.40 and 0.47 at the first. That fits the account of Wang & Liu (2021), in which the loss acts through its hardest negatives, and it bounds what our test can say: on a task that does not need the nearest neighbour, the tuned devices may be close to contrast.

C.2 An external gold

The gold lexicon of the main tables is induced from held-out text by the same mutual-best Dice statistic that produces the training links. The leaked pairs are removed, but the generator is shared, so a configuration that fits the aligner's systematic errors is rewarded for them a second time at test. Glavaš et al. (2019) make the wider point that BLI results are comparable only under one dictionary protocol. A dictionary that shares its generator with the training links is the weakest form of that protocol, so Table 7 repeats the comparisons of Tables 1 and 4 against the MUSE dictionaries (Lample et al., 2018), which were built from an internal translation tool and share nothing with the supervision. The candidate pool is every target-side type of the training text, the standard setting, and the queries are the dictionary's English words that the text contains. Per language the strict test has 97 (Japanese) to 3,842 queries, so the Japanese entries carry the widest interval. Everything else is the protocol of Section 3: the same checkpoints, the same seeds, the strict test's conventions, MRR beside BLI@1.

One property of this gold has to be held in mind. MUSE lists many pairs whose two sides are the same string; Kementchedjhieva et al. (2019) show that proper nouns and gaps in the gold, artefacts of the same generator, make dictionary-based BLI scores misleading, and identical strings are another. Copying the query verbatim scores 7.70% on the Latin-script languages of Table 7 against 0.83% on the paper's own gold. An external gold therefore enlarges the share of the test that a spelling solution can answer, and the plain Latin columns of Table 7 are more generous to such a solution than the paper's. The ciphered condition removes exactly this: the dictionary's target side passes through the corpus's cipher, an identical-string pair is no longer one, and the kept fraction is again the measurement.

What the external gold shows. Nothing reorders. On scripture the tuned VICReg scores 0.68 of contrast and the tuned SimSiam 0.70, the rank-chosen devices less, and TED2020 keeps that order. Contrast is ahead of every configuration on BLI@1 and on MRR on both corpora, and MRR orders the configurations exactly as BLI@1 does. Under the cipher contrast keeps 0.74 of its Latin score on

Table 7: The comparisons of Tables 1 and 4 against an external gold: the MUSE dictionaries (Lample et al., 2018), which share no generator with the training links, over the whole target vocabulary of the training text as candidate pool, strict reading, means over the seeds. Per language the strict test has 97 (Japanese) to 3,842 queries. Columns as in Table 1; on the ciphered corpora the dictionary's target side passes through the corpus's cipher. MUSE lists many identical-string pairs, beside the artefacts Kementchedjhieva et al. (2019) document: copying the query scores 7.70% on the Latin-script languages here against 0.83% on the paper's gold, so the plain Latin columns are more generous to a spelling solution than the paper's, and the ciphered column is the one that removes it.

	BLI@1	MRR	Latin	other	Latin, ciph.	kept
scripture						
contrast	34.22	38.28	41.47	21.63	30.86	0.74
VICReg, chosen by rank	13.07	16.10	18.81	3.11	2.47	0.13
VICReg, tuned	23.11	27.81	28.93	13.00	16.15	0.56
SimSiam, chosen by rank	15.47	18.55	20.58	6.59	5.67	0.28
SimSiam, tuned	23.90	28.51	30.53	12.38	11.03	0.36
attraction alone	5.80	6.47	8.90	0.42	0.35	0.04
no alignment loss	9.27	11.02	14.59	0.02	0.01	0.00
character-similarity baseline	11.45	14.07	18.04	0.00	0.02	0.00
Procrustes on the same links	36.31	40.85	43.88	23.16	43.88	1.00
Procrustes on the same links, CSLS	38.51	42.52	46.45	24.73	46.45	1.00
TED2020						
contrast	56.69	62.49	63.32	47.48	50.54	0.80
VICReg, chosen by rank	21.60	25.30	34.65	3.43	3.21	0.09
VICReg, tuned	34.83	41.52	44.98	20.71	22.98	0.51
SimSiam, chosen by rank	23.68	27.49	35.26	7.55	3.69	0.10
SimSiam, tuned	38.07	44.90	48.61	23.40	14.64	0.30
attraction alone	10.82	11.92	17.87	1.00	-	-
no alignment loss	15.90	18.12	26.64	0.96	0.01	0.00
character-similarity baseline	18.93	22.12	31.84	0.96	0.01	0.00
Procrustes on the same links	59.10	63.21	65.86	49.70	65.88	1.00
Procrustes on the same links, CSLS	61.25	64.69	68.51	51.15	68.54	1.00

scripture and the tuned VICReg 0.56; the tuned SimSiam, within a few points of the tuned VICReg on plain text, keeps 0.36, close to its own rank-chosen setting, and the rank-chosen devices keep less again on either corpus. The controls behave: the configuration without an alignment loss keeps 0.00, Procrustes does not move, and the character baseline goes to the floor. One thing does change. Under this gold the supervised linear map is above contrast on both corpora, pooled and on the Latin-script languages of scripture, where under the paper's gold it was above on one. The testbed remains a testbed, and the comparison between objectives is unchanged.

C.3 A DOWNSTREAM TEST

Every test so far is nearest-neighbour retrieval. Glavaš et al. (2019) show that BLI predicts downstream transfer poorly and ask that the two be measured together, and Appendix C.1 already found the gap between contrast and the tuned devices closing with retrieval depth. Table 8 is their protocol on the paper's configurations: a natural language inference classifier is trained on English MultiNLI (Williams et al., 2018) over the aligned half's word vectors, one per word type and frozen, with its epoch chosen on the English development set, and applied unchanged to the XNLI test set (Conneau et al., 2018) of each language by swapping in that language's vectors. No labelled data in any language but English is seen. The vocabulary is 145,564 word types; the byte encoder embeds all of them, so the test has no coverage rule. Two task models read the same vectors, a BiLSTM with max-pooling and a bag of vectors. On the ciphered corpora the test sentences of a ciphered language pass through the corpus's cipher.

The devices transfer as well as contrast. With the BiLSTM every device is within 2.5 points of contrast on the Latin-script languages and on the others, on both corpora; with the bag of vectors every device is above it, by 1.8 to 4.1 points on the Latin-script languages. Attraction alone and the control are well below every aligned configuration on both, so alignment is what transfers; the

Table 8: Zero-shot cross-lingual NLI with the configurations' word vectors frozen, the downstream test of Glavaš et al. (2019): a classifier is trained on English MultiNLI (392,702 pairs) over one vector per word type from the aligned half, its epoch chosen on the English XNLI development set, and applied unchanged to the XNLI test set (Conneau et al., 2018) of each language (5,010 pairs; 7 of the 8 languages, Japanese is not in XNLI). Accuracy in per cent, means over the seeds; chance is 33.3. Columns: a BiLSTM with max-pooling on English, on the Latin-script languages and on the others; the Latin-script languages read through the same configuration trained on the ciphered corpus, whose test sentences pass through the corpus's cipher; "kept" is the ciphered margin above chance over the plain one, per seed; then a bag of vectors on the same three tests. No labelled data in any language but English is seen.

	BiLSTM					bag of vectors			
	English	Latin	other	Latin, ciph.	kept	Latin	other	Latin, ciph.	kept
scripture									
contrast	64.48	52.79	47.38	52.11	0.97	46.98	45.42	46.00	0.93
VICReg, chosen by rank	63.92	52.79	47.30	51.71	0.94	50.16	46.62	48.24	0.89
VICReg, tuned	65.11	54.54	49.01	53.10	0.93	51.05	47.93	48.94	0.88
SimSiam, chosen by rank	64.05	54.83	49.91	52.87	0.91	49.01	46.02	47.10	0.88
attraction alone	51.71	43.88	40.73	42.95	0.91	37.81	35.73	37.80	1.01
no alignment loss	62.09	48.65	39.97	40.87	0.49	39.62	34.65	34.99	0.26
TED2020									
contrast	65.47	56.63	51.33	56.75	1.01	49.17	47.67	47.81	0.92
VICReg, chosen by rank	64.39	55.58	51.95	53.00	0.88	50.97	48.64	48.86	0.88
VICReg, tuned	65.33	57.71	53.02	56.44	0.95	52.46	50.60	51.39	0.94
SimSiam, chosen by rank	64.57	56.29	52.61	55.54	0.97	51.22	49.17	48.99	0.88
attraction alone	50.01	44.08	40.78	-	-	38.89	37.11	-	-
no alignment loss	62.19	48.41	39.41	40.61	0.49	39.17	34.39	34.61	0.22

control's Latin-script score sits 8.7 points above its score on the other languages, and under the cipher that advantage goes (0.9 points, kept 0.49), which is spelling doing what it can for a classifier. The aligned configurations keep it: contrast 0.97 of its margin above chance on scripture and 1.01 on TED2020, and every device at least 0.88. The rank-chosen VICReg, which keeps 0.16 of its BLI score under the cipher, keeps 0.94 of its XNLI margin. Over all configuration-seeds BLI@1 and XNLI accuracy correlate at $\rho = 0.67$ ($n = 18$) on scripture and $\rho = 0.78$ on TED2020, because attraction and the control are last on both; over contrast and the device settings the table carries alone the correlation is $\rho = -0.08$ ($n = 12$) and $\rho = 0.32$. The tuned SimSiam was not run on this task.

This is the finding of Glavaš et al. (2019) reproduced inside one testbed, and it bounds what the cipher test says. A kept fraction is a statement about the nearest neighbour: under the cipher a device's top-ranked candidate stops being the translation, which is what BLI@1 asks. A classifier over frozen vectors uses their geometry more coarsely, and at that grain the devices' alignment is as good as contrast's and as robust to the cipher. What the rank-chosen devices learned in the Latin-script languages was the cognates at rank one. Below it they hold only part of contrast's ranking, 0.64 and 0.75 of its hits at the hundredth candidate (Figure 6), but a geometry that a classifier reads as well as contrast's, which Table 8 shows on a task. What that space is worth to a nearest-neighbour use is read on the standard encoder in Table 15: out of domain the devices keep 0.01, 0.29 and 0.03 of their score under the cipher, so the rank-one loss the classifier does not see is most of what those configurations transfer to bitext mining.

D SELECTION IN FULL

This appendix backs Section 6: the development half that stands in for the whole-test selection of Section 5, and the criteria that Figure 3 leaves out.

Table 9: The comparisons of Tables 1 and 4 restricted to the gold lexicon's test half, the queries that played no part in choosing a setting (Appendix D.1). Means over the seeds in brackets. "Dev." is BLI@1 on the development half; every other column is on the test half: BLI@1, its fraction of contrast's, the Latin score plain and ciphered, the kept fraction, and the largest p over seeds of contrast's paired advantage. The SimSiam setting that placed second on the development half is shown as well.

	dev.	BLI@1	frac.	Latin	Latin, ciphered	kept	p
scripture							
contrast [3]	10.14	10.05	1.00	14.59	8.82	0.61	
VICReg, chosen by rank [3]	4.11	3.98	0.40	7.83	1.30	0.17	<0.001
VICReg, tuned [3]	7.89	8.46	0.84	12.04	6.47	0.54	0.088
SimSiam, chosen by rank [3]	4.56	4.98	0.50	7.27	2.49	0.34	<0.001
SimSiam, tuned [3]	8.55	8.49	0.85	12.69	4.30	0.34	0.031
no alignment loss [3]	2.38	2.49	0.25	5.88	0.06	0.01	<0.001
SimSiam, second on the development half [3]	7.39	7.79	0.78	11.96	3.45	0.29	<0.001
TED2020							
contrast [3]	25.80	26.00	1.00	35.12	28.77	0.82	
VICReg, chosen by rank [3]	10.83	9.63	0.37	20.32	4.23	0.21	<0.001
VICReg, tuned [3]	19.98	20.19	0.78	29.73	19.82	0.67	<0.001
SimSiam, chosen by rank [3]	11.03	11.96	0.46	20.94	4.93	0.24	<0.001
SimSiam, tuned [3]	21.82	22.50	0.87	32.88	13.27	0.40	<0.001
no alignment loss [3]	6.22	5.75	0.22	14.26	0.50	0.03	<0.001

D.1 A DEVELOPMENT HALF

The tuned settings of Section 5 were chosen by their score on the whole strict test. To measure what that choice was worth to the devices, we split the test's queries in two after the fact. A query is an English word type, and its half is decided by a seeded hash of its leading letters. The same word therefore falls in the same half in every language, on both corpora, and with and without the cipher, and the inflected forms of a word do not straddle the halves. The candidate pool is not split. No model is retrained or rescored: the score on either half is a sum over a subset of the per-query hits the test already recorded. Scripture has 2,538 development and 2,857 test queries over the 8 languages, TED2020 2,010 and 2,126.

Selection. Among the 21 VICReg settings on scripture, and among those run on TED2020, the development half selects the setting the tables call tuned. Among the 20 SimSiam settings it selects loss weight and multiplier 3, $\times 100$ (8.55% on the development half), the setting the tables call tuned; second is 2, $\times 100$ (8.04%), and the two are not distinguishable on that half ($p = 0.34$). The selected setting is the best single setting reported in Section 5. It was first run at a single seed on plain text; its other seeds, its ciphered runs, and its runs on TED2020 and at the longer budget were trained after the split, with nothing else changed, and it has 3 seeds under both conditions. At the shared seed the two score 8.82% and 7.98% on the test half.

The comparisons on the test half. Nothing reorders (Table 9). The tuned VICReg has 0.84 of contrast's score on scripture and the tuned SimSiam 0.85, the rank-chosen devices less, and TED2020 keeps the order. Under the cipher contrast keeps 0.61 of its Latin score and the tuned VICReg 0.54, against 0.34 for the tuned SimSiam, which its rank-chosen setting matches; TED2020 separates the same families. Contrast is ahead of both tuned devices in every seed. On fewer queries the test is weaker: over the tuned SimSiam, and over the tuned VICReg on TED2020, every seed remains significant ($p \leq 0.031$), and over the tuned VICReg on scripture the differences are 0.9 to 2.4 points, of which the smallest is not ($p = 0.088$).

D.2 THE CRITERIA IN FULL

Figure 7 draws the criteria Figure 3 leaves out, over the same settings: uniformity and its Wasserstein refinement, read in Section 6.1, the validation retrieval the training loop reports, over a small sample and a small pool, and training-link retrieval restricted to the pairs a spelling cannot solve. Table 10

Table 10: Label-free selection criteria over 41 non-contrastive settings (21 VICReg, 20 SimSiam; scripture, a single seed). For each criterion: its Spearman correlation with strict BLI@1 and the regret of the setting it selects, in points of BLI@1; then the same against the non-Latin languages alone, where no spelling shortcut exists. The interval beside each ρ is a bootstrap over the settings (4,000 resamples, the central 95% interval); each setting is a single run. None of the criteria reads the test. Figure 3 draws the first two columns.

	all languages			non-Latin		
	ρ	interval	regret	ρ	interval	regret
effective rank	+0.84	+0.67 to +0.93	4.47	+0.75	+0.51 to +0.91	5.20
uniformity	+0.59	+0.31 to +0.80	4.47	+0.36	+0.04 to +0.64	5.20
Wasserstein uniformity	+0.79	+0.60 to +0.90	0.00	+0.60	+0.34 to +0.79	0.84
validation retrieval, non-Latin	+0.23	-0.11 to +0.51	2.35	+0.32	-0.04 to +0.61	2.91
training-link retrieval	+0.96	+0.92 to +0.98	0.46	+0.98	+0.93 to +0.99	0.32
training-link retrieval, no-shortcut pairs	+0.91	+0.83 to +0.95	0.46	+0.97	+0.91 to +0.99	0.32

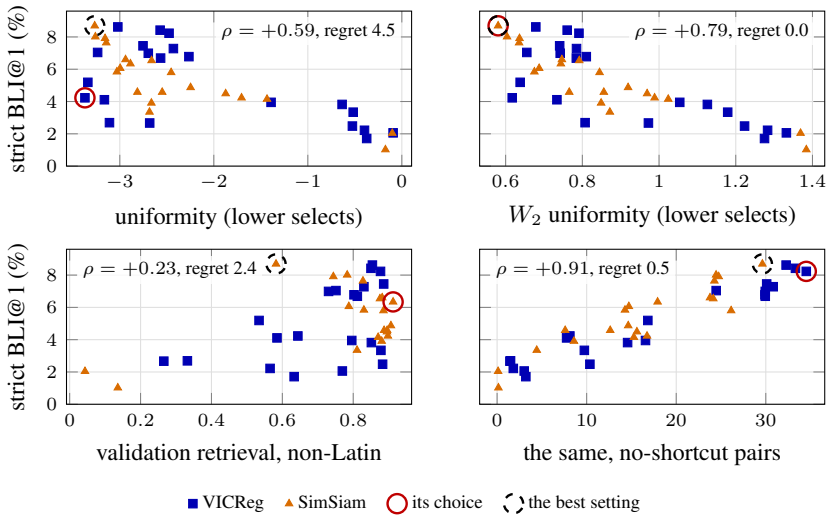


Figure 7: The remaining criteria, drawn as in Figure 3: uniformity and its Wasserstein refinement, read in Section 6.1; the validation retrieval that the training loop reports, over a small sample and a small pool; and retrieval of the training links restricted to the pairs a spelling cannot solve. In the corner, the Spearman correlation with strict BLI@1 and the regret of the selection; Table 10 has the same numbers.

has their numbers beside the others. Neither improves on the criterion of Section 6.1: restricted to those pairs it correlates with the test at $\rho = 0.91$ against 0.96, and the training loop's own retrieval, at $\rho = 0.23$, is too noisy to select with. The large pool is what makes the criterion work.

The criteria against the test half. A development half is itself a selection criterion, the supervised one, and it can be set beside the label-free criteria of Section 6.1 on the half it did not see. It correlates with the test half at $\rho = 0.96$ and its choice lies 0.2 points below the best setting. Training-link retrieval reaches $\rho = 0.98$ with a regret of 0.0, and effective rank $\rho = 0.81$ with a regret of 4.9. A criterion that reads no gold pair selects about as well as one that reads gold pairs.

E THE STANDARD ENCODER

This appendix backs Sections 5.4 and 6.2: the word-level protocol on ByT5-small and its selection, the sentence probe, a second encoder and a second corpus, an external sentence test, and the predictor objective's behaviour on this encoder.

Table 11: The testbed's word-level protocol with ByT5-small's encoder in place of the byte encoder, scripture, the test queries of the strict test, means over 3 seeds. Settings were chosen on the development half. "Kept" is the mean of the per-seed ratios; a configuration whose plain Latin score is under 2 times that of the untrained encoder in every seed has no kept fraction ("at floor"). Rank is the effective rank and "train" the retrieval of training links, each of the model trained on the plain corpus and of the one trained on the ciphered corpus.

							rank		train	
		Latin	ciphered	kept	other	other, ciph.	plain	ciph.	plain	ciph.
random init.	contrast	11.48	8.82	0.77	6.60	6.99	149	148	80	76
	VICReg, tuned	11.51	8.51	0.74	6.08	6.46	158	156	77	74
	VICReg, chosen by rank	8.26	5.68	at floor	4.01	4.55	171	172	32	30
	SimSiam, tuned	8.06	0.06	at floor	3.16	0.14	183	182	51	1
	untrained	5.34	0.17		0.12	0.12	143	146	8	0
pretrained	contrast	19.82	13.74	0.69	8.90	7.93	55	56	90	87
	VICReg, tuned	8.74	3.87	0.44	3.64	3.14	128	125	30	24
	VICReg, chosen by rank	8.11	3.62	0.45	3.75	3.12	131	129	28	23
	SimSiam, tuned	5.00	0.28	at floor	3.20	0.34	106	61	26	1
	untrained	3.31	0.08		0.12	0.12	144	144	6	0
	Procrustes, same links	22.90	22.90		15.26	15.26				

E.1 WORDS

Words. Table 11 has every configuration of the word-level protocol of Section 5.4. The learning rate was chosen on contrast and each device's setting among 20 and 24 candidates, all on the development queries (Appendix D.1); the table reports the test queries. We report no kept fraction for a configuration whose plain score is under 2 times the untrained encoder's, a rule fixed before the runs. With pretrained weights the devices reach 0.43 and 0.29 of contrast's score, against 0.97 for VICReg from random initialisation, and VICReg keeps 0.44 where contrast keeps 0.69; the VICReg setting chosen by rank keeps 0.45, so the two VICReg rows do not differ here.

SimSiam behaves differently from every other configuration. From random initialisation, under the cipher, its score is 0.06 on the Latin-script languages and falls from 3.16 to 0.14 on the other languages, whose text the cipher does not touch, in every seed and from either initialisation. Its effective rank is 183 trained on the plain corpus and 182 on the ciphered one. Retrieval of its own training links is 51 per cent against 0.6.

Selection. Over the device settings searched for Table 11, effective rank correlates with the test half at $\rho = 0.63$ from random initialisation and $\rho = 0.51$ with pretrained weights, and the setting it selects is 3.5 and 4.7 points below the best. Retrieval of the training links has $\rho = 0.89$ and $\rho = 0.95$ and selects within 0.1 and 0.1 points. Uniformity is not blind in every case: from random initialisation it has $\rho = 0.84$, and its pick is still 2.1 points short.

E.2 THE SENTENCE PROBE

The sentence probe. Table 12 is the first probe of the sentence level, at 1 seed, no tuning, 1,500 steps; a batch is 96 record pairs, and the pool is about 266 candidates of the record's language. The untrained encoder retrieves 4.2 per cent on the Latin-script languages and contrast 99.1. The cipher leaves contrast where it was (98.6). SimSiam falls from 88.5 to 13.0, and from 88.8 to 13.7 on the languages the cipher does not touch. The VICReg setting that is best on plain text falls from 94.1 to 0.4, while a weaker setting keeps 0.81 of its score. The diagnostics do not register any of this. The runs with the highest effective rank, 191 and 184, retrieve 0.6 and 0.4 per cent; contrast has rank 108. Those runs are also the most uniform of the table. The training loss does not flag them either: the VICReg run that learned nothing ends at a loss of 3.46, below the 3.70 of the same setting on plain text. Retrieval of the training pairs is within a few points of the held-out score in every row. From random initialisation this budget is too short to say anything: contrast reaches 34.6 per cent and the devices stay at the floor.

Table 12: Sentence level, pretrained ByT5-small encoder, scripture, 1 seed, untuned, 1,500 steps: retrieval of a held-out record's other side among the held-out records of its language (test half, per cent, mean over languages); then, plain and ciphered, retrieval of training pairs, effective rank and uniformity (lower is more uniform). Weights are variance / covariance.

						train pairs		rank		unif.	
	Latin	ciphered	kept	other	other, ciph.	plain	ciph.	plain	ciph.	plain	ciph.
untrained	4.2	0.7		1.2	1.2	3	1	124	116	−0.5	−0.6
contrast	99.1	98.6	1.00	95.1	94.4	99	99	108	103	−3.8	−3.8
VICReg 1/25	89.6	72.7	0.81	75.9	62.5	93	76	44	50	−0.4	−0.4
VICReg 5/100	94.1	0.4	0.00	81.2	0.3	93	0	52	184	−0.7	−1.2
VICReg 25/400	0.6	0.2		0.3	0.7	0	0	191	190	−4.0	−4.0
SimSiam, predictor ×100	53.6	1.4	0.03	35.0	1.8	55	2	60	50	−2.1	−2.1
SimSiam, predictor ×1000	88.5	13.0	0.15	88.8	13.7	88	13	87	65	−1.4	−2.1

Table 13: The sentence-level setting of Table 12 with the devices in their published form: pretrained ByT5-small, scripture, 2 seeds, 800 steps of 256 record pairs, untuned. "Head" is the map from the encoder output to the loss, the linear map of Table 12 or the published one (a projector for contrast, VICReg's expander, SimSiam's projector and predictor). The test is the hard pool read at the encoder output, where the published methods evaluate: retrieval of a held-out record's English side among every held-out English record of the test half, per cent, means over the seeds. "Kept" is the ciphered score over the plain one per seed, as its range over the seeds; a configuration that retrieves under 50 per cent of its own training pairs in every seed and on both corpora is marked collapsed. Training-pair retrieval, effective rank and uniformity are read at the encoder output, plain and ciphered.

							train pairs		rank		unif.	
	head	Latin	ciphered	kept	other	other, ciph.	plain	ciph.	plain	ciph.	plain	ciph.
contrast	linear	96.7	94.3	0.97–0.98	91.7	90.2	100	100	798	764	−2.5	−2.5
contrast	published	96.2	93.6	0.97	90.9	89.6	99	99	772	744	−1.5	−1.6
VICReg 1/25	linear	91.5	75.9	0.76–0.90	79.2	72.2	97	90	794	775	−3.2	−2.7
VICReg 5/100	linear	0.2	0.1	collapsed	0.1	0.1	1	0	1281	1257	−3.7	−3.6
VICReg, lr 3×10^{-4}	published	93.6	90.8	0.96–0.98	85.8	82.0	98	98	656	633	−2.6	−2.6
VICReg, lr 1×10^{-4}	published	92.2	81.3	0.88	80.4	74.7	97	94	535	521	−2.0	−1.9
VICReg, width 8,192	published	95.0	94.0	0.99	88.4	86.8	98	98	780	765	−2.5	−2.4
SimSiam, predictor ×1000	linear	84.5	16.6	0.01–0.35	56.1	21.5	92	42	717	698	−1.3	−1.8
SimSiam, predictor ×1	published	85.8	1.9	0.02–0.03	57.7	3.5	95	11	644	641	−1.8	−2.0
SimSiam, predictor ×10	published	94.1	23.1	0.25	78.9	28.0	99	73	719	719	−1.7	−2.0
SimSiam, predictor ×100	published	95.0	42.5	0.36–0.53	82.1	42.0	99	88	726	731	−1.6	−2.0
SimSiam, predictor ×10, lr 1×10^{-4}	published	92.9	1.4	0.01	73.0	12.3	98	12	681	709	−1.3	−1.4

Table 13 repeats the setting at 2 seeds with the published heads; the collapsed configuration there ends at the same loss plain and ciphered, so the loss comparison above has no counterpart in that batch. Published VICReg keeps 0.88 or more there in every setting and seed, so the fall of the best VICReg setting above belongs to the linear head at those weights and not to the objective.

E.3 A SECOND ENCODER AND A SECOND CORPUS

The second encoder. Table 14 repeats the sentence-level comparison with XLM-R base, a subword encoder, at 2 seeds and with the devices in their published form, as in Table 13. The ordering is ByT5's. Contrast keeps 0.92 to 0.97 of its Latin-script score on the *hard pool*, every held-out English record as candidates, at its two faster learning rates and 0.74 at the slowest; published VICReg keeps 0.84 to 0.92 in every setting. SimSiam keeps a fraction that rises with its predictor's rate, 0.20 to 0.43 at ×10, 0.57 to 0.60 at ×100 and 0.74 to 0.79 at ×1,000, with the seeds agreeing except at the slowest. One SimSiam setting is unstable on plain text, 7.1 and 72.2 across its seeds against contrast's 95.9, while every seed trains under the cipher (79.8); the cipher is not what it is sensitive to. Two seeds and one corpus do not make this a study, but the second encoder does not disagree with the first. Figure 8 is its SimSiam rows against the predictor's rate, the shape of the distillation panel of Figure 4.

The second corpus. Every row of Tables 14 and 17 is scripture. The configurations of Table 17 without the language-model term were also run on TED2020, plain and ciphered, at 2 seeds, with the

Table 14: A second encoder at sentence level: XLM-R base (Conneau et al., 2020a), pretrained, on scripture, 2 seeds, 800 steps of 256 record pairs, the devices in their published form (projection heads, and for SimSiam a predictor; every VICReg and SimSiam row). Plain and ciphered inputs hold the same bytes before tokenisation. "Hard" is retrieval of a held-out record's English side among every held-out English record of the test half, read at the encoder output where the published methods evaluate; "easy" is Table 12's test, the other side among about 266 candidates of its language, read in the head's output space where the loss is computed. Means over the seeds; "kept" is the ciphered score over the plain one per seed, printed as its range over the seeds, and blank where the plain score is not above 2 times the untrained encoder's. The last SimSiam row is unstable on plain text, its seeds 65 points apart, so its kept fraction means nothing.

	hard pool, encoder output					easy pool, head output		
	Latin	Latin, ciph.	kept	other	other, ciph.	Latin	Latin, ciph.	kept
untrained	24.2	0.1		3.7	3.7	20.0	0.9	
contrast, lr 2×10^{-5}	96.0	71.5	0.74–0.75	89.3	86.4	99.1	92.5	0.93–0.94
contrast, lr 5×10^{-5}	96.1	88.4	0.92	90.6	88.9	99.0	97.4	0.98
contrast, lr 1×10^{-4}	95.9	91.7	0.94–0.97	90.4	89.4	99.1	97.8	0.99
VICReg, lr 5×10^{-5}	92.2	77.9	0.84	82.1	78.8	96.8	89.4	0.92–0.93
VICReg, lr 1×10^{-4}	93.1	85.1	0.91–0.92	83.0	81.4	96.9	91.8	0.95
VICReg, lr 5×10^{-5}, width 8,192	94.2	85.0	0.90–0.91	85.9	83.3	97.7	94.4	0.97
SimSiam, predictor $\times 10$	90.0	28.5	0.20–0.43	74.8	54.3	94.9	65.4	0.64–0.74
SimSiam, predictor $\times 100$	93.0	54.4	0.57–0.60	82.8	76.8	97.0	84.8	0.87–0.88
SimSiam, predictor $\times 1000$	94.1	72.0	0.74–0.79	84.5	82.2	97.1	91.7	0.94–0.95
SimSiam, predictor $\times 100$, lr 1×10^{-4}	39.6	79.8	1.08	26.0	78.0	50.1	92.1	1.03

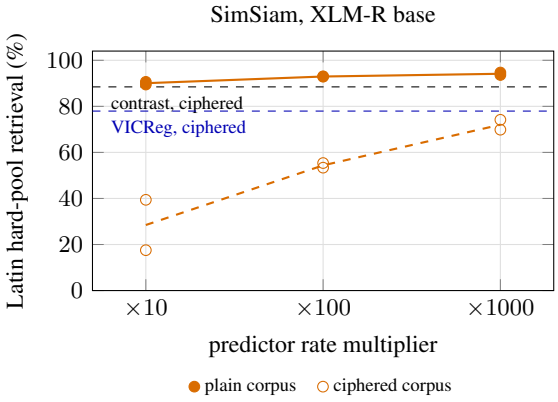


Figure 8: SimSiam on XLM-R base (Table 14) against its predictor's rate multiplier: retrieval on the hard pool of the Latin-script languages, trained on the plain corpus (filled, solid) and on the ciphered corpus (open, dashed), markers the seeds and lines their means; the dashed lines are contrast's and VICReg's ciphered scores on the same encoder at the SimSiam rows' learning rate, 5×10^{-5}, and published width. On plain text the seeds lie within 6 points of each other; the cipher orders the settings, and no seed reaches either reference.

rest of that family; Table 20 in Appendix F.6 reads them. Every trained row scores above 98.1 on plain text against 2.5 for the untrained encoder, so the pool is near its ceiling and the kept fraction is a coarser instrument than on scripture. Contrast keeps 0.99 to 1.00 and VICReg 0.94 to 0.95. SimSiam keeps 0.95 to 0.96 at the faster predictor rate and 0.87 to 0.94 at the slower, where on scripture the same settings keep 0.53 and 0.25; only the slowest predictor, at 0.47 to 0.50, loses the lexicon on this corpus.

E.4 An external sentence test

An external sentence test. The held-out records above are the corpus's own, the *corpus test* below. Table 15 scores the same configurations on the Tatoeba test set of Artetxe & Schwenk (2019): 1000 English-aligned sentences in each of the 8 languages, short and of no fixed register, and retrieval

Table 15: The configurations of Table 12 on an external sentence test: LASER's Tatoeba pairs (Artetxe & Schwenk, 2019), 1000 English-aligned sentences in each of the 8 languages, retrieval of a sentence's translation among that language's 1000 candidates, per cent, mean over languages. The runs are the same settings and seed rerun locally with the external test added; "corpus" is the rerun's own held-out Latin score, for comparison with Table 12. The other columns are Tatoeba: English to the other language on the Latin-script languages plain and ciphered, the kept fraction, the same on the other languages, and the other language to English on the Latin-script languages plain and ciphered. On the ciphered corpus the Tatoeba side in the ciphered language is ciphered with the corpus's key.

	corpus	Latin	Latin, ciph.	kept	other	other, ciph.	Latin, xx→en	ciph.
untrained	4.2	6.4	0.4		0.8	0.8	5.6	0.6
contrast	99.2	50.5	21.4	0.42	36.6	27.6	48.2	16.4
VICReg 1/25	89.5	25.0	7.3	0.29	11.4	9.6	24.8	5.1
VICReg 5/100	93.6	26.2	0.2	0.01	12.9	0.1	26.3	0.1
VICReg 25/400	1.4	0.7	0.2		0.1	0.1	0.5	0.2
SimSiam, predictor $\times 100$	23.1	4.5	0.4		2.1	0.7	4.4	0.3
SimSiam, predictor $\times 1000$	86.2	28.5	0.9	0.03	18.6	2.4	30.3	0.4

of a sentence's translation among that language's 1000 candidates. The runs of Table 12 saved no weights, so the table's rows are the same settings and seed rerun locally with the external test scored after the corpus test; each row's own corpus score is printed beside it so the rerun can be checked against Table 12. On the ciphered corpus the Tatoeba sentences of a ciphered language are ciphered with the corpus's key.

Out of domain the picture sharpens in one respect and changes in another. Contrast retrieves 50.5 per cent of the Latin-script sentences and 36.6 of the others, against 6.4 and 0.8 for the untrained encoder. The devices that learned the corpus reach 0.94, 0.90 and 0.87 of contrast's score on the corpus test and 0.52, 0.50 and 0.57 of it on Tatoeba, 0.35, 0.31 and 0.51 on the other languages. Under the cipher they keep 0.01, 0.29 and 0.03 of their Latin score, so what they transfer out of domain is what the cipher removes. Contrast is the change: on its own corpus the cipher left it where it was (0.99), on Tatoeba it keeps 0.42 of its Latin score and 0.75 of its score on the languages the cipher never touched. Part of that is a weaker shared space, since the unciphered languages fall too; part is Latin-specific. The sentence-level statement of Section 5.4 is therefore a statement about the corpus test: out of domain, contrast keeps a large part of its score where the devices keep almost none, and the ciphered contrast is still well above the best ciphered device, but it is not untouched.

E.5 The predictor objective on the standard encoder

Where the predictor objective stores its alignment. Read through its predictor, the ciphered SimSiam model of Table 12 retrieves 33.4 per cent against 13.0 at the encoder output: part of what it learned is stored in the predictor and not in a space the two languages share.

Run-to-run variance. The Tatoeba reruns of Table 15 put a number on run-to-run variance at this budget: SimSiam with the intermediate predictor rate scores 23.1 on the corpus test there against 53.6 in Table 12 at the same seed, while contrast, the working VICReg settings, SimSiam with the fast predictor and the floor land within a few points of their earlier runs.

SimSiam tuned under the cipher. Table 16 varies the published SimSiam of Table 13 on the ciphered corpus: the predictor's rate multiplier, the encoder's learning rate, the predictor's schedule and the budget, 13 settings at 1 seed. The kept fraction rises with the predictor's rate at both budgets, from 0.71 at best within 800 steps to 0.58, 0.83 and 0.96 at 1600 steps as the multiplier rises, and the longer budget raises it at every multiplier. The best setting, the multiplier $\times 1{,}000$ at 1600 steps, fits 99 per cent of its ciphered training pairs. At the slower encoder rate no multiplier keeps more than 0.02 while every such run scores over 93.7 on plain text, and a decayed predictor schedule does not help. The dependence is therefore the setting's and not the objective's, and the search says something else too: the 11 settings outside the fastest encoder rate score between 93.7 and 95.0 on plain text while keeping between 0.02 and 0.96, so selection on plain text cannot tell them apart, and

Table 16: SimSiam tuned under the cipher: the published SimSiam of Table 13 on scripture, 1 seed, with the predictor's rate multiplier, the learning rate, the predictor's schedule (constant, or decayed with the encoder's) and the budget varied, 13 settings. The test is the hard pool at the encoder output, as in Table 13; "kept" is the ciphered score over the plain one. Training-pair retrieval is plain / ciphered.

predictor	lr	schedule	steps	Latin	Latin, ciph.	kept	other	other, ciph.	train pairs
$\times 10$	3×10^{-4}	constant	800	93.7	23.0	0.25	78.7	25.6	99 / 67
$\times 100$	3×10^{-4}	constant	800	94.8	47.0	0.50	81.4	43.5	98 / 88
$\times 1000$	3×10^{-4}	constant	800	94.9	66.1	0.70	84.1	61.9	99 / 95
$\times 10000$	3×10^{-4}	constant	800	95.0	67.5	0.71	88.0	61.4	99 / 96
$\times 100$	1×10^{-4}	constant	800	93.9	2.0	0.02	77.9	16.9	98 / 14
$\times 1000$	1×10^{-4}	constant	800	94.7	1.8	0.02	81.6	15.0	98 / 18
$\times 10000$	1×10^{-4}	constant	800	94.3	1.9	0.02	81.4	15.7	98 / 22
$\times 10$	1×10^{-3}	constant	800	78.5	3.7	0.05	46.7	4.7	90 / 18
$\times 100$	1×10^{-3}	constant	800	90.5	9.8	0.11	70.8	13.0	97 / 42
$\times 100$	3×10^{-4}	decayed	800	94.6	40.5	0.43	84.4	44.8	98 / 85
$\times 10$	3×10^{-4}	constant	1600	94.2	54.6	0.58	80.5	50.8	99 / 90
$\times 100$	3×10^{-4}	constant	1600	94.4	78.8	0.83	85.7	68.0	99 / 98
$\times 1000$	3×10^{-4}	constant	1600	94.4	90.5	0.96	88.7	78.7	99 / 99

Table 17: An anchored encoder: pretrained ByT5-small at sentence level, scripture, 1 seed, 800 steps, with a masked-byte language-model term beside the pair loss at equal gradient norm (+ *LM*) and without it, the same seed. Retrieval of a held-out record's English side among every held-out English side (test half, per cent, encoder output, mean over languages), plain and ciphered, and the fraction kept; then retrieval of training pairs and effective rank, plain and ciphered.

						train pairs		rank	
	Latin	ciphered	kept	other	other, ciph.	plain	ciph.	plain	ciph.
contrast	96.7	94.7	0.98	91.3	90.3	100	100	796	767
contrast + LM	95.9	94.2	0.98	90.5	90.2	100	100	773	748
VICReg	93.1	90.9	0.98	85.6	81.8	98	98	655	637
VICReg + LM	93.1	90.9	0.98	86.0	81.1	98	98	629	598
SimSiam, predictor $\times 10$	93.7	23.0	0.25	78.7	25.6	99	67	718	708
SimSiam, predictor $\times 10$ + LM	92.9	26.8	0.29	76.6	31.6	98	54	684	650
SimSiam, predictor $\times 100$	94.8	50.5	0.53	81.4	48.1	98	91	729	728
SimSiam, predictor $\times 100$ + LM	93.3	46.4	0.50	80.9	45.3	99	76	699	669

the development half picks the multiplier $\times 10{,}000$ at 800 steps, which keeps 0.71. One seed on one corpus: what this establishes is that the good region exists, not where its edges are.

F AN ANCHORED ENCODER

The encoders of Section 5.4 are trained by the alignment loss alone. A cross-lingual encoder in practice is anchored by something else as well: a language-model loss beside the pair loss, or, in the negative-free recipe that is used in practice, a frozen teacher whose vectors the student must match (Reimers & Gurevych, 2020). Table 17 is a first probe of the first case: the pretrained ByT5-small encoder at sentence level, in the protocol of Table 14 (Appendix E.3), with masked-byte prediction beside the pair loss. The term's gradient is scaled to the pair gradient's norm on the encoder at every step, so that it is the same fraction of the update for every objective; unscaled, it swamps SimSiam's pair gradient and the comparison is one of scales, not of terms. The rows without the term are the same configurations at the same seed.

F.1 THE LANGUAGE-MODEL TERM

The language-model term. The term changes nothing that the cipher measures. Contrast keeps 0.98 of its Latin-script score with the term and 0.98 without; VICReg 0.98 and 0.98; SimSiam 0.29 and 0.25 at predictor multiplier $\times 10$, 0.50 and 0.53 at $\times 100$. No kept fraction moves by more than 0.04 and no plain score by more than 1.5 points, while both seeds of SimSiam at $\times 100$ without the term differ by 0.17. At this weight and budget a language-model term neither rescues the predictor objective nor harms the others.

F.2 A FROZEN TEACHER

The frozen teacher. The second case is distillation to a frozen teacher (Reimers & Gurevych, 2020), the anchor the devices lack by construction. The teacher reads the English side only; the student reads both sides and matches the teacher's vector at the encoder output by mean squared error. English is never ciphered, so the teacher's space is the same under both corpora, and the cipher changes only what the student must map onto it. Table 18 runs it in the setting of Table 17 at 2 seeds with a trained and with an untrained teacher: a contrast run of the same script on plain scripture, an English sentence space in the role the recipe gives its English teacher, and the pretrained encoder as is. The learning rate is chosen on the plain-scripture development half.

Section 7 reports the trained teacher: the development half selects 1×10^{-3}, where the student keeps 0.98; a decade slower it keeps 0.54 to 0.55, with the plain scores of every rate within 0.8 points of one another. Out of domain the selected student is at or above its teacher on Tatoeba and keeps 0.34 to 0.35 of that under the cipher, less than the contrast of Table 15. With the untrained teacher the student reaches at most 14.9 on the hard pool and keeps at most 0.14: the masked-language-model space is a poor target, and the little that was learned was spelling. XLM-R from its own pretrained space, at 1 seed, is well above the floor and keeps 0.24 to 0.51, more at the faster rate.

The trained teacher above is a run of this paper. It reads the English side only and the student starts from the pretrained encoder, so no vector of the other language and no weight passes from teacher to student; but its English space was shaped by contrast against these very pairs, and the recipe as used has a teacher trained apart. The *external* block of Table 18 runs the recipe with all-MiniLM-L6-v2 as the frozen teacher, a 384-coordinate MiniLM encoder (Wang et al., 2020) fine-tuned as a sentence transformer (Reimers & Gurevych, 2019) on English pairs alone (Sentence-Transformers, 2021), which never saw this corpus or any of its other languages; it reads the English side as text, cut at the student's input length in bytes. The widths differ, so the student carries what the other rows lack, a linear head to the teacher's width; the loss is at the head, and the scores, as everywhere in the table, are at the encoder output; the head gives the same reading. The development half selects 1×10^{-3}, where the student keeps 0.93 to 0.96 and the other languages barely move: the lexicon is learned from a teacher that could not have supplied it. The kept fraction falls with the rate, to 0.14 to 0.15 at 1×10^{-4}, and here, unlike with the trained teacher, plain text sees the difference: the selected rate leads the slowest by 15 points, and the slowest retrieves fewer of its training pairs on plain text. Out of domain the selected student is above the student whose teacher was trained on this corpus on Tatoeba, and keeps 0.32 to 0.34 of that score.

F.3 THE SUBWORD ENCODER AS STUDENT

XLM-R from the same teacher, at 2 seeds and 3 rates, reads otherwise. The development half selects 5×10^{-5}, the slowest, where the kept fraction is 0.06 to 0.08; no other rate keeps more than 0.35 at the encoder output in a seed that trained, and at the fastest rate 1 seed diverges. What the subword encoder learned from a fixed English target at this budget was, at every rate, mostly the cognates, though the same encoder keeps the lexicon under contrast (Table 14) and keeps 0.51 from its untrained teacher. Plain text gives a faint sign: at the selected rate the student retrieves fewer of its training pairs than every other row that trained, and far fewer under the cipher, so the criterion of Section 6.1 flags it faintly on plain text and plainly under the cipher. At the fastest rate the seed that trained keeps 0.76 at the head against 0.35 at the encoder output: there the lexicon, such as it is, lives in the anchored space and not in the encoder's.

Table 18: Distillation to a frozen teacher (Reimers & Gurevych, 2020) in the setting of Table 17: the teacher reads the English side only, the student reads both sides and matches the teacher's vector at the encoder output by mean squared error, so the teacher's space is the same under both corpora. Teachers: a contrast run of the same script on plain scripture (*trained*); all-MiniLM-L6-v2 (Wang et al., 2020; Reimers & Gurevych, 2019), an English sentence encoder trained apart from this corpus (*external*; the widths differ, so the student carries a linear head to the teacher's width, the loss is at the head and the scores stay at the encoder output); and the pretrained encoder as is (*untrained*); ByT5-small at 2 seeds (means; the kept fraction as its range over the seeds), XLM-R base at 1 from the untrained teacher and 2 from the external one, where the fastest rate diverged in 1 seed, which enters the means and blanks the kept cell. The XLM-R external block is repeated at a longer budget, 1600 steps, at the rates that trained; selection was not redone there, so no row of it is marked. The last block is the decoder anchor of LASER (Artetxe & Schwenk, 2019) on ByT5-small at 2 seeds: the sentence vector is the whole memory of the model's own decoder, which must spell the English side from either side of a record; no teacher and no negatives; its last row is the selected rate at a longer budget. Columns as Table 17, then Tatoeba English to the other language, plain and ciphered. The learning rate is chosen on the plain-scripture development half; the row it selects is marked.

						train pairs		rank		Tatoeba	
	Latin	ciphered	kept	other	other, ciph.	plain	ciph.	plain	ciph.	plain	ciph.
contrast (the trained teacher)	96.7	94.7	0.98	91.3	90.3	100	100	796	767	38.7	–
trained teacher											
lr 1×10^{-4}	95.4	51.8	0.54–0.55	84.3	82.6	99	81	747	749	41.7	4.8
lr 3×10^{-4}	95.8	88.1	0.91–0.92	88.8	88.6	99	97	728	727	41.6	9.6
lr 1×10^{-3} (selected)	96.0	94.4	0.98	91.1	90.7	99	99	710	708	44.7	15.4
lr 1×10^{-4}, longer budget	95.7	82.0	0.85–0.86	87.7	86.4	99	96	733	733	41.3	7.5
lr 1×10^{-3}, longer budget	96.0	94.9	0.99	91.9	91.6	99	100	704	701	43.0	16.4
external teacher, trained apart from this corpus											
lr 1×10^{-4}	72.0	10.5	0.14–0.15	38.7	26.0	88	41	649	621	28.3	1.6
lr 3×10^{-4}	89.4	62.8	0.68–0.72	65.6	62.0	97	90	802	792	45.0	5.6
lr 1×10^{-3} (selected)	91.9	87.0	0.93–0.96	75.2	74.6	99	98	945	955	58.8	19.4
untrained teacher											
lr 1×10^{-4}	8.0	0.6	0.06–0.09	0.9	0.6	36	9	763	766	8.1	1.0
lr 3×10^{-4}	9.3	0.9	0.10	1.5	0.9	48	14	758	761	8.8	1.0
lr 1×10^{-3} (selected)	14.4	2.0	0.14	2.7	2.2	58	30	746	749	10.8	1.5
XLM-R, untrained teacher											
lr 5×10^{-5}	77.3	18.4	0.24	38.4	34.5	97	77	420	419	66.8	2.4
lr 1×10^{-4} (selected)	80.9	41.6	0.51	44.5	41.3	98	92	416	415	69.6	4.0
XLM-R, external teacher											
lr 5×10^{-5} (selected)	80.3	5.6	0.06–0.08	36.9	31.4	93	32	285	287	72.0	1.8
lr 1×10^{-4}	81.6	19.4	0.15–0.33	39.8	40.7	96	51	325	328	76.9	4.4
lr 3×10^{-4}	32.2	11.4		12.6	18.1	49	48	362	206	38.3	6.3
XLM-R, external teacher, longer budget											
lr 5×10^{-5}	85.7	31.6	0.36–0.38	44.7	50.7	96	71	314	318	77.6	6.0
lr 1×10^{-4}	78.8	44.8	0.43–0.71	42.1	42.7	98	69	376	374	76.9	9.2
decoder anchor (LASER), no teacher											
lr 1×10^{-4}	4.8	0.3	0.04–0.09	0.3	0.1	22	4	722	702	24.7	1.5
lr 3×10^{-4}	46.6	5.3	0.11–0.12	9.0	4.1	50	16	865	849	35.8	2.8
lr 1×10^{-3} (selected)	80.1	58.8	0.72–0.74	41.6	36.3	92	79	979	971	55.7	15.4
longer budget, same rate	88.9	79.6	0.88–0.91	62.8	57.4	100	98	1054	1049	56.4	22.9
untrained encoder	1.5	–		0.2	–	–	–	691	–	9.7	–

At a longer budget, 1600 steps at the rates that trained and both seeds, every row keeps more than at the shorter, 0.36 to 0.38 at the selected rate; no row keeps over 0.71, and the selected rate has still not fitted its training pairs under the cipher. What a budget beyond that would do was not run.

F.4 THE DECODER ANCHOR

The decoder anchor. The third case is the decoder anchor of LASER (Artetxe & Schwenk, 2019), the other non-contrastive recipe in use for cross-lingual sentence encoders. The sentence vector is the whole memory of ByT5-small's own decoder, which must spell the record's English side, teacher-

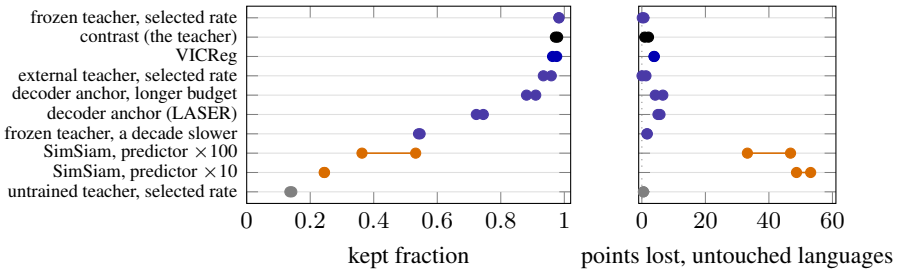


Figure 9: What each anchor keeps. Pretrained ByT5-small at sentence level on scripture, the objectives of Tables 17 and 18 at their selected settings and the same budget (the decoder anchor also at its longer budget), one dot per seed: the fraction of the Latin-script score kept under the cipher, and the points the languages the cipher does not touch lose. The fixed targets, a frozen teacher whether trained on this corpus or apart from it, and contrast itself, and VICReg keep alike; the decoder anchor keeps less and recovers most of the difference with budget; the predictor objective keeps least of the objectives that learned; the untrained teacher keeps nothing. Contrast and the frozen teacher leave the untouched languages where they were; VICReg and the decoder anchor lose a few points there; the predictor objective loses most of its score in the untouched languages as well.

Table 19: The target-momentum sweep. Distillation as in Table 18 with the teacher a moving average of the student at momentum m, initialised at the trained contrast run, with the frozen teacher at one end and the student under stop-gradient at the other. The loss (one minus cosine), a predictor on the student side, the rate, the budget, the data and the seeds are the same in every row, so m is the one variable. ByT5-small at 2 seeds (means; the kept fraction as its range over the seeds, none when the plain score is under the floor rule of Table 18). Columns as Table 17, then retrieval of the training pairs over the pool of every language and uniformity, each plain and ciphered.

						train pairs		rank		train, all languages	uniformity	
m	Latin	ciphered	kept	other	other, ciph.	plain	ciph.	plain	ciph.	plain	ciph.	plain ciph.
1	96.0	95.2	0.99–1.00	91.5	91.7	100	100	892	887	97.8	97.7	-2.49 -2.44
0.9999	96.2	95.5	0.99–1.00	91.6	91.5	100	100	894	890	97.9	97.7	-2.49 -2.44
0.999	96.0	95.3	0.99–1.00	91.1	90.5	100	100	907	904	97.8	97.6	-2.37 -2.32
0.99	94.1	84.1	0.88–0.91	73.7	61.1	99	99	943	940	95.5	90.8	-1.83 -1.58
0.9	1.1	0.6	–	0.1	0.4	2	3	85	211	0.5	0.5	-0.10 -0.15
0.5	0.4	0.2	–	0.1	0.1	1	1	14	81	0.4	0.2	-0.13 -0.43
0	0.3	0.1	–	0.0	0.0	1	1	28	15	0.2	0.1	-0.28 -0.06

forced, from either side of the record: the other side translates and the English side reconstructs through the one-vector bottleneck, and the shared target is what pulls the two together, as in LASER, where every input language decodes into the same target languages. There is no teacher, no predictor and no negative. The last block of Table 18 runs it in the same setting at 2 seeds, the learning rate chosen on the development half. It selects 1×10^{-3}, where the run keeps 0.72 to 0.74: more than the predictor objective at either of its rates in Table 17, less than the distilled student at its selected rate. Two things separate it from the fixed teacher. The languages the cipher does not touch fall by 5.1 to 5.7 points, and retrieval of the training pairs is 92 per cent on plain text and 79 under the cipher, so at this budget the decoder is still fitting its pairs and the ciphered run is the slower fit. The last row of the block runs the selected rate for 1600 steps. The training pairs are then fitted under both corpora, and the kept fraction rises to 0.88 to 0.91, while the untouched languages still fall by 4.2 to 6.6 points. Most of the shortfall at the shorter budget was budget; what remains at the longer budget, in the ciphered languages and the untouched ones alike, is a slower fit that more training did not close, and this probe does not say whether more still would. The slower rates do not show distillation's silent failure: plain text separates the selected rate from the others, and the middle rate keeps 0.11 to 0.12. On Tatoeba the selected run is above the distilled student on plain text and keeps 0.25 to 0.31; at the longer budget it keeps 0.37 to 0.44. Figure 9 orders the anchors of this appendix by what they keep and by what the untouched languages lose.

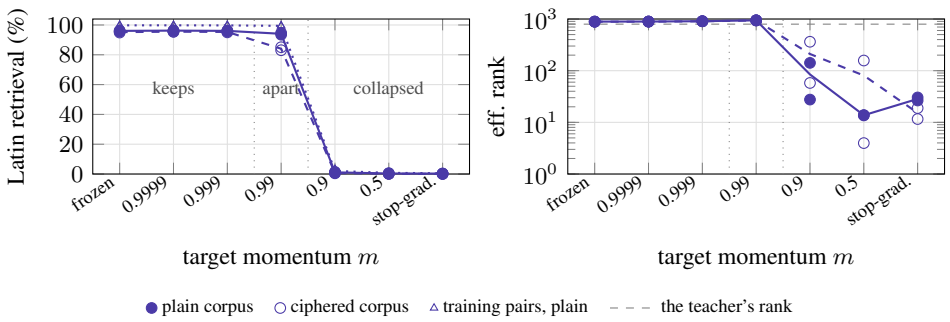


Figure 10: Whether the target may move. Distillation as in Figure 4 with the teacher a moving average of the student, against its momentum from frozen to stop-gradient (Table 19), under a cosine loss with a predictor, so the frozen end is not the selected rate of Figure 4. Left: Latin hard-pool retrieval trained on the plain corpus (filled, solid) and on the ciphered corpus (open, dashed), the plain runs' training-pair retrieval dotted. Right: effective rank at the encoder output, the teacher's dashed. A fast target collapses in view of every diagnostic, and the one momentum between is apart on plain text already.

F.5 Varying the target alone

Varying the target alone. The reading above, that the target's freedom to move is what separates the recipes, compares recipes that differ in loss, architecture and target at once. Table 19 and Figure 10 vary the one thing it names. Distillation is run as in Table 18 with the teacher a moving average of the student at momentum m, initialised at the trained contrast run, with the frozen teacher at one end and the student under stop-gradient at the other. The loss is one minus cosine, so that a moving target is not met by shrinking norms; a predictor on the student side is present in every row, at the predictor rate of Table 17; the learning rate is the one selected for distillation in Table 18 and is not re-tuned per momentum, which would be a second variable. The English side is the target for both sides, so below the frozen end the English side is compared with a copy of itself; that term is what a single-target design costs, and it is one reason the fast end may collapse rather than fail silently. The reading was fixed before the runs: the sweep would carry the paper's claim only if the kept fraction fell with the target's speed while the plain-text diagnostics stayed flat.

It does not. Every momentum at or above 0.999 keeps 0.99 or more in both seeds from a plain score above 96.0. Every momentum at or below 0.9 collapses on plain text, to at most 1.4 with effective rank at most 363, which every diagnostic in the table sees. The one momentum between, 0.99, keeps 0.88 to 0.91 of a plain Latin-script score of 94.1, and plain text already sets it apart: its untouched languages score 73.7 against 91.1 or more at the slower momenta, its uniformity is -1.83 against -2.49, and its effective rank 943 against 892. The training-pair criterion over every language registers it under the cipher, 90.8 against 97.6 or more, as it should; on plain text it is not needed. A target that follows the student either keeps the lexicon or collapses in view; in this design there is no speed at which it learns the cognates silently. The sentence in Section 7 is therefore a reading across the recipes tested and not a controlled result, and the paper states it as such.

F.6 The second corpus and the longer budget

The frozen teacher on TED2020 is the second block of Table 20, in the setting of the devices above it: a contrast run of the same script on plain TED2020 as teacher, the student at the rates of Table 18, 2 seeds. The plain-TED2020 development half puts 1×10^{-4} first and 1×10^{-3} second by 0.2 points, a tie under the rule fixed before the runs: a difference under a point on the development half is a tie, and a tie marks the rate scripture selected. At 1×10^{-3} the student keeps 0.98, and a decade slower 0.39 to 0.40. On plain text the rates are alike in Latin score, in training-pair retrieval and in effective rank; under the cipher the slow student retrieves 74 per cent of its training pairs against 100. On Tatoeba the marked student is above its teacher and keeps 0.47 to 0.49. The frozen teacher's silent failure is on both corpora; the fixed-versus-moving reading of Section 7 is a reading of scripture at this budget.

Table 20: The second corpus at sentence level: pretrained ByT5-small on TED2020 in the setting of Table 17, 2 seeds (means; the kept fraction as its range over the seeds). The first block is the objectives of Table 17 without the language-model term, and SimSiam at a slower predictor rate besides, run alongside the sentence-level configurations of Section 5.4 and not scored on Tatoeba. The second block is distillation to a frozen teacher as in Table 18, the teacher a contrast run of the same script on plain TED2020, the learning rate chosen on the plain-TED2020 development half by the rule of Table 18, a tie by that rule's margin falling to the rate scripture selected; the marked row is that rate. Columns as Table 18.

						train pairs		rank		Tatoeba	
	Latin	ciphered	kept	other	other, ciph.	plain	ciph.	plain	ciph.	plain	ciph.
the objectives of Table 17, TED2020											
contrast	99.9	99.5	0.99–1.00	99.0	98.8	100	100	767	740	–	–
VICReg	98.8	93.5	0.94–0.95	90.3	89.2	100	99	688	651	–	–
SimSiam, predictor $\times 1$	98.3	47.8	0.47–0.50	87.8	66.8	100	78	629	625	–	–
SimSiam, predictor $\times 10$	99.3	90.1	0.87–0.94	92.7	88.2	100	99	674	689	–	–
SimSiam, predictor $\times 100$	99.5	95.0	0.95–0.96	94.0	92.2	100	99	665	699	–	–
distillation, trained teacher (a contrast run on plain TED2020)											
contrast (the teacher, rerun)	99.8	–		99.1	–	100	–	769	–	45.0	–
lr 1×10^{-4}	99.8	39.2	0.39–0.40	97.4	97.0	100	74	743	746	42.6	6.5
lr 3×10^{-4}	99.8	87.0	0.87	98.3	98.5	100	98	726	727	47.7	14.3
lr 1×10^{-3} (selected)	99.8	98.2	0.98	98.4	98.5	100	100	710	709	50.4	24.2
lr 1×10^{-4}, longer budget	99.8	74.5	0.74–0.75	98.0	98.2	100	93	731	733	44.8	10.9
lr 1×10^{-3}, longer budget	99.8	99.1	0.99–1.00	98.6	98.5	100	100	704	702	47.7	24.9
untrained encoder	2.5	–		0.4	–	–	–	707	–	9.7	–

The longer budget. The rows marked "longer budget" in Tables 18 and 20 run the marked rate and the slowest rate for 1600 steps on both corpora at 2 seeds, everything else held. The marked rate keeps 0.99 or more in every seed on both corpora: the longer budget does not undo what was learned. The slow student keeps 0.85 to 0.86 on scripture and 0.74 to 0.75 on TED2020, from 0.54 to 0.55 and 0.39 to 0.40 at the shorter budget, a rise of 0.31 to 0.36; under the cipher it retrieves 96 and 93 per cent of its training pairs against 100 or more for the marked rate, so its fit is still unfinished at this budget too. On plain text the Latin scores of the two rates lie within 0.5 points on each corpus, training-pair retrieval at 99 per cent or above and effective rank between 701 and 733. The languages the cipher does not touch do separate the rates on plain scripture, by 4.2 points, and not on plain TED2020, by 0.6: the silence of plain text in Section 7 is a statement about the Latin-script measures. Part of what the slow student lost at the shorter budget was budget; this probe does not say whether the rest is.