SynPair: Pairing Unpaired Antibody Chains at Billion-Sequence Scale With Contrastive Learning

Oliver M. Turnbull [1] **Charlotte M. Deane** [1]

Abstract

Large-scale antibody sequence datasets, such as the Observed Antibody Space (OAS), contain billions of unpaired heavy (VH) and light (VL) chain sequences but fewer than 0.2% paired sequences, limiting the performance of antibody language models trained on these resources. Existing computational antibody pairing models, such as ImmunoMatch, achieve promising accuracy but rely on computationally intensive cross-encoder architectures, making large-scale synthetic pairing infeasible. Here, we reframe antibody chain pairing as a dense retrieval problem and introduce SynPair, a dual-encoder model trained with contrastive InfoNCE loss that achieves state-of-the-art pairing accuracy while dramatically reducing computational requirements. SynPair can pair the entire unpaired OAS corpus—over 2 billion sequences—in less than 24 hours on standard HPC resources, a task previously computationally intractable. The synthetically paired libraries generated by SynPair closely match naturally occurring antibody pairing distributions, providing the potential for a biologically realistic, massively expanded paired dataset for antibody language model pre-training.

1. Introduction

Antibodies play a crucial role in the immune response and are an increasingly important class of therapeutic (Raybould et al., 2024). They consist of two sets of heavy and light chains with antigen binding mediated by the Fv region of each chain (VH and VL, respectively) (Chiu et al., 2019). The majority of the diversity in antibodies is located in six hyper-variable loops within the Fv region known as complementarity-determining regions (CDRs). The light chain and heavy chain each contain 3 CDR loops (CDRL 1-3 and CDRH 1-3).

Recent advancements in protein language models (PLMs) have been effectively applied to a variety of antibody-specific tasks, including de novo sequence generation (Turnbull et al., 2024; Shuai et al., 2023), sequence optimisation (Olsen et al., 2024), and antibody property prediction (Turnbull et al., 2024; Olsen et al., 2024; Kenlay et al., 2024), and now form an integral part of many antibody discovery pipelines. A large portion of PLM success is their ability to train on large corpora of sequences efficiently. This is particularly important in the case of antibody PLMs due to the very large space of possible antibody sequences, driven by the biological processes of recombination, somatic hypermutation, and random pairing of VH and VL chains (Chiu et al., 2019).

Public datasets such as the Observed Antibody Space (OAS) (Olsen et al., 2022) provide curated sets of antibody sequences and are used for training the majority of publicly released antibody PLMs. OAS contains over 2.2 billion unpaired human VH and VL sequences, but only 3 million paired VH/VL sequences. Models trained on both paired and unpaired sequences outperform unpaired models on downstream tasks (Burbach & Briney, 2025; Olsen et al., 2024; Turnbull et al., 2024; Burbach & Briney, 2023), likely due to the richer sequence representation provided by the paired format. However, unpaired sequences provide a source of diversity that is currently not found in paired sequence libraries.

Therefore, an open question in the field is how to best take advantage of the diversity of unpaired sequences, combined with the richer representation of paired sequences. Current approaches involve finetuning, either in a two-stage approach (Turnbull et al., 2024), or using curriculum learning (Burbach & Briney, 2025; Kenlay et al., 2024). However, both approaches risk catastrophic forgetting of the diversity present in unpaired pertaining data and a lack of transferability from the different data distributions of unpaired to paired sequences. Additionally, this approach introduces issues when integrating structural information, current antibody structure predictors require paired antibody input (Abanades et al., 2023; Ruffolo et al., 2023).

[1] Department of Statistics, University of Oxford, UK. Correspondence to: Charlotte Deane <deane@stats.ox.ac.uk>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

An alternative approach we propose is artificial pairing of unpaired sequences with a computational predictor of VH-VL compatibility. Recent models, including ImmunoMatch (Guo et al., 2025) and Humatch (Chinery et al., 2024) achieve promising pairing accuracy but rely on cross-encoder architectures which make the combinatorial scoring required by large scale pairing computationally intractable.

We reframe antibody pairing as a large-scale dense retrieval problem, inspired by successful dual-encoder approaches from information retrieval research. We present SynPair, a dense encoder model capable of pairing the entire unpaired OAS dataset within hours, not years, while surpassing classification accuracy of current cross-encoder approaches. SynPair allows for over 3 orders of magnitude increase in the paired pre-training data for language models. Additionally, SynPair can efficiently identify alternative VL chain partners from natural repertoires, supporting developability optimization.

2. Background and Related Work

2.1. Cross-encoder pairing models

Current state-of-the-art pairing predictors encode concatenated VH and VL jointly, scoring every candidate pair individually. **ImmunoMatch** (Guo et al., 2025) fine-tunes an antibody-specific transformer (AntiBERTa2 (Barton et al., 2024)) to classify cognate versus random pairs, while **Humatch** uses an aligned representation input to a CNN (Chinery et al., 2024).

Although accurate, the compute scales quadratically: evaluating the unpaired OAS corpus of 1.9×10^9 VH against 3.5×10^8 VL requires $\mathcal{O}(NM)$ forward passes, making exhaustive pairing with the required $\approx 7 \times 10^{17}$ combinations infeasible. Using a single A100 GPU this would take over 500,000,000 years.

2.2. Dense retrieval with dual encoders

Natural-language information retrieval side-steps this bottleneck by representing queries and documents with *separate* encoders whose outputs live in a shared embedding space; nearest-neighbour search replaces exhaustive scoring. The paradigm originated with Siamese networks for semantic similarity (Reimers & Gurevych, 2019), was popularised for open-domain QA by Dense Passage Retrieval (DPR) (Karpukhin et al., 2020), and has since powered large-scale retrieval in vision–language models such as CLIP (Radford et al., 2021). Key to training such models is a *contrastive* objective.

Given a batch of B true VH–VL pairs $\{(h_i, \ell_i)\}_{i=1}^{B}$, the InfoNCE loss (Oord et al., 2019) maximises cosine similarity between cognate embeddings while pushing away all in-batch non-cognates.

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp(\text{sim}(f(h_i), g(\ell_i))/\tau)}{\sum_{j=1}^{B} \exp(\text{sim}(f(h_i), g(\ell_j))/\tau)}$$

2.3. Approximate nearest-neighbour search at the scale of billions

Embedding the entire VL corpus once reduces pairing to nearest-neighbour search for each VH. Exact search over hundreds of millions of vectors remains memory- and compute-heavy, so most systems adopt *Approximate Nearest Neighbour* (ANN) indices. We build upon **Faiss**—a GPU-accelerated library that implements the IVF–PQ family of indices (Douze et al., 2025). IVF partitions the space into n_{list} Voronoi cells; Product Quantisation (PQ) compresses residuals into compact byte codes, yielding sub-millisecond query times with $< 1\%$ recall loss when properly tuned. While ANN search has been applied to protein structural alignment (Hamamsy et al., 2024), SynPair is the first to employ a dual-encoder + ANN pipeline for VH–VL pairing.

3. Method

We cast VH–VL pairing as a *dense-retrieval* problem: given a heavy-chain query h, retrieve the most compatible light chains ℓ from a large corpus. Our solution couples a **dual encoder** trained with a self-supervised contrastive loss (Section 3.2) to a GPU-accelerated approximate-nearest-neighbour (ANN) index (Section 3.4), allowing sub-millisecond search over billions of candidates.

3.1. Datasets

We train on paired sequences taken from OAS, using the same train, validation, and test sets publicly released by Humatch (Chinery et al., 2024). The Humatch set contains both ”true” pairings taken from paired OAS, as well as ”fake” pairings created from unpaired OAS, intended to be biologically unrealistic pairings. We trained on only ”true” pairings, but augmented our test set with fake pairings to increase the difficulty of synthetic pairing.

As a held-out set, we use sequences from donor 3 from the Dieudonné et al. study 2024, which was not trained on by ImmunoMatch nor included in the Humatch dataset. To create the negative dataset we randomly paired VH and VL chains from within the test set, giving an equal number of positive and negative examples. ImmunoMatch provides separate κ and λ models which we used for testing. We used ANARCI (Dunbar & Deane, 2015) to split our test set into κ and λ.

3.2. Dual-encoder architecture

Backbone. Both VH and VL chains are embedded by the same frozen AntiBerta2 model ϕ (token length mean 128 ± 19), as used by ImmunoMatch. The final hidden layer output is averaged across the sequence. The resulting 1536-dimensional sequence representation is then passed to a trainable projection layer.

Chain-specific projection head:

$$\mathbf{z}_H = f(h) = \text{norm}(W_H^{(2)}\sigma(W_H^{(1)}\phi(h))),$$
$$\mathbf{z}_L = g(\ell) = \text{norm}(W_L^{(2)}\sigma(W_L^{(1)}\phi(\ell))),$$

where $W^{(1)} \in \mathbb{R}^{512\times1536}$, $W^{(2)} \in \mathbb{R}^{128\times512}$, σ = GELU, and $\text{norm}(\mathbf{v}) = \mathbf{v}/\|\mathbf{v}\|_2$.

A separate projection head is trained for VH and VL sequences.

Similarity function. Pair compatibility is the cosine similarity $s(h, \ell) = \langle \mathbf{z}_H, \mathbf{z}_L \rangle$.

3.3. Contrastive objective

Given a minibatch of B true pairs $\{(h_i, \ell_i)\}_{i=1}^B$ we apply the InfoNCE loss (Oord et al., 2019):

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{B}\sum_{i=1}^{B}\log\frac{\exp(s(h_i,\ell_i)/\tau)}{\sum_{j=1}^{B}\exp(s(h_i,\ell_j)/\tau)},$$

where τ is a learnable temperature (initial $\ln 0.07$). All in-batch off-diagonals act as negatives, yielding $B(B-1)$ negatives per step without hand-crafted sampling.

Optimisation. We train for 100 epochs with AdamW ($lr = 1\text{e}{-4}$, β=0.9, 0.98, weight-decay 10^{-2}), batch size 512, a cosine-decay schedule, and fp16 mixed precision. We monitored mean reciprocal rank (MRR), and recall@20 for hyperparameter optimisation.

3.4. Approximate nearest-neighbour retrieval

All embeddings were stored as fp16. We build a Faiss IVF–PQ index (n_{list}=4096, m=16 8-bit sub-quantisers). At query time we probe n_{probe}=16 clusters, then scan PQ codes on GPU, returning the top-k light chains.

3.5. Complexity and memory

Encoding cost is $\mathcal{O}(N+M)$ forward passes (one per chain), and ANN search is $\mathcal{O}(\log N)$ per VH query.

3.6. Paired Set Testing

For the ”SynPair Pairings” set we sampled 2000 random VH chains from our test set, and performed a similarity search against test set VL chains. We found that optimum diversity was achieved with top_k=2 and random sampling of k=1,2 for each VH chain. We also discarded any pairings with a similarity score <0.3. For the ”True Pairings” set we used the same set of VH chains combined with their true pairing.

We used ANARCI (Dunbar & Deane, 2015) to determine the mutation rates of V-gene segments and the V-gene allele of our SynPair, True, and Random set. For the germline correlation plot, we counted the frequency of each V-gene allele combination in our SynPair, True and Random set, and subtracted the Random counts from the SynPair and True set.

3.7. Implementation details

All experiments use PyTorch 2.3 and HuggingFace Transformers 4.41. Training, embedding and GPU search run on an NVIDIA A100-80 GB. Code, pretrained weights, and full unpaired OAS SynPair pairings will be released upon publication.

4. Results

4.1. SynPair Achieves State-of-the-Art Paired Classification

First, we investigated SynPair’s classification accuracy for predicting true vs random pairings. We benchmarked performance on our held-out test set containing mature paired sequences not included in the train set of SynPair or ImmunoMatch. Pseudo-negative data was generated by random pairing of VH and VL sequences (see Methods). SynPair achieved ROC–AUC 0.792 versus ImmunoMatch’s 0.697 on the external-donor test set (Fig. 1), demonstrating stronger generalisation beyond the training distribution. This validates SynPair’s state-of-the-art ability to differentiate between true and random pairings.

4.2. SynPair Creates Paired Libraries With Natural Characteristics

Next, we investigated SynPair’s ability to generate novel and realistic VH/VL pairings. As a case study, we embedded and paired sequences from an augmented test set comprising 320,000 true pairs and 320,000 non-cognate (‘fake’) pairings, provided by Humatch, totalling 1.2 million sequences. We included fake pairings within the search space to increase the difficulty of the pairing task and better resemble the full unpaired case.

Embedding all sequences took 7 minutes on a single NVIDIA A100 GPU; subsequent pairing via IVF–PQ indexing took approximately one second. The embedding and retrieval process scales linearly, enabling practical pairing even at unpaired OAS scale. Embedding the 2.2 B-sequence

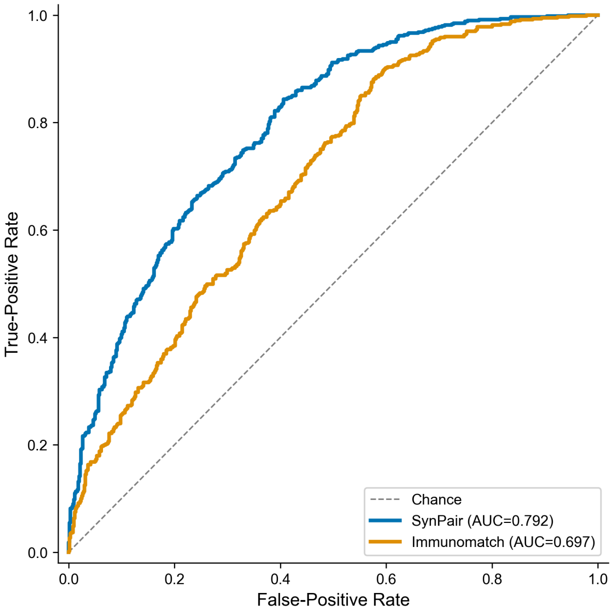

Figure 1. ROC curves on the external-donor test set. ImmunoMatch values are averaged over κ and λ models.

human uOAS corpus would require 9 GPU-days on a single A100, on 32 GPUs this falls below 8 hours. IVF–PQ search for all VH queries completes in <40 minutes.

To validate our generated pairings, we took 2000 synthetically paired sequences and calculated ImmunoMatch pairing scores for this set. We also calculated ImmunoMatch pairing scores for 2000 true pairings, and 2000 VH/VL sequences we randomly paired from the test set (Fig 2). We found that true and SynPair-paired sequences showed a very similar distribution of ImmunoMatch scores. Randomly paired sequences also showed a proportion of higher ImmunoMatch scores, explainable by the known promiscuity of VL/VH pairing. However, random pairs also showed a distinctive lower tail not present in SynPair or True pairings.

We also investigated the correlation of mutation rate across the VL and VH chains. We expect a strong correlation for biologically meaningful pairings, as VH and VL chains should undergo a similar rate of somatic hypermutation when originating from mature B-cells. We observe strong V-gene identity correlation for SynPair pairs (Pearson $r = 0.794$) comparable to true pairs ($r = 0.801$), whereas random pairs show no correlation (Appendix Fig. 3).

Finally, we looked at the frequency of V-gene VH/VL combinations in both True and SynPair paired sequences, relative to a random baseline (Appendix Fig. 4. VH/VL chains have been shown to have weak V-gene pairing preferences (Jayaram et al., 2012). We found a significant Pearson correlation (0.563) between the V-gene pairing frequencies of SynPair and True pairings, relative to the random set. This means that V-gene combinations that are seen more frequently than a random baseline in True pairings are also seen more frequently than baseline in SynPair pairings.

Collectively, our results indicate that our generated pairings are biologically plausible, and also predicted to be good pairings by an independently trained pairing model, Immunomatch.

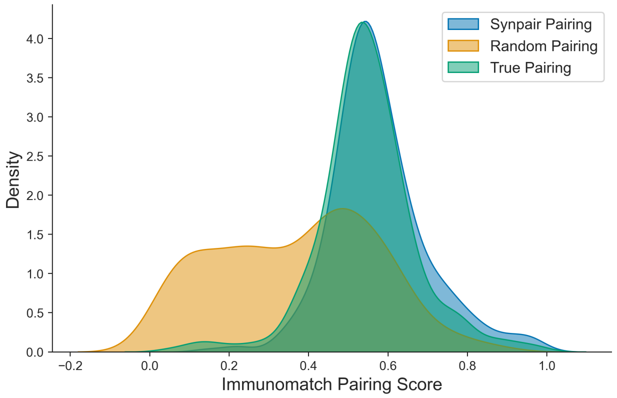

Figure 2. ImmunoMatch pairing scores for true, randomly paired, and synthetically paired VH/VL sequences taken from the test set.

5. Conclusion & Future Work

In this paper, we introduced SynPair, a dual-encoder retrieval model capable of computationally efficient and biologically plausible pairing of antibody heavy (VH) and light (VL) chains at the scale of billions. SynPair exceeds the accuracy of state-of-the-art cross-encoder pairing models, achieving ROC–AUC scores of 0.79 on held-out data, while reducing the computational time required to pair the entire unpaired OAS from an estimated 500 million GPU-years to approximately two GPU-weeks. This advance allows rapid generation of synthetic paired datasets at the scale of the entire Observed Antibody Space (OAS), greatly expanding the availability of paired data for antibody language model pre-training and downstream antibody design applications.

Acknowledgements

This work was supported by funding from the UK Engineering and Physical Sciences Research Council (OT, reference EP/S024093/1) and AstraZeneca.

References

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., and Deane, C. M. ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Communications Biology*, 6(1): 1–8, May 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04927-7. URL https://www.nature.com/articles/s42003-023-04927-7. Number: 1 Publisher: Nature Publishing Group.

Barton, J., Galson, J. D., and Leem, J. Enhancing Antibody Language Models with Structural Information, January 2024. URL https://www.biorxiv.org/content/10.1101/2023.12.12.569610v1. Pages: 2023.12.12.569610 Section: New Results.

Burbach, S. M. and Briney, B. Improving antibody language models with native pairing, August 2023. URL http://arxiv.org/abs/2308.14300. arXiv:2308.14300 [q-bio].

Burbach, S. M. and Briney, B. A curriculum learning approach to training antibody language models, March 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.27.640641v1. Pages: 2025.02.27.640641 Section: New Results.

Chinery, L., Jeliazkov, J. R., and Deane, C. M. Humatch - fast, gene-specific joint humanisation of antibody heavy and light chains, September 2024. URL https://www.biorxiv.org/content/10.1101/2024.09.16.613210v1. Pages: 2024.09.16.613210 Section: New Results.

Chiu, M. L., Goulet, D. R., Teplyakov, A., and Gilliland, G. L. Antibody Structure and Function: The Basis for Engineering Therapeutics. *Antibodies*, 8(4): 55, December 2019. ISSN 2073-4468. doi: 10.3390/antib8040055. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6963682/.

Dieudonné, Y., Lorenzetti, R., Rottura, J., Janowska, I., Frenger, Q., Jacquel, L., Vollmer, O., Carbone, F., Chengsong, Z., Luka, M., Depauw, S., Wadier, N., Giorgiutti, S., Nespola, B., Herb, A., Voll, R. E., Guffroy, A., Poindron, V., Ménager, M., Martin, T., Soulas-Sprauel, P., Rizzi, M., Korganow, A.-S., and Gies, V. Defective germinal center selection results in persistence of self-reactive B cells from the primary to the secondary repertoire in Primary Antiphospholipid Syndrome. *Nature Communications*, 15(1):9921, November 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-54228-8. URL https://www.nature.com/articles/s41467-024-54228-8. Publisher: Nature Publishing Group.

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., and Jégou, H. The Faiss library, February 2025. URL http://arxiv.org/abs/2401.08281. arXiv:2401.08281 [cs].

Dunbar, J. and Deane, C. M. ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics*, pp. btv552, September 2015. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btv552. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv552.

Guo, D., Dunn-Walters, D. K., Fraternali, F., and Ng, J. C. F. ImmunoMatch learns and predicts cognate pairing of heavy and light immunoglobulin chains, February 2025. URL https://www.biorxiv.org/content/10.1101/2025.02.11.637677v1. Pages: 2025.02.11.637677 Section: New Results.

Hamamsy, T., Morton, J. T., Blackwell, R., Berenberg, D., Carriero, N., Gligorijevic, V., Strauss, C. E. M., Leman, J. K., Cho, K., and Bonneau, R. Protein remote homology detection and structural alignment using deep learning. *Nature Biotechnology*, 42(6):975–985, June 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01917-2. URL https://www.nature.com/articles/s41587-023-01917-2. Publisher: Nature Publishing Group.

Jayaram, N., Bhowmick, P., and Martin, A. C. R. Germline VH/VL pairing in antibodies. *Protein engineering, design & selection: PEDS*, 25(10):523–529, October 2012. ISSN 1741-0134. doi: 10.1093/protein/gzs043.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and Yih, W.-t. Dense Passage Retrieval for Open-Domain Question Answering, September 2020. URL http://arxiv.org/abs/2004.04906. arXiv:2004.04906 [cs].

Kenlay, H., Dreyer, F. A., Kovaltsuk, A., Miketa, D., Pires, D., and Deane, C. M. Large scale paired antibody language models, March 2024. URL http://arxiv.org/abs/2403.17889. arXiv:2403.17889 [q-bio].

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Olsen, T. H., Boyles, F., and Deane, C. M. Observed Antibody Space: A diverse database of cleaned, annotated, and translated unpaired and paired antibody sequences. *Protein Science*, 31(1):141–146, 2022. ISSN 1469-896X. doi: 10.1002/pro.4205.

URL https://onlinelibrary.wiley.com/doi/abs/10.1002/pro.4205. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pro.4205.

Olsen, T. H., Moal, I. H., and Deane, C. M. Addressing the antibody germline bias and its effect on language models for improved antibody design, February 2024. URL https://www.biorxiv.org/content/10.1101/2024.02.02.578678v1. Pages: 2024.02.02.578678 Section: New Results.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation Learning with Contrastive Predictive Coding, January 2019. URL http://arxiv.org/abs/1807.03748. arXiv:1807.03748 [cs].

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Raybould, M. I. J., Turnbull, O. M., Suter, A., Guloglu, B., and Deane, C. M. Contextualising the developability risk of antibodies with lambda light chains using enhanced therapeutic antibody profiling. *Communications Biology*, 7(1):1–13, January 2024. ISSN 2399-3642. doi: 10.1038/s42003-023-05744-8. URL https://www.nature.com/articles/s42003-023-05744-8. Number: 1 Publisher: Nature Publishing Group.

Reimers, N. and Gurevych, I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, August 2019. URL http://arxiv.org/abs/1908.10084. arXiv:1908.10084 [cs].

Ruffolo, J. A., Chu, L.-S., Mahajan, S. P., and Gray, J. J. Fast, accurate antibody structure prediction from deep learning on massive set of natural antibodies. *Nature Communications*, 14(1):2389, April 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38063-x. URL https://www.nature.com/articles/s41467-023-38063-x. Publisher: Nature Publishing Group.

Shuai, R. W., Ruffolo, J. A., and Gray, J. J. IgLM: Infilling language modeling for antibody sequence design. *Cell Systems*, 14(11):979–989.e4, November 2023. ISSN 2405-4712. doi: 10.1016/j.cels.2023.10.001. URL https://www.sciencedirect.com/science/article/pii/S2405471223002715.

Turnbull, O. M., Oglic, D., Croasdale-Wood, R., and Deane, C. M. p-IgGen: a paired antibody generative language model. *Bioinformatics*, 40(11):btae659, November 2024. ISSN 1367-4811. doi: 10.1093/bioinformatics/btae659. URL https://doi.org/10.1093/bioinformatics/btae659.

A. Mutation Rate Correlation.

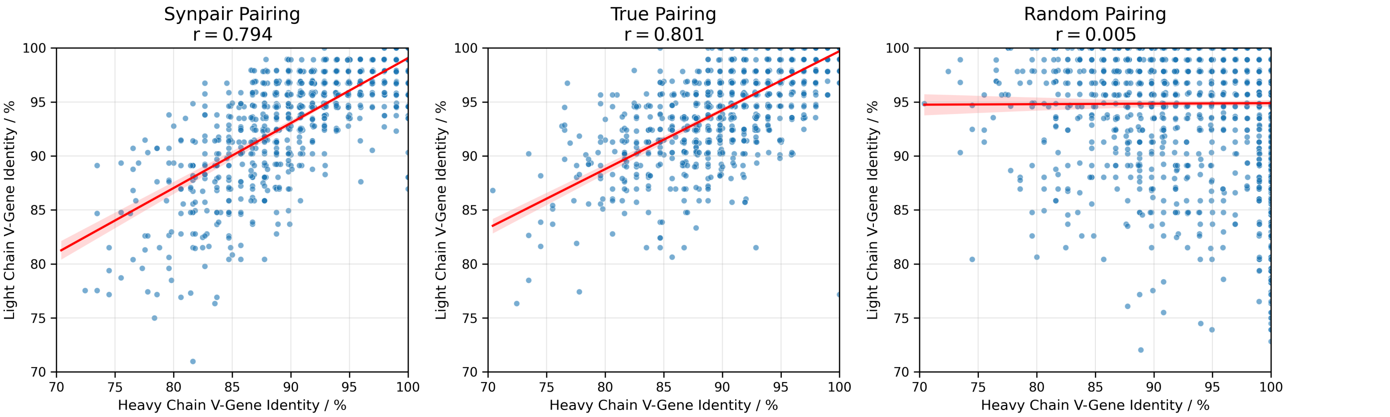

Figure 3. Pearson correlation between germline identity for heavy chain v-gen and light chain V-gene for SynPair Pairings, True Pairings, and Random Pairings.

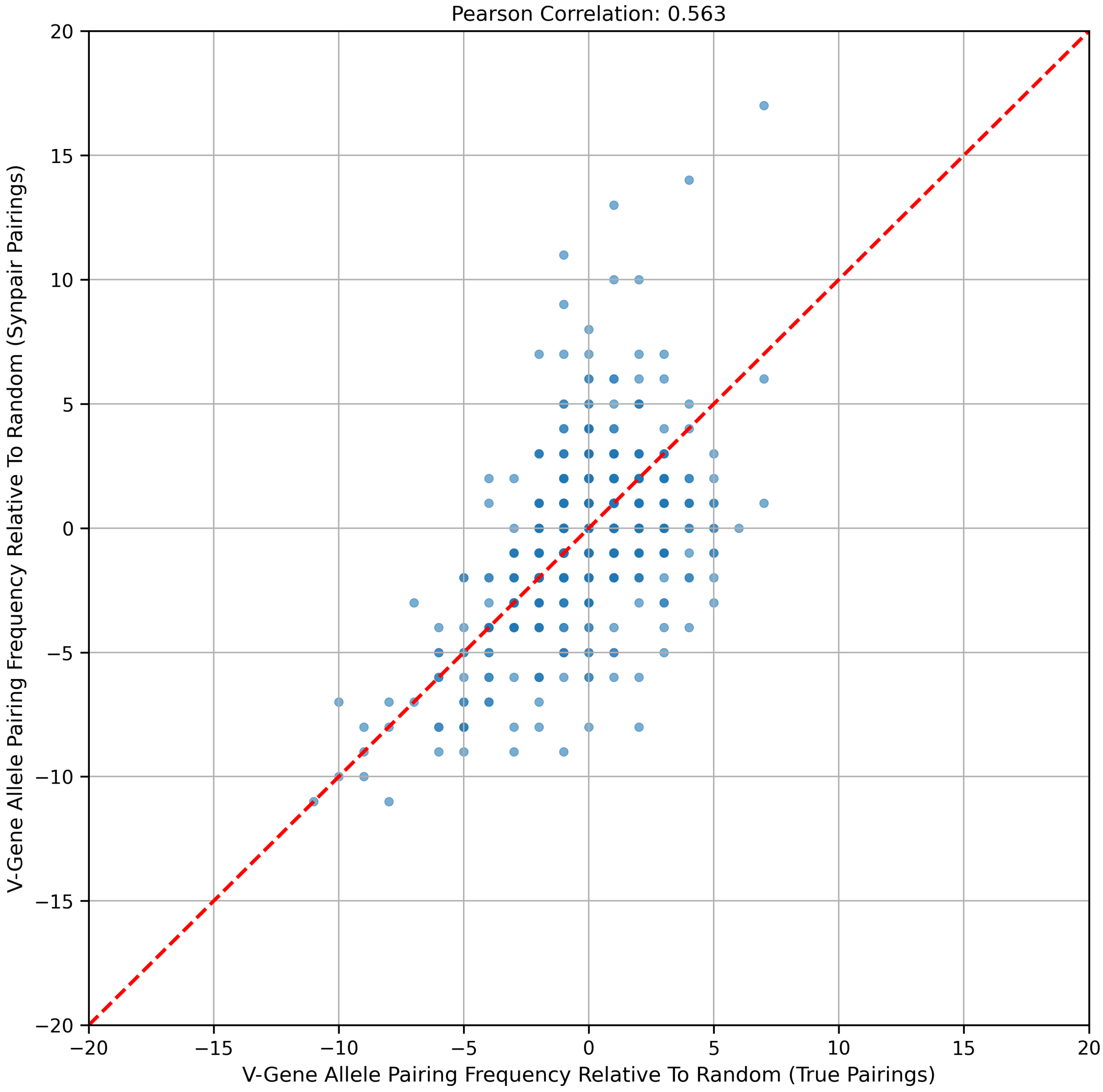

Figure 4. Correlation of the frequency of different VH/VL V-gene combinations in SynPair and True pairings, relative to the Random set pairings.