SaNano - Structure Aware Transfer Learning For Data Limited Protein Modality

Hugo Frelin [1] **Andrei Kamenski** [1] **Paolo Marcatili** [1]

Abstract

Single domain antibodies such as camelid VHH (or NANOBODY®) are increasingly adopted as therapeutics due to their compact architecture, stability, and favorable developability. However, their distinct structural features limit transferability from general protein language models, and the limited number of available sequences constrains training large VHH-specific foundation models from scratch. We introduce SaNano, a structure-aware VHH language model that unifies sequence and Foldseek 3Di structural tokens, fine-tuned from SaProt. Trained on curated VHH sequences paired with NanobodyBuilder2 structures and mixed with SwissProt/AlphaFold data, SaNano interleaves structure-conditioned and sequence-only masked language modeling to internalize structural priors while retaining strong sequence-only inference. SaNano outperforms general protein and antibody/VHH baseline models on few-shot contact prediction, biophysical property prediction, and sequence reconstruction (pseudo-perplexity), with especially large gains in the highly variable complementarity-determining region 3 (CDR3). Crucially, structure-aware fine-tuning improves sequence-only performance, reducing reliance on costly structure prediction in high-throughput screening. SaNano is available at https://huggingface.co/novonordisk-red/SaNano.

1. Introduction

Variable heavy domain of heavy chain (VHH) antibodies, commonly known as NANOBODY®, have seen rapidly increasing adoption in biopharmaceutical research and development. Their compact single-domain architecture offer superior tissue penetration, excellent thermal and chemical stability, favorable developability profiles, as well as high antigen-binding affinity comparable to conventional monoclonal antibodies, making them attractive modalities for therapeutics and diagnostics (Jin et al., 2023).

The advent of Transformer architectures has driven rapid progress in protein modeling, with masked language modeling enabling general-purpose protein language models such as ESM-1b and ESM-2 (Vaswani et al., 2017; Rives et al., 2021; Lin et al., 2023; Nambiar et al., 2020). Building on this foundation, modality-specific sequence models have been developed for antibodies and VHHs, including AbLang, IgBert, NanoBert, and VHHBert (Olsen et al., 2022; Kenlay et al., 2024; Hadsund et al., 2024; Tsuruta et al., 2024). In parallel, advances in structure prediction and the availability of large structural databases have enabled models that learn from backbone geometry (Jumper et al., 2021; Varadi et al., 2022; 2023; Høie et al., 2025; Hsu et al., 2022). The ever growing protein structure databases created a need for efficient database search, something Foldseek addressed through the 3Di structural alphabet. SaProt subsequently combined amino acid and 3Di tokens into a unified structure-aware vocabulary supporting both sequence-only and structure-conditioned inference (van Kempen et al., 2024; Su et al., 2024). However, existing VHH-focused models are sequence-only, while existing structure-aware models are not specialized to the distinctive geometry of VHHs.

VHHs possess multiple unique structural features compared to conventional antibodies, most of which can be viewed as adaptations to the loss of the light chain. These include the presence of hallmark residues in FR2 at the light chain interface positions, on average longer CDR3, and the ability of VHH CDR3 to adopt unconventional conformations, often resulting in distinctive, more convex, paratope shapes (Jin et al., 2023; Bahrami Dizicheh, 2023). Because of these unique features, we hypothesize that VHH prediction and generation will benefit from a multimodal protein language model trained explicitly on both VHH sequences and structures. While general protein language models capture broad structural principles and antibody language models capture

[1]Novo Nordisk. Correspondence to: Hugo Frelin <ahyf@novonordisk.com>, Andrei Kamenski <endk@novonordisk.com>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

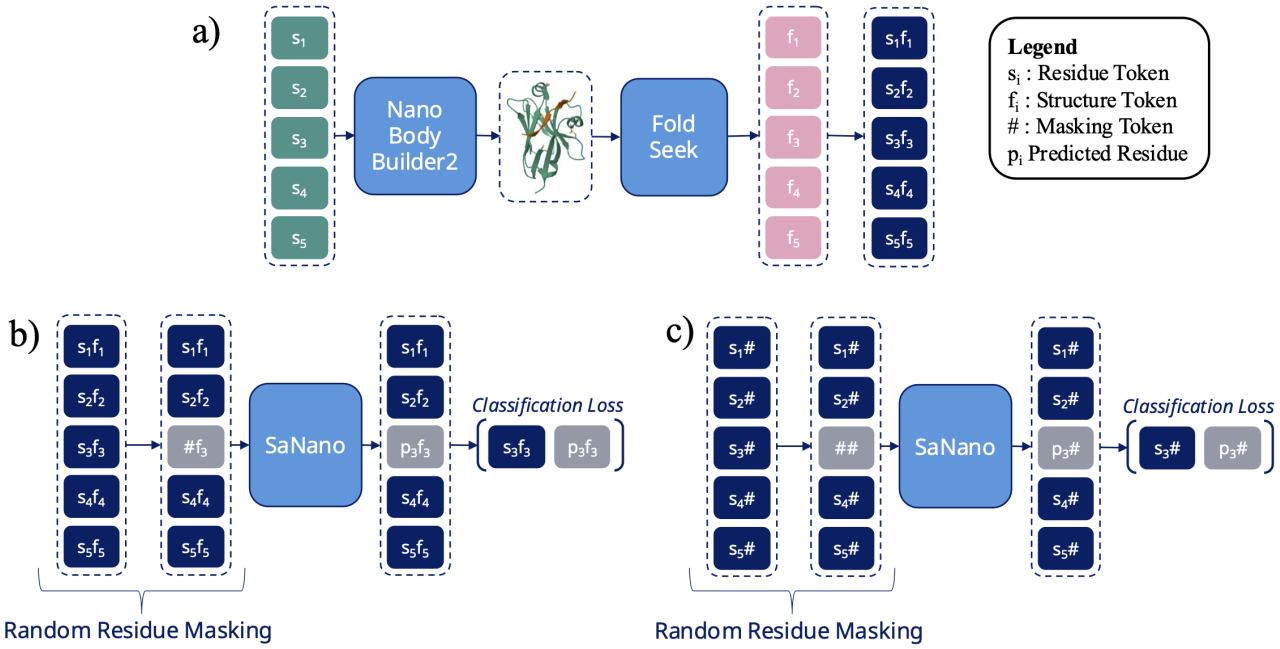

Figure 1. a) Shows the pipeline for generating SA tokens from VHH sequences using NanoBodyBuilder2 followed by Foldseek (van Kempen et al., 2024; Abanades et al., 2023). b) Displays structure-conditioned MLM using SA tokens. In this scenario, the structural character f_3 is visible and only the residue character s_3 is masked, yielding the masked token $\#f_3$. c) Displays sequence-only MLM, where all the structural characters are masked. In this case the masked token is simply $\#\#$. To ensure SaNano performs well in both usage scenarios (with and without structure) these two methods, b) and c), are mixed in training.

features specific to conventional antibodies, their implicit knowledge may not efficiently transfer to VHHs due to their distinct geometric properties. To our knowledge, no existing VHH-specific model integrates both sequence and structure within a unified framework. To address this gap, we present the following contributions:

- We introduce SaNano, the first VHH-specific language model that processes both sequence and structure tokens.
- We demonstrate that structure-aware fine-tuning significantly improves sequence-only inference, allowing the model to internalize structural priors without requiring expensive structure prediction at runtime.
- We show through comprehensive benchmarking that SaNano outperforms both general protein models and specialized VHH and antibody models on contact prediction, biophysical property prediction, and sequence reconstruction.

2. Methods

2.1. Base Model Architecture

SaNano builds upon SaProt, a 33-layer BERT-style encoder-only transformer with 650M parameters, identical to ESM2-650M (Devlin et al., 2018; Lin et al., 2023). SaProt extends standard amino acid tokenization with the SA vocabulary, allowing the model to process both sequence-only and structure-conditioned inputs using a single encoder (Su et al., 2024).

We fine-tune SaProt using Low-Rank Adaptation (LoRA), and introduce trainable low-rank matrices into the key, query, value and dense matrices of the transformer block while freezing the base model weights (Hu et al., 2021). This reduces computational cost and mitigates catastrophic forgetting, allowing SaNano to specialize for VHH sequences while retaining general protein understanding.

2.2. Training and Tokenisation

SaNano adopts the SA vocabulary (Su et al., 2024). We fine-tune using MLM with a 15% masking rate and two input modes displayed in Figure 1 b) and c). The first mode is structure-conditioned, in which both the residues and their paired 3Di tokens are provided. The second one is sequence-only, in which all the structural characters are masked. This design encourages the model to learn structure-informed representations while maintaining strong sequence-only performance at inference time, where large-scale structure prediction is computationally expensive. Full training and model details are in Appendix A.2.

2.3. Data

SaNano was trained on the 257,000 VHH sequences remaining of the Integrated Database of Nanobodies for Immunoinformatics (INDI1) after clustering and filtering, with structures predicted using NanobodyBuilder2 (Deszyński et al., 2021; Abanades et al., 2023). To mitigate overfitting and preserve general protein understanding, we supplemented the training data with sequences from SwissProt and their corresponding AlphaFold 2 structures (Varadi et al., 2022; 2023). This was done in a 3:1 ratio of VHH to general protein sequences. Data pre processing details can be found in Appendix A.1.

3. Experiments and Results

We evaluate SaNano on five biophysical regression tasks (three proprietary and two from NbBench), few-shot contact prediction on SAbDab, and pseudo-perplexity on 26,000 non-redundant INDI1 VHH sequences; dataset details are provided in Appendix A.3.

3.1. Contact Prediction

Table 1. The AUC results of contact prediction experiments. Green highlight indicate best result, and bold text indicate best sequence only result.

Model	Modality	Short	Medium	Long
ESM2-650M	Seq	0.563	0.732	0.781
SaProt	Seq	0.588	0.730	0.824
	Seq+Str	0.641	0.796	0.866
SaNano	Seq	**0.609**	**0.753**	**0.830**
	Seq+Str	0.633	0.792	0.859

Contact prediction is a few-shot learning task to evaluate how well a PLM has learned structural relationships through unsupervised pretraining (Rao et al., 2020). We define two residues to be in contact if their Cα atoms are within 8Å of each other. Following standard practice, we freeze the base model and fit a logistic regression classifier to the attention maps between residue pairs. With only 10 training and 20 validation samples, the linear probe identifies which attention heads encode physical proximity.

Table 1 presents the contact prediction results. Consistent with prior work, SaProt and SaNano achieves its best performance when it is conditioned on structure (Su et al., 2024). This is comes as no surprise, as it directly predicts structure from structure. However, SaNano demonstrates superior structural understanding of VHHs compared to both SaProt and ESM2-650M in the sequence-only setting, indicating that fine-tuning has enhanced its implicit structural representations.

Table 2. Pseudo-perplexity on holdout VHH test set across CDR and framework regions. Lower values indicate better sequence reconstruction abilities. Bold text indicate best sequence only result.

Model	Modality	CDR1	CDR2	CDR3	FW
ESM2-150M	Seq	12.12	11.28	19.55	4.86
ESM2-650M	Seq	11.57	10.32	18.08	3.36
AbLang2	Seq	10.44	10.08	20.40	3.16
SaProt	Seq	6.99	8.73	14.17	2.06
SaNano	Seq	**5.24**	**6.16**	**11.14**	**1.83**

3.2. Biophysical Property and Antigen Binding Prediction

For biophysical property and antigen binding prediction, we trained ridge regression models on frozen, mean-pooled embeddings. Regularization hyperparameters were selected using the validation set, and Table 3 reports Spearman Correlations on the test sets. SaNano achieves the best performance across all seven benchmarks. Notably, while SaProt shows improved thermostability prediction on our proprietary data when provided with structural information, SaNano performs equally well with and without structure, demonstrating robust sequence-only representations. The ability of SaNano to match or exceed structure-conditioned performance in sequence-only mode is particularly valuable for high-throughput screening scenarios where structure prediction for millions of candidates would be computationally prohibitive. Seq+Str regime was not used for the antigen binding predictions due to the high structural similarity of the VHH variants within the datasets.

Additionally, SaNano outperforms both VHH-specific models (NanoBert, VHHBert) and antibody-focused models (AbLang-Heavy, IgBert), despite these models being trained on substantially more immunoglobulin data (Hadsund et al., 2024; Tsuruta et al., 2024; Olsen et al., 2022; Kenlay et al., 2024). This suggests that the combination of structural awareness and domain-specific fine-tuning is more effective than sequence-only training on specialized datasets alone and sequence-only transfer learning.

3.3. Pseudo-Perplexity

To assess SaNano's potential as a variation generator for drug discovery, we evaluated pseudo-perplexity on a held-out test set of NGS VHH sequences from INDI1 (Deszyński et al., 2021). Lower perplexity indicates better modeling of the sequence distribution and higher confidence in generative predictions. As shown in Table 2, SaNano achieves significantly lower perplexity than both SaProt, AbLang2 and ESM2 across all regions in both sequence-only. Notably, it has a significant improvement in the highly variable CDR3. This demonstrates SaNano's capability to generate plausible VHH mutations, a key requirement for computa-

Table 3. VHH property performance comparison, reported in Spearman's rank correlation (ρ), between protein representation models. Highlighted in green is the best result for a given dataset, and bold text shows the best sequence-only performance.

		Internal Proprietary Data			Public			
					Thermo		Binding	
Model	**Modality**	**Hydrophobicity**	**Expression**	**Thermo**	**Thermo-seq**	**Thermo-tm**	**VHH72-CoV2**	**VHH72-CoV1**
VHH Models								
NanoBert	Seq	0.64	0.56	0.49	0.58	0.76	0.58	0.59
VHHBert	Seq	0.56	0.54	0.45	0.61	0.68	0.63	0.63
AntiBody Models								
AbLang-Heavy	Seq	0.59	0.62	0.61	0.71	0.74	0.49	0.50
IgBert-Unpaired	Seq	0.57	0.55	0.62	0.70	0.82	0.72	0.72
IgBert	Seq	0.62	0.57	0.55	0.69	0.85	0.72	0.71
Protein Models								
ESM2-150M	Seq	0.66	0.62	0.56	0.75	0.76	0.70	0.72
ESM2-650M	Seq	0.65	0.61	0.56	0.71	0.72	0.74	0.75
SaProt	Seq	0.66	0.64	0.60	0.76	0.81	0.74	0.74
	Seq+Str	0.66	0.64	0.65	0.74	0.73	-	-
Fine-Tuned Model								
SaNano	Seq	**0.67**	**0.68**	**0.68**	**0.78**	**0.86**	**0.76**	**0.76**
	Seq+Str	0.67	0.67	0.68	0.79	0.74	-	-

tional VHH design.

4. Discussion

A central finding of this work is that fine-tuning a structure-aware base model on domain-specific data improves performance even when structural information is withheld at inference time. In our contact prediction benchmarks, SaNano outperformed both the general-purpose ESM2-650M and the base SaProt model in the sequence-only setting. This suggests that by training on multimodal inputs (sequence and structure), the model internalizes the structural priors specific to the VHH modality; it learns an improved implicit mapping of sequence to structure, compared to the general protein models. Furthermore, SaNano's representation are in best in class for a series of important VHH developability prediction targets, indicating that its latent space can effectively represent VHHs.

The ability to suggest novel, biologically plausible mutations is critical for de novo antibody design. Our evaluation on the held-out test set provides strong evidence of SaNano's generative capabilities. While the general ESM2 model and the antibody specific AbLang2 model struggle to model the hypervariable CDR3 distribution (PPLs of $\approx$ 18.3 and $\approx$ 20.4 respectively), SaNano achieved a perplexity of 11.0. This significant reduction indicates that the model has successfully learned the "grammar" of VHH CDR3s and can effectively narrow the search space for design candidates. Crucially, this performance was achieved on a non-redundant test set, indicating that the model is generalizing to unseen clonotypes rather than merely memorizing training examples.

4.1. Limitations

Our work relies on predicted structures from NanobodyBuilder2 for the training set, as experimental crystal structures for VHHs are scarce. While we mitigate error propagation by mixing in high-quality AlphaFold structures from SwissProt, the model's understanding of fine-grained structural details is ultimately bounded by the accuracy of the structure predictor. Additionally, our training data is heavily biased towards Llama and Alpaca sequences from the INDI1 database.

4.2. Conclusion

In this work, we introduced SaNano, a VHH-specific protein language model developed through structure-aware transfer learning. By fine-tuning a multimodal base model on a curated dataset of VHH sequences and structures, we achieved state-of-the-art performance on both representation learning and generative tasks. Our results demonstrate that incorporating structural information during pre-training offer significant benefits, even for downstream tasks that rely solely on sequence inputs. We believe this method of leveraging structure-aware base models with domain-specific fine tuning provides a generalizable blueprint for developing specialized AI tools across the protein universe.

Acknowledgements

We thank the reviewers for their constructive feedback. We also thank Teresa Sousa Sotto Mayor and Jonas Blomquist Jørgensen for their assistance with data preprocessing and for insightful discussions. We are grateful to colleagues at Novo Nordisk for helpful feedback and support throughout this project.

Impact Statement

This work aims to advance machine learning for VHH protein modeling and therapeutic discovery. Potential benefits include more efficient biologic design, while potential risks include misuse for harmful design tasks, dataset bias, and overreliance on computational predictions without sufficient experimental validation. We therefore recommend that such models be used only with appropriate safeguards and downstream experimental verification.

Trademark Notice

NANOBODY® is a registered trademark of Sanofi or an affiliate.

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): 575, May 2023. ISSN 2399-3642. doi: 10.1038/s42003-023-04927-7. URL https://doi.org/10.1038/s42003-023-04927-7.

Agarwal, A., Harrang, J., Noble, D., McGowan, K., Lange, A., Engelhart, E., Lahman, M., Adamo, J., Yu, X., Serang, O., Minch, K., Wellman, K., Younger, D., Lopez, R., and Emerson, R. AlphaBind, a domain-specific model to predict and optimize antibody–antigen binding affinity. *mAbs*, 17:2534626, July 2025. URL https://doi.org/10.1080/19420862.2025.2534626.

Bahrami Dizicheh, Z., C. I. . K. P. VHH CDR-H3 conformation is determined by VH germline usage. *Communications Biology*, 6:863, 2023. URL https://doi.org/10.1038/s42003-023-05241-y.

Deszyński, P., Młokosiewicz, J., Volanakis, A., Jaszczyszyn, I., Castellana, N., Bonissone, S., Ganesan, R., and Krawczyk, K. Indi—integrated nanobody database for immunoinformatics. *Nucleic Acids Research*, 50 (D1):D1273–D1281, 11 2021. ISSN 0305-1048. doi: 10.1093/nar/gkab1021. URL https://doi.org/10.1093/nar/gkab1021.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Hadsund, J. T., Satława, T., Janusz, B., Shan, L., Zhou, L., Röttger, R., and Krawczyk, K. nanobert: a deep learning model for gene agnostic navigation of the nanobody mutational space. *Bioinformatics Advances*, 4(1):vbae033, 03 2024. ISSN 2635-0041. doi: 10.1093/bioadv/vbae033. URL https://doi.org/10.1093/bioadv/vbae033.

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. Learning inverse folding from millions of predicted structures. *bioRxiv*, 2022. doi: 10.1101/2022.04.10.487779. URL https://www.biorxiv.org/content/early/2022/04/10/2022.04.10.487779.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., and Chen, W. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021. URL https://arxiv.org/abs/2106.09685.

Høie, M. H., Hummer, A. M., Olsen, T. H., Aguilar-Sanjuan, B., Nielsen, M., and Deane, C. M. Antifold: improved structure-based antibody design using inverse folding. *Bioinformatics Advances*, 5(1):vbae202, 03 2025. ISSN 2635-0041. doi: 10.1093/bioadv/vbae202. URL https://doi.org/10.1093/bioadv/vbae202.

Jin, B.-K., Odongo, S., Radwanska, M., and Magez, S. Nanobodies: A review of generation, diagnostics and therapeutics. *International Journal of Molecular Sciences*, 24:5994, 03 2023. doi: 10.3390/ijms24065994.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. Highly accurate protein structure prediction with alphafold. *Nature*, 596 (7873):583–589, Aug 2021. ISSN 1476-4687. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Kenlay, H., Dreyer, F. A., Kovaltsuk, A., Miketa, D., Pires, D., and Deane, C. M. Large scale paired antibody language models. 2024.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos

Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., and Ritz, A. Transforming the language of life: Transformer neural networks for protein prediction tasks. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '20, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379649. doi: 10.1145/3388440.3412467. URL https://doi.org/10.1145/3388440.3412467.

Olsen, T. H., Moal, I. H., and Deane, C. M. Ablang: An antibody language model for completing antibody sequences. *bioRxiv*, 2022. doi: 10.1101/2022.01.20.477061. URL https://www.biorxiv.org/content/early/2022/01/22/2022.01.20.477061.

Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., and Rives, A. Transformer protein language models are unsupervised structure learners. *bioRxiv*, 2020. doi: 10.1101/2020.12.15.422761. URL https://www.biorxiv.org/content/early/2020/12/15/2020.12.15.422761.

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., and Fergus, R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021. doi: 10.1073/pnas.2016239118. URL https://www.pnas.org/doi/abs/10.1073/pnas.2016239118.

Schneider, C., Raybould, M. I. J., and Deane, C. M. SAbDab in the age of biotherapeutics: updates including SAbDab-nano, the nanobody structure tracker. *Nucleic Acids Res.*, 50(D1):D1368–D1372, January 2022.

Steinegger, M. and Söding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, 35(11):1026–1028, November 2017.

Su, J., Han, C., Zhou, Y., Shan, J., Zhou, X., and Yuan, F. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6MRm3G4NiU.

Tsuruta, H., Yamazaki, H., Maeda, R., Tamura, R., and Imura, A. A SARS-CoV-2 interaction dataset and VHH sequence corpus for antibody language models. 2024.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L. M., Söding, J., and Steinegger, M. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, 42(2):243–246, Feb 2024. ISSN 1546-1696. doi: 10.1038/s41587-023-01773-0. URL https://doi.org/10.1038/s41587-023-01773-0.

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., and Velankar, S. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, Jan 2022. ISSN 0305-1048. doi: 10.1093/nar/gkab1061. URL https://doi.org/10.1093/nar/gkab1061.

Varadi, M., Bertoni, D., Magana, P., Paramval, U., Pidruchna, I., Radhakrishnan, M., Tsenkov, M., Nair, S., Mirdita, M., Yeo, J., Kovalevskiy, O., Tunyasuvunakool, K., Laydon, A., Žídek, A., Tomlinson, H., Hariharan, D., Abrahamson, J., Green, T., Jumper, J., Birney, E., Steinegger, M., Hassabis, D., and Velankar, S. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 11 2023. ISSN 0305-1048. doi: 10.1093/nar/gkad1011. URL https://doi.org/10.1093/nar/gkad1011.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. Huggingface's transformers:

State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Zhang, Y. and Tsuda, K. NbBench: Benchmarking language models for comprehensive nanobody tasks. 2025.

A. Appendix

A.1. Training Data

A.1.1. Data Sources

VHH sequences were obtained from the Integrated Database of Nanobodies for Immunoinformatics (INDI1) (Deszyński et al., 2021), while protein sequences and their predicted structures were retrieved from SwissProt via the AlphaFold Protein Structure Database (Varadi et al., 2022; 2023). Structures for VHH sequences were predicted using NanobodyBuilder2 (Abanades et al., 2023).

A.1.2. VHH Sequence Filtering and Curation

The INDI1 NGS dataset was filtered through a multi-step pipeline to ensure sequence quality and diversity. First, we retained only sequences with CDR3 regions observed at least twice in the dataset to reduce sequencing artifacts and low-confidence entries. Redundant sequences sharing identical CDR3s were collapsed by retaining the longest full-length representative.

To control sequence similarity while preserving diversity, we performed clustering using MMSeqs2 (Steinegger & Söding, 2017) with a minimum sequence identity threshold of 95% and coverage mode 1, retaining the longest sequence from each cluster as the representative. Finally, we removed sequences with truncations in framework regions FR1 or FR4 to ensure complete structural context for modeling. This filtering procedure yielded approximately 240,000 high-quality VHH sequences for training. To this set of sequences, an additional approximately 17,000 patented VHH sequences were added, resulting in a VHH training datasets of about 257,000 sequences.

A.1.3. Structure-Aware Token Generation

VHH structures predicted by NanobodyBuilder2 and AlphaFold structures from SwissProt were processed using Foldseek to generate 3Di structural tokens (van Kempen et al., 2024). These 3Di tokens were combined with amino acid identities to form composite structure-aware (SA) tokens following the SaProt vocabulary (Su et al., 2024). During training, each sequence was randomly assigned either sequence-only tokens (standard amino acids + structure masking token) or SA tokens (amino acid + 3Di token) with equal probability, enabling the model to learn robust representations in both sequence-only and structure-conditioned settings.

A.2. Training Details

The model hyperparameters are displayed in Table 4. Training was implemented using the HuggingFace Transformers library and the PEFT (Parameter-Efficient Fine-Tuning) library for LoRA adaptation (Wolf et al., 2020; Mangrulkar et al., 2022; Hu et al., 2021).

A.2.1. Model Architecture and Adaptation

SaNano is built on the SaProt base model, a 33-layer BERT-style encoder with 650 million parameters pretrained on 40 million sequences and structures (Su et al., 2024). Apart from the tokenisation, the model architecture is identical as ESM2-650m (Lin et al., 2023). We applied LoRA to the query, key, value, and dense projection matrices in all transformer layers, introducing approximately 24 million trainable parameters (3.7% of the base model) while keeping the remaining weights frozen. This parameter-efficient approach aims to preserve the general protein understanding of the pretrained model while enabling domain-specific adaptation to VHH sequences.

A.2.2. Data Sampling and Modality Mixing

In training, the VHH sequences were randomly mixed up with randomly sampled SwissProt sequences and structures at a 3:1 VHH to protein ratio. To avoid out-of-memory issues, only proteins of 200 residues or less were used. The SwissProt sequences were, just like the VHHs, also randomly set to either have SA tokens, or the only have the residues and all structure tokens masked.

Table 4. Training hyperparameters and configuration for SaNano self-supervised fine-tuning.

Category	Parameter	Value
Base Model		
	Architecture	SaProt (33-layer BERT)
	Parameters	650M (frozen base + LoRA adapters)
Data		
	Training set	VHH sequences (≈257k) + SwissProt
	VHH:SwissProt ratio	3:1 by sampling weight
	Structure tokens	50% Seq+Str, 50% Seq only
	Train/Val/Test split	90% / 5% / 5%
LoRA Configuration		
	Rank (r)	32
	Alpha (α)	64
	Dropout	0.1
	Target modules	Query, Key, Value, Dense
	Bias	None
MLM		
	Masking rate	15%
	Masking strategy	Uniform
Optimization		
	Optimizer	AdamW
	Learning rate	2×10^{-5}
	Scheduler	Cosine decay
	Warmup steps	1,000
	Weight decay	0.01
	Steps	5000
Training Setup		
	Effective batch size	64
	Mixed precision	FP16

A.2.3. MASKING STRATEGY

We employed a simplified masking strategy where all selected tokens (15%) were replaced with the masking token, departing from the standard BERT protocol of 80% masking, 10% random replacement, and 10% unchanged tokens (Devlin et al., 2018). Preliminary experiments in our setup showed that including random replacements and identity retention did not yield performance improvements for this fine-tuning task, likely because the SaProt base model was already pretrained with the full BERT masking scheme and has learned robust representations. Thus, we opted for the simplified approach to focus the training objective purely on context-based reconstruction.

An SA token $s_i f_i$ consist of sequence character s_i and its Foldseek structure character f_i. The masking token used is $\#$. Thus, when in sequence-only modality (when the all structure characters are masked the token is and the tokens are $s_i\#$), masking token i yields the $\#\#$. If however structure is provided, only the residue token is masked, yielding $\#f_i$. This is because FoldSeek tokens of predicted structures are too noisy to be predicted, and is the masking method used by Su et al. to the train the base SaProt model.

A.2.4. OPTIMIZATION AND TRAINING REGIME

The model was trained for one epoch using the AdamW optimizer with a learning rate of 2×10^{-5}, a cosine decay schedule, and 1,000 warmup steps. We used an effective batch size of 64 (16 per device with 4 gradient accumulation steps) and applied a weight decay of 0.01 to prevent overfitting. Training employed mixed-precision (FP16) computation to reduce memory footprint and accelerate convergence. The model was trained for 5000 steps, which is just around 1 epoch, to avoid over-fitting.

A.3. Evaluation Datasets

Table 5. Evaluation datasets for SaNano representation benchmarking. There are seven VHH property regression tasks, one contact prediction dataset, and a test set of CDR3 non-redundant VHH sequences for sequence reconstruction.

Name	Task	Granularity	Source	Data Split (train/val/test)
Thermostability	Regression	Sequence	Proprietary	5-fold CV
Expression	Regression	Sequence	Proprietary	5-fold CV
Hydrophobicity	Regression	Sequence	Proprietary	5-fold CV
Thermo-seq	Regression	Sequence	NbBench (Zhang & Tsuda, 2025)	522/92/147
Thermo-tm	Regression	Sequence	NbBench (Zhang & Tsuda, 2025)	396/57/114
VHH72-CoV2	Regression	Sequence	open-alphaseq, YM0549 (Agarwal et al., 2025)	5-fold CV
VHH72-CoV1	Regression	Sequence	open-alphaseq, YM0549 (Agarwal et al., 2025)	5-fold CV
Contact Prediction	Classification	Residue	SAbDab (Schneider et al., 2022)	10/20/669
VHH NGS Alpaca	Reconstruction	Residue	INDI1 (Deszyński et al., 2021)	0/0/26,000

A.4. Evaluation Tasks

A.4.1. CONTACT PREDICTION

The contact prediction experiments fits a logistic regression model function $f : \mathbb{R}^{nd} \to \mathbb{R}$, where n is the number of attention heads and d the model depth. Thus nd is the number of times two given residues attend to each other. This simple linear probe acts as a selector, identifying which specific attention heads have specialized to encode 3D proximity during pretraining. The model is trained in a few-shot setting (10 training structures) to ensure it relies on existing structural knowledge rather than learning new features.

The structures used are a non-redundant set of VHH structures, both bound and unbound, with a maximum sequence similarity of of 90% from the The Structural Antibody Database (Schneider et al., 2022).

Table 6 displays the Precision at Length (P@L) metrics for three different ranges. For a protein of length L, P@L is the precision of the top L predicted contacts for a protein. Contacts are categorized by sequence separation as short ($6 \leq$ separation < 12), medium ($12 \leq$ separation < 24) and long (separation ≥ 24).

Table 6. Contact prediction performance measured using P@L metrics for short, medium and long range contacts.

		Short Range 6 ≤ separation <12			**Medium Range** 12 ≤ separation <24			**Long Range** 24 ≤ separation		
Model	**Modality**	**P@L**	**P@L/2**	**P@L/5**	**P@L**	**P@L/2**	**P@L/5**	**P@L**	**P@L/2**	**P@L/5**
ESM2-650M	Seq	0.327	0.542	0.829	0.553	0.747	0.862	0.601	0.792	**0.948**
SaProt	Seq	0.346	0.586	0.844	0.531	0.735	0.918	0.690	0.830	0.940
	Seq+Str	0.364	0.660	0.917	0.596	0.815	0.949	0.762	0.887	0.928
SaNano	Seq	**0.355**	**0.606**	**0.879**	**0.562**	**0.770**	**0.944**	**0.704**	**0.847**	0.903
	Seq+Str	0.358	0.643	0.912	0.592	0.812	0.958	0.763	0.880	0.902

A.4.2. Biophysical Property Prediction

Accurate prediction of biophysical properties is critical for assessing the developability and therapeutic potential of VHH drug candidates. To evaluate the quality of the learned representations, we trained ridge regression models on frozen, mean-pooled embeddings extracted from the last hidden layer of each model. This simple downstream evaluation protocol ensures that performance differences reflect the quality of the pre-trained representations rather than the capacity of the prediction head.

We utilized three proprietary datasets measuring key early-stage developability parameters: thermostability, hydrophobicity, and expression levels. For these internal datasets, we report the average Spearman correlation from a five-fold nested cross-validation. To prevent data leakage and ensure rigorous evaluation, the cross-validation partitions were generated using MMSeqs2 clustering with a sequence identity threshold of 85%, ensuring that structurally similar sequences do not appear in both training and validation sets (Steinegger & Söding, 2017).

In addition to proprietary data, we evaluated performance on two public thermostability benchmarks: Thermo-Seq and Thermo-Tm, which are part of the NbBench suite (Zhang & Tsuda, 2025).

A.4.3. Antigen Binding Prediction

Antigen binding affinity is a key parameter for the selection of therapeutic or tool VHHs in discovery campaigns. The performance of PLM embeddings for affinity prediction was evaluated using the same methodology as employed for biophysical property prediction.

A publicly available dataset comprising binding data of variants of a broadly neutralizing SARS-CoV VHH72 was used for evaluation (Agarwal et al., 2025). Evaluation was carried out on two antigens, SARS-CoV-1 and SARS-CoV-2. Random partitions were used due to the dataset being composed of closely related variants of VHH72, where meaningful sequence identity-based splits were difficult to achieve without significant loss of data.

A.4.4. Pseudo Perplexity

Pseudo perplexity, PPL, is used to quantify a protein language models level of surprise of a given sequence. We used this metric to evaluate a models ability to generate plausible mutations. The lower the PPL, the better the model is at understanding and predicting the region. The pseudo perplexity of a region R in a sequence is calculated as:

$$PPL = \exp\left(-\frac{1}{|R|}\sum_{i\in R}\log p(s_i|s_{j\neq i})\right) \tag{1}$$

Since we use the SA vocabulary, there are 21 tokens for each residue, 20 for each structural token f_i and the masking token $\#$. The probability of a given residue s_i is thus the sum of the probabilities of all 21 tokens it is represented by. Let the full set of structural tokens be $\mathcal{F}$. Thus, the PPL calculation for SaNano and SaProt is:

$$PPL = \exp\left(-\frac{1}{|R|}\sum_{i\in R}\log\sum_{\mathcal{F}\in f} p(s_i, f|s_{j\neq i})\right) \tag{2}$$

To better understand the models' abilities to generate valuable mutations, the reported PPL is calculated on test set of about 26,000 sequences with non-redundant CDR3s. They are Alpaca NGS VHHs are retrieved from INDI1, and originate from NCBI sequencing projects PRJDB2382, PRJDB7792, PRJNA516512 (Deszyński et al., 2021).