TOPH (True Retrieval Of Proteins Homologs): Adapting A Constrastive Question-Answering Framework for Protein Search

Ron Boger [*1] **Amy Lu** [*1] **Seyone Chithrananda** [*1] **Kevin Yang** [1] **Petr Skopintsev** [1] **Ben Adler** [1] **Eric Wallace** [1] **Peter Yoon** [1] **Pieter Abbeel** [1] **Jennifer Doudna** [1]

Abstract

Protein homology detection is a pivotal aspect of bioinformatics, enabling insights into protein functions and evolution; however, detection of remote homologs - proteins with low sequence similarity - has proven challenging in classical bioinformatic methods. In this work, we present a novel approach that employs open domain question answering to retrieve remote homologs, aided by a carefully curated dataset with biologically relevant hard negatives for Dense Passage Retrieval (DPR) training. To evaluate our approach, we introduce a diverse CRISPR-Cas and evolutionary related nucleases protein dataset, providing a robust testbed for algorithmic improvement. Additionally, we offer a user-friendly web interface to streamline protein homology search.

1. Introduction

Remote homology detection is a crucial task in computational biology that seeks to identify proteins that are evolutionarily related but share little sequence similarity. Homology refers to proteins that share a common evolutionary origin. Remote homologs often perform similar functions or have similar structures, despite their sequences differing significantly due to evolutionary divergence over time. Identifying these relationships can provide insights into protein function, structure, and evolution. However, the task is challenging due to the vast sequence space and the subtlety of the signals that indicate homology.

In NLP, searching and retrieving over large collections of documents is traditionally performed using methods such as TF-IDF, which work using word matching and alignment between a user's query and each document in the collection. In recent years, search has been dramatically improved by shifting to deep-learning-based approaches, which largely combine two ingredients: (1) pre-trained language models, which can extract powerful semantic features from text, and (2) carefully-curated training sets that contain pairs of related documents as well as false positives (non-related documents) returned by traditional word-overlap methods. The advantage of deep learning approaches for search is that they provide high accuracy and fast speeds, due to embedding documents into low-dimensional semantic vectors that can be quickly searched over using approximate nearest neighbor methods.

Biology, on the other hand, continues largely to use traditional tools. BLAST and Hidden Markov Models have a long history of use searching over large databases of protein sequences scoring by residue overlap and alignment-based features. Structure based methods such as DALI (Holm, 2020) and TM-align (Zhang & Skolnick, 2005) have long conferred higher sensitivity to find remote homologs, but struggled to capture more widespread adoption due to their speed and number of available protein structures. With the advent of accurate protein structure prediction methods such as AlphaFold2 (Jumper et al., 2021), using prior tools to search through homologous structures has become all but untenable. Deep learning based methods such as Foldseek (van Kempen et al., 2023), TM-vec (Hamamsy et al., 2022), SMAMPNN (Trinquier et al., 2022), Progres (Greener & Jamali, 2022) have sought to meet this gap, but do not yet rival the sensitivity of DALI or speed of sequence searches (Steinegger & Söding, 2017).

Our work has 3 main contributions. First, we apply open domain question answering for highly sensitive and rapid retrieval of remote homologs, for which we are to the first to do to our knowledge. To this end, we carefully curate a dataset of hard negative proteins for DPR style training. Second, we introduce a dataset of single-effector (class 2) CRISPR-Cas proteins and evolutionary related nucleases with high sequence and structural diversity for evaluation of remote homology algorithms. Finally, we release a web interface for scientist to quickly and easily search for homologous proteins.

[*]Equal contribution [1]UC Berkeley. Correspondence to: Ron Boger <ronb@berkeley.edu>, Jennifer Doudna <doudna@berkeley.edu>.

The 2023 ICML Workshop on Computational Biology. Baltimore, Maryland, USA, 2023.

2. Methods

2.1. DPR (Dense Passage Retrieval)

Dense passage retrieval (DPR) (Karpukhin et al., 2020) is an innovative approach employed in open-domain question answering (QA) systems, which uses fine-tuned language models to improve information retrieval. We focus on the retriever component Given a question q and passage p, DPR fine tunes a BERT model to maximize dot product similarity $sim(q, p) = E_Q^T(q)E_P(p)$, where E_Q and E_P are the question and passage encoders respectively. A contrastive objective is used to train:

$$L(q_i, p_i^+, p_{i,1}^-, \ldots, p_{i,n}^-) = -\log \frac{e^{\text{sim}(q_i, p_i^+)}}{e^{\text{sim}(q_i, p_i^+)} + \sum_j e^{\text{sim}(q_i, p_{i,j}^-)}}$$

At inference time, E_p is applied to all the passages and indexed with FAISS (Johnson et al., 2019). Given a question q, its embedding is derived and top k passages with the closest embeddings are retrieved.

Hard negative sampling and in-batch negative sampling are essential techniques employed in the training of Dense Passage Retrieval (DPR) models. Hard negative sampling involves selecting samples that are difficult for the model to classify correctly, thereby pushing it to learn more nuanced features. For example, hard negatives might come from non-related documents that traditional word-overlap methods, like BM25, incorrectly return as relevant, or from errors in a trained DPR model itself. In-batch negative sampling, on the other hand, leverages the other negative examples within the same training batch as the current query-document pair. Both strategies are designed to strengthen the model's ability to discern between truly relevant documents (the Gold standard) and false positives, enhancing the precision of the retrieval process.

We adapt DPR to the protein search in several ways. We fine-tune on protein sequence models, such as ESM-2 (Lin et al., 2022), instead of BERT. We treat full proteins as both the question and passage, and proteins with known homology are considered positive examples. Instead of sourcing hard negatives from BM25, we choose proteins misclassified by existing protein search models.

2.2. Training and Evaluation

For training, we use sequences in the Astral Structural Classification of Proteins (SCOPe) database, using 40% sequence identity threshold on version 2.08 (Fox et al., 2013). We train on 40% sequence identity pairs in order to enrich for hard positives: homologous proteins with low sequence similarity. In contrast to related works like Progres (Greener & Jamali, 2022) that use higher sequence identity thresholds, we hypothesize that the ESM embeddings we feed into DPR implicitly capture sequence-structure relationships for domains where high sequence similarity indicates high structural similarity. This leaves 15,177 domains in the training set across over 4693 families. For a given protein pair within the same family, we select its hard negative by retrieving the protein within a different fold with the highest TM-vec score to the pair.

For evaluation, we use the same test set of 400 domains as Progres, which are ensured to have $< 30\%$ sequence identity to every protein in the training set. The domains in the set are filtered to have at least one other family, superfamily and fold member. For each domain, we measure the fraction of true positives (TPs) detected until the first incorrect family/fold/superfamily detected. ESM-2 embedding-based baselines use both the `esm2_t33_650M_UR50D` and `esm2_t36_3B_UR50D` models, which have 1280 and 2560-dimensional embeddings, respectively, ESM-2 in GPU mode was run using a NVIDIA RTX A6000 GPU.

We trained two models, fine-tuning either the `esm2_t6_8M_UR50D` or `esm2_t33_650M_UR50D` as the question and passage encoders with the DPR contrastive objective. We train the DPR model initialized with the `ESM_t6` weights (with 8M parameters) for 35 epochs, fine-tuning the final 3 layers. We train TOPH on initializing with `ESM_t33` weights (with 650M parameters) for 2 epochs, fine tuning the final 8 layers. Both models are trained on a single NVIDIA A100 GPU and with $lr = 1e - 4$.

2.3. CRISPR-Cas datasets for Remote Homology Detection

CRISPR-Cas, well-known for its revolutionary role in genome editing, is a key defense mechanism in adaptive bacterial immunity against foreign genetic elements (Jinek et al., 2012). We utilize a diverse CRISPR-Cas and evolutionary related nucleases protein dataset for remote homology detection. Our dataset draws from (Makarova et al., 2020) (Pausch et al., 2020) (Urbaitis et al., 2022) (Al-Shayeb et al., 2022), in addition to hand-curation from structural biologists. This choice is underpinned by the fact that the remote homologs are verifiable due to the unique biological characteristic of Cas genes being upstream of CRISPR loci. Furthermore, the ancestral and other evolutionary related RNA-guided nucleases originating from mobile genetic elements, such as TnpB (Sasnauskas et al., 2023), IscB (Schuler et al., 2022), and IsrB (Hirano et al., 2022), demonstrate similarity in structural organization and RNA-guided cleavage mechanism. This ensures a high level of confidence in our ground truth for homology detection. The related proteins in the dataset are highly varied in length, containing both long (which existing protein models struggle on; up to 1600 a.a.)

and short sequences (as low as 400 a.a.). The sequence and structural diversity of Cas and evolutionary related nucleases further renders them a valuable and challenging testbed for the evaluation of remote homology detection algorithms, given the complexity and subtlety of the signals indicating homology. Furthermore, anecdotes from structural biologists suggest Foldseek fails to find Cas proteins which can be found by DALI.

3. Results

3.1. SCOPe 2.08 sensitivity

We first assessed our model on sensitivity using the SCOPe 2.08 holdout set. Sensitivity is measured as the fraction of true positives (TPs) until the first incorrect fold, following the analysis in (van Kempen et al., 2023) and (Greener & Jamali, 2022). We consider TPs as same family, same superfamily but different family, and same fold but different superfamily for the family, superfamily, and fold tasks respectively. We chose to assess results in comparison to methods using both classical bioinformatic and deep learning approach on either structure or sequence.

Our results are comparable to Foldseek, without the need to search or process through folded structures. TM-Vec, another deep learning model learning on sequence pairs, shows a marginal improvement over TOPH in sensitivity on the SCOPe2.08 holdout set, but trains on the much larger CATH dataset. We have not done extensive hyperparameter tuning nor training on multiple GPUs. Furthermore, we suspect that a model trained on the fold/superfamily/family task should outperform a model that well-estimates TM-alignment scores. Classical structural alignment algorithms TM-Align and DALI perform similarly, with DALI returning the highest sensitivity of all models, but suffer from long run-times. We also benchmark against a class of ESM models that are not fine tuned on the family detection task. TOPH outperforms all ESM models, showing the benefit in DPR style training and fine tuning. Interestingly, we observe that when using the cosine similarities among ESM-embedded SCOPe proteins as a way to rank protein sequences, a surprising initial decrease in family/subfamily/fold prediction performance occurs going from the ESM 8M to 650M parameter models, before an increase when benchmarking the predictive power of the ESM_3B embeddings.

	Family	Superfamily	Fold
ESM2 (8M)	0.412	0.265	0.010
ESM2 (650M)	0.314	0.134	0.010
ESM2 (3B)	0.477	0.221	0.014
MMseqs2	*0.433*	*0.165*	*0.001*
TM-Vec	0.848	0.596	0.121
TM-Align (avg)	0.868	0.619	0.163
DALI	***0.885***	***0.709***	***0.168***
Foldseek	*0.821*	*0.578*	*0.070*
Progres	*0.878*	*0.680*	*0.144*
TOPH (ESM-650M)	0.818	0.528	0.065
TOPH (ESM-8M)	0.571	0.392	0.0376

Table 1. Retrieval sensitivity for homologous proteins in SCOPe. Sensitivity measures fraction of true positives (TPs) up to the first incorrect fold *(higher is better)*. We consider TPs as same family, same superfamily but different family, and same fold but different superfamily for the family, superfamily, and fold tasks respectively. We split the table by sequence, structure, and our models. Italicized results are reported from (Greener & Jamali, 2022). Results in bold are the best performing models on the family, superfamily, and fold result respectively. TM-align (avg) refers to the average of TM scores 1 and 2 outputted from the structural alignment.

3.2. CRISPR-Cas identification

Cas12 identification. Cas12 is a compact and efficient protein that creates staggered cuts in dsDNA, conferring great potential for genome editing. We curated a list of 436 Cas12 proteins from recent efforts in Cas12 effector identification (Makarova et al., 2020) (Pausch et al., 2020) (Urbaitis et al., 2022) (Al-Shayeb et al., 2022). As a frame of comparison, we included 8 unique TnpB proteins, which are Cas12 ancestors only recently shown to be RNA guided endonucleases (Shmakov et al., 2017) (Karvelis et al., 2021). When passed forward through our model, we observed clear separation between Cas12 subtypes 1. Many Cas12 subtypes clustered together suggesting similar embedded structures. Of particular interest, biphyletic groups Cas12a and Cas12b (Makarova et al., 2020) were indeed separated into two groups. Similarly, the polypheletic Cas12f (Harrington et al., 2018), was the least distinct cluster and co-located with TnpB proteins. Of particular note, Cas12k which guides targeting Tn7 transposition (Strecker et al., 2019) formed the most separated clade. Curiously, remaining uncharacterized Cas12 proteins Cas12U2, Cas12U3, and Cas12U4 were also highly separable from all other Cas12 proteins within this embedding. Potentially, this indicates a distinct biological role for these proteins.

Skopintsev dataset. In addition, we composed a dataset comprising of protein sequences which correctness was validated structurally. Specifically, the structures of Cas9, Cas12, or the ancestral TnpB, IscB, IsrB were initially included. The sequences were further enriched with HMMER (Zimmermann et al., 2018) and BLAST search (Altschul et al., 1990), which were additionally validated for structural integrity with AF2 (Jumper et al., 2021). Other proteins coming from mobile genetic elements having putative similar mechanism were processed in a similar fashion (Altae-Tran

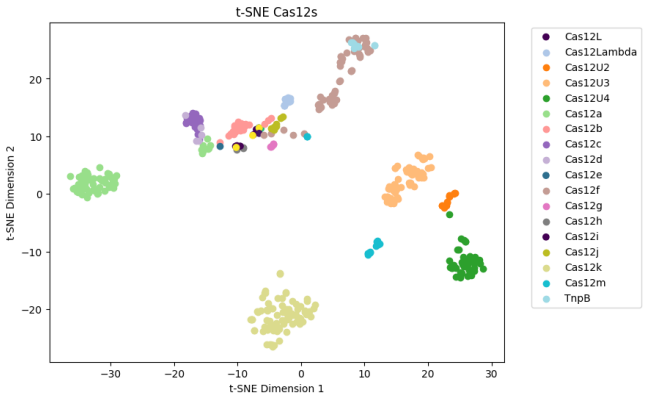

Figure 1. t-SNE plot of TOPH embeddings on subset of (Makarova et al., 2020), focusing on Cas12 proteins.

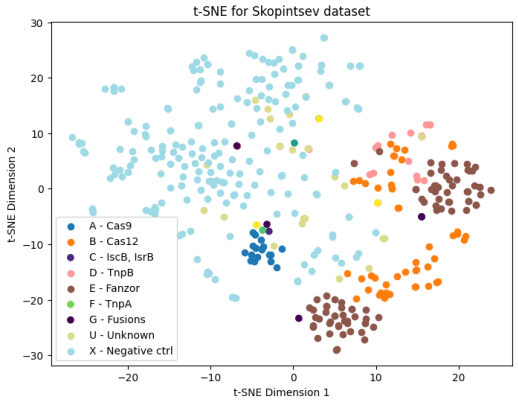

Figure 2. t-SNE plot of TOPH embeddings on the Skopintsev dataset.

et al., 2021). As expected, Cas9's and Cas12's fell into distinct groups 2. Interestingly, Cas12's showed higher diversity than Cas9's and separated into a few distinct, but related clusters, some of which were categorized together with the evolutionary related TnpBs, and other putative RNA-guided nucleases. The negative control ("X") dataset was composed of random Protein Data Bank entries (wwp, 2019), and did not co-localize with the Cas9's or Cas12's clusters.

4. Discussion

We have demonstrated preliminary results applying ideas from open domain question answering to the problem of detecting remote homologs. Our model, without being trained on a larger dataset, any hyperparameter tuning, or use of protein structures, is able to achieve comparable performance to Foldseek. Furthermore, we are able to show our model is able to reproduce evolutionary classifications of CRISPR-Cas proteins produced by biologists, despite being trained on single protein domains in SCOPe. We hope that our CRISPR-Cas datasets can serve as a valuable benchmark dataset for the remote homology problem, and plan to quantify the dataset further with metrics such as sequence identities.

In our DPR training, we make use of its split encoder architecture, which is currently unnecessary because *questions* and *passages* are both simply SCOPe domains. Classically, DPR models involve using a retriever to get possible contexts the contain an answer to a given question. For the protein universe, we can think of this as a question being a single- domain protein and passage being a multi-domain protein. Further incorporating DPR analogies, we can then build a **reader** to select an *answer*, ie annotate protein domains within a sequence or structure.

There are numerous ways to improve model performance, aside from doing hyperparameter tuning. We chose to only train on sequences, as sequences are far more abundant in public databases and require less computational power to process. However, we may condition on both sequence and structure, as per (Lu et al., 2023), or use embeddings from inverse folding models (Hsu et al., 2022). It may be likely to further improve model performance by training on a larger dataset such CATH. We hope to adopt a curriculum learning strategy, gradually fine-tuning on increasingly complex tasks within the SCOPe or CATH hierarchy. Although we do not evaluate speed directly for this workshop, our method runs with comparable speeds to the ESM models that are used to initialize it. Similar to DPR, we will integrate the FAISS package (Johnson et al., 2019) for the speed benefits it confers.

In summary, we apply techniques from QA to detect homologous proteins, introduce a new CRISPR-Cas dataset for evaluation, and have created a user-friendly web interface for biologists. These preliminary results, we hope, can help springboard biology into the era of deep learning for retrieval.

References

Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research*, 47(D1): D520–D528, 2019.

Al-Shayeb, B., Skopintsev, P., Soczek, K. M., Stahl, E. C., Li, Z., Groover, E., Smock, D., Eggers, A. R., Pausch, P., Cress, B. F., et al. Diverse virus-encoded crispr-cas systems include streamlined genome editors. *Cell*, 185 (24):4574–4586, 2022.

Altae-Tran, H., Kannan, S., Demircioglu, F. E., Oshiro, R., Nety, S. P., McKay, L. J., Dlakić, M., Inskeep,

W. P., Makarova, K. S., Macrae, R. K., et al. The widespread is200/is605 transposon family encodes diverse programmable rna-guided endonucleases. *Science*, 374(6563):57–65, 2021.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

Fox, N. K., Brenner, S. E., and Chandonia, J.-M. SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Research*, 42(D1):D304–D309, 12 2013. ISSN 0305-1048. doi: 10.1093/nar/gkt1240. URL https://doi.org/10.1093/nar/gkt1240.

Greener, J. G. and Jamali, K. Fast protein structure searching using structure graph embeddings. *bioRxiv*, pp. 2022–11, 2022.

Hamamsy, T., Morton, J. T., Berenberg, D., Carriero, N., Gligorijevic, V., Blackwell, R., Strauss, C. E., Leman, J. K., Cho, K., and Bonneau, R. Tm-vec: template modeling vectors for fast homology detection and alignment. *bioRxiv*, pp. 2022–07, 2022.

Harrington, L. B., Burstein, D., Chen, J. S., Paez-Espino, D., Ma, E., Witte, I. P., Cofsky, J. C., Kyrpides, N. C., Banfield, J. F., and Doudna, J. A. Programmed dna destruction by miniature crispr-cas14 enzymes. *Science*, 362(6416):839–842, 2018.

Hirano, S., Kappel, K., Altae-Tran, H., Faure, G., Wilkinson, M. E., Kannan, S., Demircioglu, F. E., Yan, R., Shiozaki, M., Yu, Z., et al. Structure of the omega nickase isrb in complex with ωrna and target dna. *Nature*, 610(7932):575–581, 2022.

Holm, L. Using dali for protein structure comparison. *Structural Bioinformatics: Methods and Protocols*, pp. 29–42, 2020.

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. In *International Conference on Machine Learning*, pp. 8946–8970. PMLR, 2022.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., and Charpentier, E. A programmable dual-rna–guided dna endonuclease in adaptive bacterial immunity. *science*, 337(6096):816–821, 2012.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7 (3):535–547, 2019.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and tau Yih, W. Dense passage retrieval for open-domain question answering, 2020.

Karvelis, T., Druteika, G., Bigelyte, G., Budre, K., Zedaveinyte, R., Silanskas, A., Kazlauskas, D., Venclovas, Č., and Siksnys, V. Transposon-associated tnpb is a programmable rna-guided dna endonuclease. *Nature*, 599 (7886):692–696, 2021.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv*, 2022.

Lu, A., Abeel, P., and Yang, K. Multimodal generation of protein sequence and structure by latent diffusion. 2023.

Makarova, K. S., Wolf, Y. I., Iranzo, J., Shmakov, S. A., Alkhnbashi, O. S., Brouns, S. J., Charpentier, E., Cheng, D., Haft, D. H., Horvath, P., et al. Evolutionary classification of crispr–cas systems: a burst of class 2 and derived variants. *Nature Reviews Microbiology*, 18(2):67–83, 2020.

Pausch, P., Al-Shayeb, B., Bisom-Rapp, E., Tsuchida, C. A., Li, Z., Cress, B. F., Knott, G. J., Jacobsen, S. E., Banfield, J. F., and Doudna, J. A. Crispr-casϕ from huge phages is a hypercompact genome editor. *Science*, 369(6501):333–337, 2020.

Sasnauskas, G., Tamulaitiene, G., Druteika, G., Carabias, A., Silanskas, A., Kazlauskas, D., Venclovas, Č., Montoya, G., Karvelis, T., and Siksnys, V. Tnpb structure reveals minimal functional core of cas12 nuclease family. *Nature*, pp. 1–6, 2023.

Schuler, G., Hu, C., and Ke, A. Structural basis for rna-guided dna cleavage by iscb-ωrna and mechanistic comparison with cas9. *Science*, 376(6600):1476–1481, 2022.

Shmakov, S., Smargon, A., Scott, D., Cox, D., Pyzocha, N., Yan, W., Abudayyeh, O. O., Gootenberg, J. S., Makarova, K. S., Wolf, Y. I., et al. Diversity and evolution of class 2 crispr–cas systems. *Nature reviews microbiology*, 15(3):169–182, 2017.

Steinegger, M. and Söding, J. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature biotechnology*, 35(11):1026–1028, 2017.

Strecker, J., Ladha, A., Gardner, Z., Schmid-Burgk, J. L., Makarova, K. S., Koonin, E. V., and Zhang, F. Rna-guided dna insertion with crispr-associated transposases. *Science*, 365(6448):48–53, 2019.

Trinquier, J., Petti, S., Feng, S., Söding, J., Steinegger, M., and Ovchinnikov, S. Swampnn: End-to-end protein structures alignment. *Machine Learning for Structural Biology Workshop, NeurIPS*, 2022.

Urbaitis, T., Gasiunas, G., Young, J. K., Hou, Z., Paulraj, S., Godliauskaite, E., Juskeviciene, M. M., Stitilyte, M., Jasnauskaite, M., Mabuchi, M., et al. A new family of crispr-type v nucleases with c-rich pam recognition. *EMBO reports*, 23(12):e55481, 2022.

van Kempen, M., Kim, S. S., Tumescheit, C., Mirdita, M., Lee, J., Gilchrist, C. L., Söding, J., and Steinegger, M. Fast and accurate protein structure search with foldseek. *Nature Biotechnology*, pp. 1–4, 2023.

Zhang, Y. and Skolnick, J. Tm-align: a protein structure alignment algorithm based on the tm-score. *Nucleic acids research*, 33(7):2302–2309, 2005.

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A. N., and Alva, V. A completely reimplemented mpi bioinformatics toolkit with a new hhpred server at its core. *Journal of molecular biology*, 430(15):2237–2243, 2018.