Computational and Systems Biology

Sensitive remote homology search by local alignment of small positional embeddings from protein language models

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted
October 27, 2023 (this version)

Sent for peer review
August 8, 2023

Posted to bioRxiv
July 29, 2023

Sean R. Johnson, **Meghana Peshwa**, **Zhiyi Sun**

New England Biolabs Inc., 240 County Road, Ipswich, MA 01938, United States

Abstract

Accurately detecting distant evolutionary relationships between proteins remains an ongoing challenge in bioinformatics. Search methods based on primary sequence struggle to accurately detect homology between sequences with less than 20% amino acid identity. Profile- and structure-based strategies extend sensitive search capabilities into this twilight zone of sequence similarity but require slow pre-processing steps. Recently, whole-protein and positional embeddings from deep neural networks have shown promise for providing sensitive sequence comparison and annotation at long evolutionary distances. Embeddings are generally faster to compute than profiles and predicted structures but still suffer several drawbacks related to the ability of whole-protein embeddings to discriminate domain-level homology, and the database size and search speed of methods using positional embeddings. In this work, we show that low-dimensionality positional embeddings can be used directly in speed-optimized local search algorithms. As a proof of concept, we use the ESM2 3B model to convert primary sequences directly into the 3Di alphabet or amino acid profiles and use these embeddings as input to the highly optimized Foldseek, HMMER3, and HH-suite search algorithms. Our results suggest that positional embeddings as small as a single byte can provide sufficient information for dramatically improved sensitivity over amino acid sequence searches without sacrificing search speed.

eLife assessment

This work presents the application of protein language models in combination with current methods for the detection of distant evolutionary relationships. While the results are **important**, they are supported by **incomplete**, preliminary evidence. A more comprehensive benchmark and clarification of technical details may turn this exploratory study into an algorithm ready for wide use.

Introduction

A common method for assigning a putative function to a protein sequence is to find sequences with experimentally determined functions that have similarities in sequence, structure, or evolutionary origin to the unannotated sequence (Loewenstein et al., 2009). Direct comparisons of primary sequence, for example using BLASTP (Camacho et al., 2009), are fast and reliable but show poor ability to detect homologs with less than about 20% identity to the query (Rost, 1999). Popular approaches for higher sensitivity sequence searches involve using sequence profiles, for example with PSI-BLAST (Altschul et al., 1997), HMMER3 (Eddy, 2011), HH-suite3 (Steinegger et al., 2019), or MMseqs2 (Steinegger and Söding, 2017). Sequence profiles are derived from multiple sequence alignments (MSAs) and are often modeled as profile hidden Markov models (HMMs), for example in HMMER and HH-suite. Profile HMMs model each position as the probability of each amino acid at the position together with insertion and deletion probabilities. Because of their reliance on the construction of MSAs, profile-based methods can have high computational overhead for database construction, query preparation, or both.

Protein structure searches also show higher sensitivity than sequence searches (Jambrich et al., 2023). Until recently, the utility of structure searches for protein annotation was limited by the lack of extensive reference databases and the inability to predict structures quickly and reliably for sequences lacking experimentally determined structures. In the past several years accurate protein structure prediction programs such as AlphaFold2 (Jumper et al., 2021) and ESMFold (Lin et al., 2023) have led to a massive increase in the size of databases of predicted protein structures. Coupled with fast structure search algorithms such as Foldseek (van Kempen et al., 2023), RUPEE (Ayoub and Lee, 2019), and Dali (Holm, 2022), structure prediction programs provide another powerful tool for remote homology detection. Foldseek achieves fast structure search by encoding the tertiary interactions of each amino acid in the 20-letter 3D interaction (3Di) alphabet. By using a structure alphabet of the same size as the amino acid alphabet, Foldseek can leverage optimized sequence search algorithms originally developed for amino acid sequences (Steinegger and Söding, 2017; van Kempen et al., 2023). Structure search methods suffer from some of the same drawbacks as profile-based methods, including the computational cost of converting primary sequences to predicted structures.

Emerging methods for protein annotation and remote homology detection rely on deep neural networks taking protein sequences as inputs and producing either a classification from a controlled vocabulary (Bileschi et al., 2022; Sanderson et al., 2023), a natural language description (Gane et al., 2022), positional embeddings, or a sequence embedding. Positional embeddings are fixed length vectors for each amino acid position of the protein. Positional embeddings produced by popular protein language models (pLMs) usually have large dimensions such as 1024 for ProtT5-XL-U50 (Elnaggar et al., 2021) and 2560 for ESM-2 3B (Lin et al., 2023). Sequence embeddings represent an entire sequence and are often calculated by element-wise averaging of the positional embeddings. Positional and sequence embeddings can be used for remote homology detection by using them to calculate substitution matrices in pairwise local alignments (Kaminski et al., 2022; Pantolini et al., 2022; Ye and Iovino, 2023), or by k-nearest neighbors searches (Hamamsy et al., 2022; Schütze et al., 2022), respectively.

While each of these emerging methods shows promise for improving sensitivity of protein search and annotation, they suffer various limitations. Classification models and methods relying on sequence embeddings struggle at discriminating individual domains of multi-domain proteins. Methods relying on large positional embeddings are space inefficient, and current search implementations are slow compared to other methods. Smaller positional embeddings would be more amenable to algorithmic optimizations using single instruction multiple data (SIMD) capabilities of central processing units (CPUs) that contribute to the speed of optimized sequence search algorithms (Buchfink et al., 2021; Eddy, 2011; Steinegger et al., 2019).

We recognized that profile HMMs and 3Di sequences are types of positional embeddings with dimensionality as low as 1 (3Di sequences) to about 25 (profile HMMs, including amino acid frequencies and state transition probabilities). We test the hypothesis that the ESM-2 3B pLM (Lin et al., 2023) could be used to directly convert primary amino acid sequences into profile HMMs compatible with HMMER3 or HH-suite, and 3Di sequences compatible with Foldseek, providing a sequence search workflow leveraging the speed and sensitivity advantages of profile and structure search algorithms with an accelerated query preparation step enabled by the pLM.

Results

ESM-2 was already pretrained on the masked language modeling task (Devlin et al., 2019; Lin et al., 2023) of predicting amino acid distributions at masked positions of input sequences, therefore no additional fine-tuning was necessary to induce it to produce probabilities compatible with profile HMM tools (**Figure 1A**). Positional amino acid frequencies predicted by ESM-2 3B resembled those found in MSAs built from sequence searches (**Figure 2**).

To produce Foldseek-compatible 3Di sequences, we trained a two-layer 1D convolutional neural network (CNN) to convert positional embeddings from the last transformer layer of ESM-2 3B into 3Di sequences, we then unfroze the last transformer layer and fine-tuned it together with the CNN. The fine-tuned model, which we call ESM-2 3B 3Di (**Figure 1B**), converted amino acids to 3Di with an accuracy of 64% compared to a test set of 3Di sequences derived from AlphaFold2 predicted structures.

To evaluate the capacity of small embeddings to improve search sensitivity, we generated predicted profiles and 3Di sequences from clustered Pfam 32 splits (Bileschi et al., 2022) and converted them into formats compatible with various search tools (**Figure 1C**). Pfam (Mistry et al., 2021) is a set of manually curated multiple sequence alignments of families of homologous protein domains. Some families presumed to have a common evolutionary origin are further grouped into clans. In the clustered splits, each family is divided into train and test groups such that each sequence in the test group has less than 25% identity to the most similar protein in the train group (Bileschi et al., 2022). The sensitivity of a search algorithm is evaluated by the ability to match sequences from the test groups to their corresponding train group at the family or clan level.

Methods using predicted profiles and those using predicted 3Di sequences both showed greater accuracy than phmmer (amino acid to amino acid) searches across all identity bins, and hmmscan (amino acid to profile HMM) searches on test sequences below 20% identity to the closest train sequence (**Figure 3**, compare lines 6 and 7 with lines 1 and 2). Foldseek searches where both the queries and database consisted of 3Di sequences produced by ESM-2 3B 3Di (line 7) performed the best overall. Foldseek considers both 3Di and amino acid sequences in its alignments and can therefore be conceptualized as using a 2-byte embedding. Running Foldseek in 3Di-only mode (line 8), a 1-byte embedding, led to a decrease in accuracy but still outperformed phmmer across all identity bins, and hmmscan on bins below 20% identity. We also tried creating HMMER3 profiles from predicted 3Di sequences (line 9). These performed worse than single-sequence Foldseek searches. Patching HMMER3 to use 3Di-derived background frequencies and Dirichlet priors (line 10) did not improve performance. Furthermore, we noticed that HMMER3 runs very slowly on 3Di sequences and profiles, presumably because the prefiltering steps were not optimized with 3Di sequences in mind.

While faster than MSA construction or full structure prediction, pLM embedding still have non-trivial computational overhead. This limits the possibility of using methods based on pLM embeddings to perform sensitive homology searches against large metagenomic databases such as the 2.4 billion sequence Mgnify database (Richardson et al., 2023). It would be desirable to have

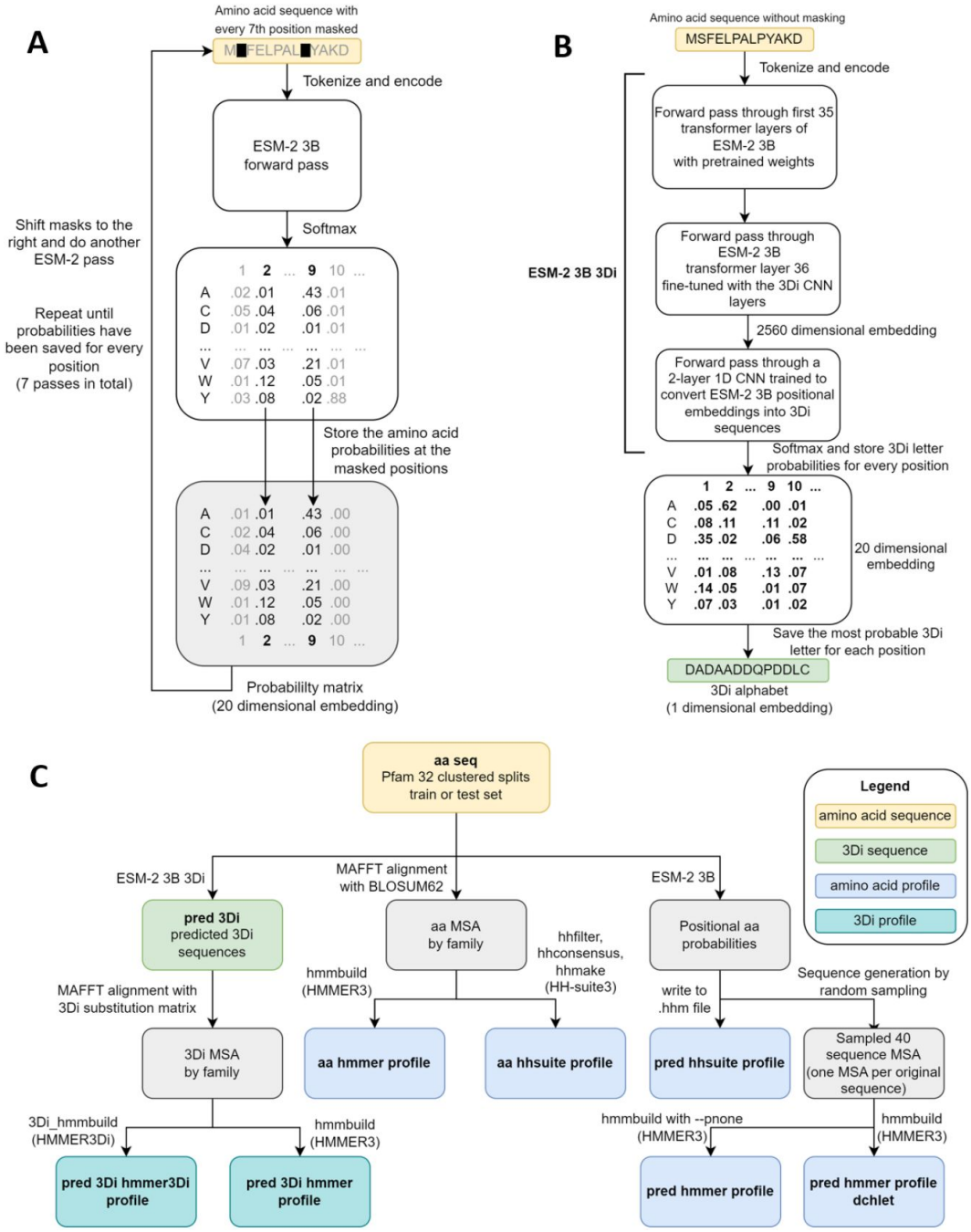


Figure 1.

Schematics of embedding models and the experimental design.

(**A**) ESM-2 3B can be directly used to predict amino acid probability distributions at masked positions. Our implementation uses seven passes. The second pass is shown in the figure. (**B**) ESM-2 3B 3Di, a fine-tuned ESM-2 3B with a small CNN top model can be used to predict 3Di sequences from amino acid sequences. (**C**) Data flow from amino acid sequences through embedding models and other programs to produce files used in homology searches. Bold words correspond to line labels in **Figure 3**.

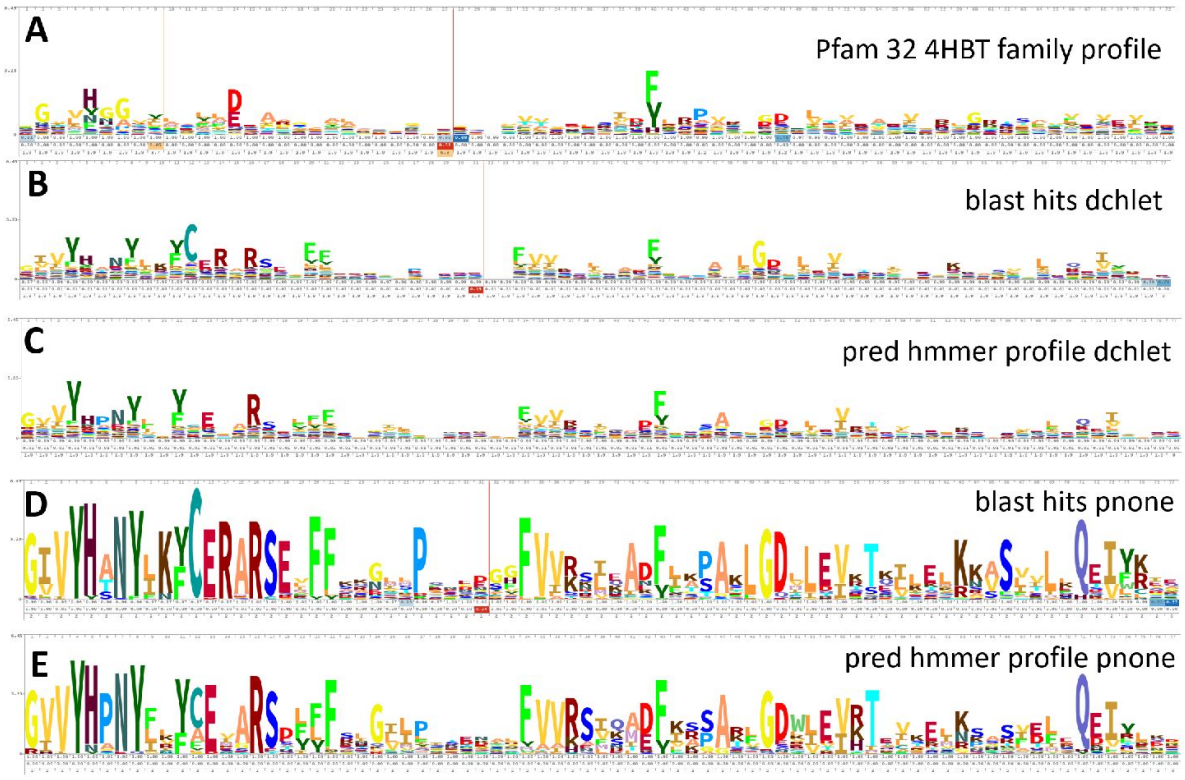
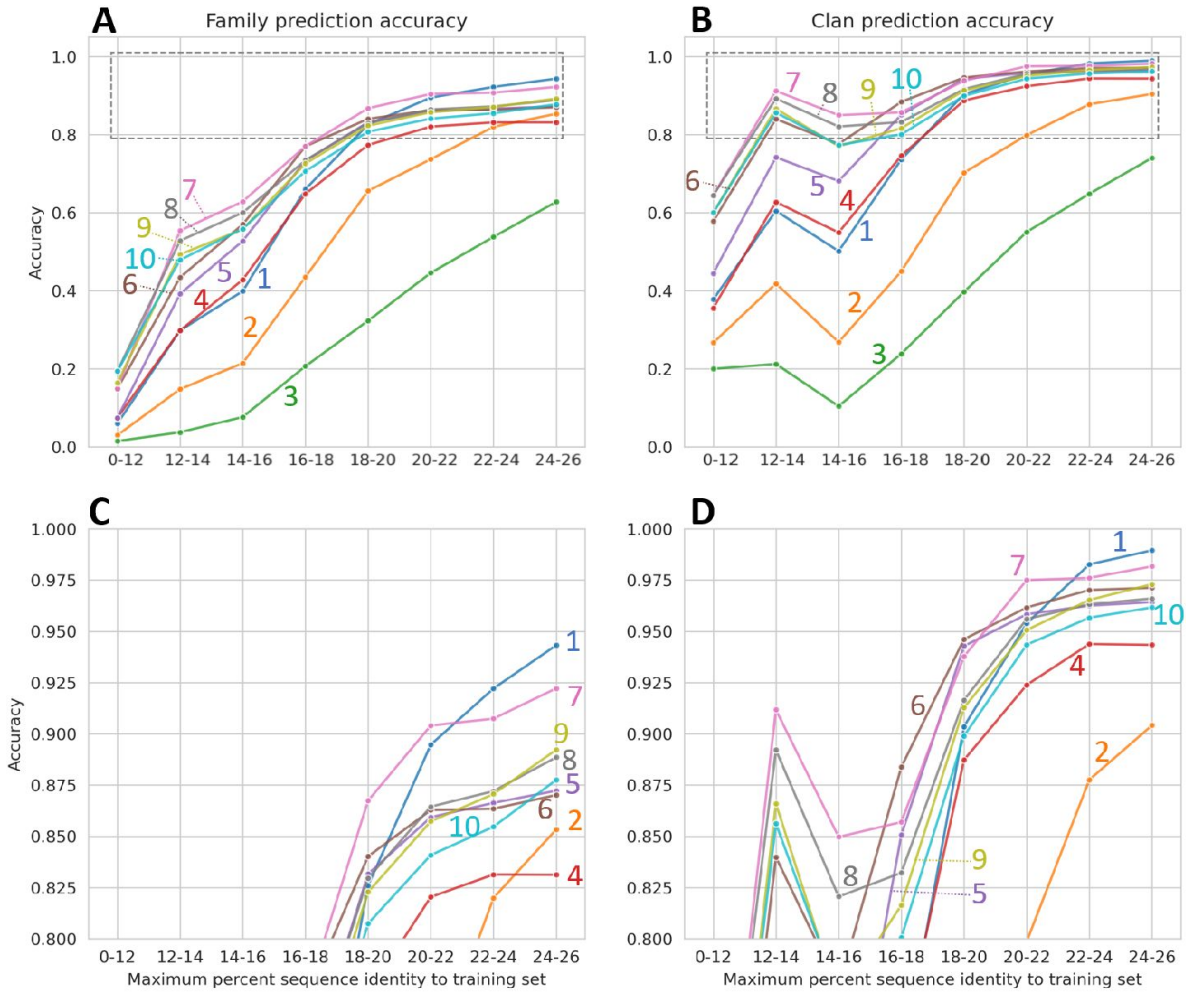


Figure 2.

Logos related to the example test sequence YBGC_HELPY__14-90 from the 4HBT family.

(**A**) 4HBT family hmm from Pfam 32. (**B**) hmmbuild with default settings on an MSA of the top 100 hits supplied by an online blast search (*blast.ncbi.nlm.nih.gov (http://blast.ncbi.nlm.nih.gov/)*) of YBGC_HELPY__14-90 against the NCBI clustered nr database. (**C**) hmmbuild with default settings on the MSA sampled from the ESM-2 3B positional probabilities for YBGC_HELPY__14-90. (**D**) and (**E**) hmmbuild with Dirichlet priors disabled on the same MSAs as for (**B**) and (**C**), respectively. All logos were generated by uploading the corresponding .hmm file to *skylign.org (http://skylign.org/)* (Wheeler et al., 2014).



Color	Line #	Algorithm	Query (test split)	Target (train split)	Family error rate	Number of errors	Clan error rate	Number of errors
	1	hmmscan	aa seq	aa hmmer profile	17.6	3744	10.7	1305
	2	phmmer	aa seq	aa seq	32.6	6942	27.3	3341
	3	hmmsearch	pred hmmer profile	aa seq	59.8	12740	50.9	6235
	4	hmmsearch	pred hmmer profile dchlet	aa seq	23.8	5078	12.7	1559
	5	hhblits	pred hhsuite profile	aa hhsuite profile	18.7	3979	7.8	955
	6	hhblits	pred hhsuite profile	pred hhsuite profile	17.7	3778	6.2	760
	7	foldseek	pred 3Di	pred 3Di	14.0	2982	5.5	672
	8	foldseek 3Di-only	pred 3Di	pred 3Di	17.6	3753	7.4	903
	9	hmmscan	pred 3Di	pred 3Di hmmer profile	18.4	3909	8.0	982
	10	3Di_hmmscan	pred 3Di	pred 3Di hmmer3Di profile	19.9	4234	9.0	1101

Figure 3.

Homology detection accuracy.

Test sequences were binned based on percent identity to the closest training sequence in the same family and annotated based on the top scoring hit from a search against the entire set of training sequences or training sequence family profiles, depending on the algorithm. (**A**) family recovery accuracy by bin. (**B**) clan recovery accuracy. (**C**) and (**D**) are detail views of the 0.8 to 1.0 accuracy range. There are a total of 21,293 test sequences. 12,246 test sequences have clan assignments.

search methods where the database can remain as amino acid sequences or other cheaply-calculated representations and the queries can be processed with more expensive methods. To this end, we tested hmmsearch using ESM2 3B generated profiles as queries against amino acid databases (lines 3 and 4). This is similar to a two-iteration PSI-BLAST (Altschul et al., 1997) or JackHMMER (Johnson et al., 2010) search where the first search and MSA-building step is replaced by a pLM embedding step. Curiously, hmmsearch using profiles built directly from the pLM probabilities (line 3) had the lowest accuracy of any algorithm. Nevertheless, profiles processed with hmmbuild from HMMER3, applying HMMER3 Dirichlet priors on top of the pLM probabilities (line 4), had better family prediction accuracy than phmmer at all but the highest identity bin, and accuracy on par with hmmscan at 18% identity and lower.

Discussion

Our results show that small positional embeddings produced by pLMs can enhance search sensitivity, potentially with less computational overhead than MSA construction or full structure prediction. While this manuscript was being submitted, another group reported an amino acid to 3Di translation model, ProstT5, that also showed strong performance on remote homology detection (Heinzinger et al., 2023). There are many possible directions for future development of improved embeddings, faster conversion and search programs, and comprehensive reference databases. Of particular interest could be conversion to profile HMM-style embeddings in a single pass, rather than the seven we required, and predicting state transition probabilities. Small positional embeddings optimal for local alignment could also be learned directly from a differentiable alignment algorithm (Petti et al., 2023) instead of relying on proxy tasks of amino acid or 3Di prediction. Finally, asymmetric architectures where embedding a database sequence is cheaper than embedding a query, analogous to searches with profile HMM queries against primary sequence databases, could be a powerful method for improving search sensitivity against large and growing reference databases. We've made our model training, search, and data analysis code publicly available. We hope our results and code will serve as a springboard for further exploration of the utility of low-dimensionality positional embeddings of protein sequences.

Methods

Alignments

Unless otherwise noted, multiple sequence alignments were made using MAFFT (v7.505) (Katoh and Standley, 2013) with options --anysymbol --maxiterate 1000 --globalpair. Protein alignments used the BLOSUM62 substitution matrix (Henikoff and Henikoff, 1992). 3Di alignments used the 3Di substitution matrix from Foldseek (van Kempen et al., 2023) (*https://github.com/steineggerlab/foldseek/blob/master/data/mat3di.out*).

Patching HMMER3 with background frequencies and Dirichlet priors for 3Di

We created a fork of the HMMER3 program, replacing amino acid background frequencies and Dirichlet priors with values calculated from the 3Di alphabet instead of the amino acid alphabet (*https://github.com/seanrjohnson/hmmer3di*). To generate a set of 3Di MSAs, we converted the AlphaFold UniProt Foldseek database (Jumper et al., 2021; van Kempen et al., 2023; Varadi et al., 2022) to a 3Di fasta file. We then looked up every sequence name from the Pfam 35 seed file in the UniProt 3Di fasta file and, for cases where the corresponding sequence was identifiable, extracted the sub-sequence corresponding to the Pfam 35 seed. 3Di seeds from each profile were aligned using MAFFT. MSA columns with more than 10 rows were used to calculate background

frequencies and Dirichlet priors using the HMMER3 program esl-mixdchlet fit with options -s 17 9 20. Amino acid background frequencies and Dirichlet priors in the HMMER3 source code were then replaced with the newly calculated 3Di background frequencies and Dirichlet priors. We call the patched HMMER3 as HMMER3Di.

Fine tuning ESM-2 3B to convert amino acid sequences into 3Di sequences

A 1D CNN was added on top of ESM-2 3B. The CNN takes as input position-wise embeddings from the last transformer layer of ESM-2 3B. The CNN consists of two layers, the first layer has 2560 input channels (the size of the embeddings from ESM-2 3B), and 300 output channels, kernel size 5, stride 1, padding 3. The second layer has 300 input channels and 21 output channels (one for each 3Di symbol plus a padding symbol), kernel size 5, stride 1, padding 1. The model was trained with a weighted cross-entropy loss function using weights of 0.1 * the diagonal from 3Di substitution matrix. The neural network was implemented in PyTorch (Paszke et al., 2019).

Training data was derived from the Foldseek AlphaFold2 UniProt50 dataset (Jumper et al., 2021; van Kempen et al., 2023; Varadi et al., 2022), a reduced-redundancy subset of the UniProt AlphaFold2 structures. The Foldseek database was downloaded using the Foldseek "databases" command line program, converted into protein and 3Di fasta files, filtered to remove sequences smaller than 120 amino acids and larger than 1000 amino acids, and split into train, validation, and test subsets, 90%:5%:5%, (33,924,764: 1,884,709: 1,884,710 sequences)

With ESM-2 layers frozen, the CNN was trained on the task of converting amino acids to 3Di sequences using the AdamW optimizer with learning rate 0.001, weight decay 0.001, and exponential learning rate decrease (gamma 0.98, applied every 100 batches). Training sequences were randomly selected in batches of 15 sequences. Training proceeded for 1301 batches, leading to a training accuracy of about 58%.

The last transformer layer of ESM-2 was then unfrozen and training restarted from the saved weights, with the same training parameters. Training continued for another 24001 batches of 10 random training sequences. Accuracy on the final training batch was 65%. Using the final trained weights, test sequences were converted to 3Di at an accuracy of 64.4%. We call the fine-tuned model ESM-2 3B 3Di.

The trained weights are available on Zenodo (*https://doi.org/10.5281/zenodo.8174959*). The training code is available on Github (*https://github.com/seanrjohnson/esmologs*). The model training code should be useful for fine-tuning ESM-2 to convert amino acid sequences to various other kinds of sequences, such as secondary structure codes, etc.

Pfam 32 clustered splits

Pfam 32 clustered splits (Bileschi et al., 2022) were downloaded from: *https://console.cloud.google.com/storage/browser/brain-genomics-public/research/proteins/pfam/clustered_split*. Data for mapping of individual Pfam 32 families to clans was downloaded from: *https://ftp.ebi.ac.uk/pub/databases/Pfam/releases/Pfam32.0/*. The sequences from the train, dev, and test splits were sorted into unaligned fasta files according to their split and family (**aa seq**).

Predicting 3Di sequences

Each training and test sequence was converted into a predicted 3Di sequence (**pred 3Di**) using the ESM-2 3B 3Di model described above. MAFFT alignments were made of both the amino acid and predicted 3Di training and test sequences for each family. HMMER3 profiles were built from the alignments using either unaltered HMMER3 (**pred 3Di hmmer profile**), or HMMER3Di (**pred 3Di hmmer3Di profile**).

Predicting profiles to generate HH-suite hhm files and HMMER3 hmm files from single sequences

Positional amino acid probabilities were calculated for unaligned train and test sequence using pre-trained ESM-2 3B in the following algorithm.

1. Prepend sequence with M. This is because ESM-2 3B has a strong bias towards predicting M as the first amino acid in every sequence, and most Pfam domains don't start with M.
2. Mask the first, eighth, etc. position in the sequence
3. Run a forward pass of ESM-2 3B over the masked sequence.
4. Save the logits for each of the 20 amino acid tokens for each masked position.
5. Shift the masks one position to the right.
6. Repeat steps 3 through 5 another 6 times until logits have been saved for every position.
7. Use the softmax function to calculate the amino acid probabilities at each position from the logits.

Note that in our actual implementation, a single forward pass was run on a batch of seven copies of the input sequence, each with different masking.

The positional probabilities were written directly as an HH-suite compatible .hhm file (**pred hhsuite profile**). A 40 sequence fasta MSA file was written where each sequence was randomly sampled from the probability distribution. Hmmbuild was run with default settings on the sampled MSA (**pred hmmer profile dchlet)** and with the -pnone setting, which disables adjustments to the probability distribution based on Dirichlet priors (**pred hmmer profile**).

Building HH-suite hhm and HMMER3 hmm profiles from train amino acid MSAs

The amino acid MSA for each training family were converted into an HH-suite database (**aa hhsuite profile**) with the following bash script:

```
echo '#'$profile_name > msa/${profile_name}.a3m hhfilter -i $MSA_FASTA -a msa/${profile_name}.a3m -M 50 -id 90; hhconsensus -i msa/${profile_name}.a3m -o consensus/${profile_name}.a3m hhmake -name $base -i consensus/${profile_name}.a3m -o hhm/${base}.hhm ffindex_build -s db_hhm.ffdata db_hhm.ffindex hhm ffindex_build -s db_a3m.ffdata db_a3m.ffindex consensus cstranslate -f -x 0.3 -c 4 -I a3m -i db_a3m -o db_cs219
```

HMMER3 profiles were built by calling hmmbuild with default settings on the training family amino acid MSAs (**aa hmmer profile**).

Building HH-suite databases from predicted profiles

HH-suite databases were built from predicted profiles .hhm files and the corresponding sampled 40 sequence MSA fasta files using the following bash script:

```
ffindex_build -s db_hhm.ffdata db_hhm.ffindex esm2_3B_profiles ffindex_build -s db_a3m.ffdata db_a3m.ffindex esm2_3B_sampled_msas cstranslate -f -x 0.3 -c 4 -I a3m -i db_a3m -o db_cs219
```

Building Foldseek databases from predicted 3Di sequences

Amino acid and predicted 3Di fasta files were converted into Foldseek-compatible databases using a new script, fasta2foldseek.py, available from the esmologs python package (see below).

Hmmscan, phmmer, and hmmsearch HMMER3 searches

In an attempt to mimic the Top pick HMM strategy reported by (Bileschi et al., 2022) we ran all HMMER3 searches in up to two iterations. The first iteration was run with default settings. For test sequences where no hits were detected among the training sequences or profiles, depending on the program, a second iteration was run with the addition of parameters intended to maximize sensitivity at the expense of search speed:

--max -Z 1 --domZ 1 -E 1000000 --domE 1000000

It should be noted that while our phmmer results are directly comparable to the phmmer results reported by Bileschi et al., our hmmscan results are not directly comparable to the reported "Top Pick HMM" results because we re-aligned the training sequences for each family instead of using the Pfam seed alignments. Still our results were very similar. We observed a 17.6% error rate (3744 test sequences with mispredicted family assignments) by hmmscan, compared to the reported 18.1% error rate (3844 mispredictions).

3Di_hmmscan HMMER3Di searches

3Di_hmmscan searches were performed using the same two iteration method described above for searches using standard HMMER3 programs.

Hhblits HH-suite searches

Hhbilts was run with the options:

-tags -n 1 -v 0

Foldseek searches

After converting both query and target amino acid and predicted 3Di fasta files into Foldseek compatible databases (see above), Foldseek searches were run with the following commands:

foldseek search test_db train_db foldseek_results tmpFolder foldseek convertalis test_db train_db aln_3Di_only ../hits.tsv --format- output query,target,bits

For 3Di-only searches, the option --alignment-type 0 was added to the search call.

Data and Code availability

HMMER3 patched with 3Di background frequencies and Dirichlet priors: *https://github.com/seanrjohnson/hmmer3di*

Code for neural network training, sequence searches, and data analysis: *https://github.com/seanrjohnson/esmologs*

Model weights for ESM-2 3B 3Di, predicted profiles and 3Di sequences from the Pfam 32 clustered splits, and other data necessary to reproduce the analyses: *https://doi.org/10.5281/zenodo.8174959*

Acknowledgements

We thank Sergey Ovchinnikov for helpful discussion about protein language models, Sean Eddy for helpful discussion about HMMER3, and Gary Smith for IT support.

References

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs** *Nucleic Acids Research* **25**:3389–3402 https://doi.org/10.1093/nar/25.17.3389

Ayoub R, Lee Y. (2019) **RUPEE: A fast and accurate purely geometric protein structure search** *PLOS ONE* **14** https://doi.org/10.1371/journal.pone.0213712

Bileschi ML, Belanger D, Bryant DH, Sanderson T, Carter B, Sculley D, Bateman A, DePristo MA, Colwell LJ (2022) **Using deep learning to annotate the protein universe** *Nat Biotechnol 1–6* https://doi.org/10.1038/s41587-021-01179-w

Buchfink B, Reuter K, Drost H-G. (2021) **Sensitive protein alignments at tree-of-life scale using DIAMOND** *Nat Methods* **18**:366–368 https://doi.org/10.1038/s41592-021-01101-x

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) **BLAST+: architecture and applications** *BMC Bioinformatics* **10** https://doi.org/10.1186/1471-2105-10-421

Devlin J, Chang M-W, Lee K, Toutanova K. (2019) **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**

Eddy SR (2011) **Accelerated Profile HMM Searches** *PLOS Computational Biology* **7** https://doi.org/10.1371/journal.pcbi.1002195

Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Yu W, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B. (2021) **ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing** *IEEE Transactions on Pattern Analysis and Machine Intelligence 1–1* https://doi.org/10.1109/TPAMI.2021.3095381

Gane A, Bileschi ML, Dohan D, Speretta E, Héliou A, Meng-Papaxanthos L, Zellner H, Brevdo E, Parikh A, Orchard S. (2022) **Gane A, Bileschi ML, Dohan D, Speretta E, Héliou A, Meng-Papaxanthos L, Zellner H, Brevdo E, Parikh A, Orchard S. 2022. ProtNLM: Model-based Natural Language Protein Annotation. https://www.uniprot.org/help/ProtNLM**

Hamamsy T, Morton JT, Berenberg D, Carriero N, Gligorijevic V, Blackwell R, Strauss CEM, Leman JK, Cho K, Bonneau R. (2022) **Hamamsy T, Morton JT, Berenberg D, Carriero N, Gligorijevic V, Blackwell R, Strauss CEM, Leman JK, Cho K, Bonneau R. 2022. TM-Vec: template modeling vectors for fast homology detection and alignment. doi:10.1101/2022.07.25.501437** https://doi.org/10.1101/2022.07.25.501437

Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Steinegger M, Rost B. (2023) **Heinzinger M, Weissenow K, Sanchez JG, Henkel A, Steinegger M, Rost B. 2023. ProstT5: Bilingual Language Model for Protein Sequence and Structure. doi:10.1101/2023.07.23.550085** https://doi.org/10.1101/2023.07.23.550085

Henikoff S, Henikoff JG (1992) **Amino acid substitution matrices from protein blocks** *Proc Natl Acad Sci U S A* **89**:10915–10919

Holm L. (2022) **Dali server: structural unification of protein families** *Nucleic Acids Research gkac387* https://doi.org/10.1093/nar/gkac387

Jambrich MA, Tusnady GE, Dobson L. (2023) **Jambrich MA, Tusnady GE, Dobson L. 2023. How AlphaFold shaped the structural coverage of the human transmembrane proteome. doi:10.1101/2023.04.18.537193** https://doi.org/10.1101/2023.04.18.537193

Johnson LS, Eddy SR, Portugaly E. (2010) **Hidden Markov model speed heuristic and iterative HMM search procedure** *BMC Bioinformatics* **11** https://doi.org/10.1186/1471-2105-11-431

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 SAA, Ballard AJ, 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 AW, Kavukcuoglu K, Kohli P, Hassabis D. (2021) **Highly accurate protein structure prediction with AlphaFold** *Nature 1–11* https://doi.org/10.1038/s41586-021-03819-2

Kaminski K, Ludwiczak J, Alva V, Dunin-Horkawicz S. (2022) **Kaminski K, Ludwiczak J, Alva V, Dunin-Horkawicz S. 2022. pLM-BLAST –distant homology detection based on direct comparison of sequence representations from protein language models. doi:10.1101/2022.11.24.517862** https://doi.org/10.1101/2022.11.24.517862

Katoh K, Standley DM (2013) **MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability** *Molecular Biology and Evolution* **30**:772–780 https://doi.org/10.1093/molbev/mst010

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, Rives A. (2023) **Evolutionary-scale prediction of atomic-level protein structure with a language model** *Science* **379**:1123–1130 https://doi.org/10.1126/science.ade2574

Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. (2009) **Protein function annotation by homology-based inference** *Genome Biology* **10** https://doi.org/10.1186/gb-2009-10-2-207

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, Finn RD, Bateman A. (2021) **Pfam: The protein families database in 2021** *Nucleic Acids Research* **49**:D412–D419 https://doi.org/10.1093/nar/gkaa913

Pantolini L, Studer G, Pereira J, Durairaj J, Schwede T. (2022) **Pantolini L, Studer G, Pereira J, Durairaj J, Schwede T. 2022. Embedding-based alignment: combining protein language models and alignment approaches to detect structural similarities in the twilight-zone. doi:10.1101/2022.12.13.520313** https://doi.org/10.1101/2022.12.13.520313

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. (2019) **Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A, Köpf A, Yang E, DeVito Z, Raison M, Tejani A, Chilamkurthy S, Steiner B, Fang L, Bai J, Chintala S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. doi:10.48550/arXiv.1912.01703** https://doi.org/10.48550/arXiv.1912.01703

Petti S, Bhattacharya N, Rao R, Dauparas J, Thomas N, Zhou J, Rush AM, Koo P, Ovchinnikov S. (2023) **End-to-end learning of multiple sequence alignments with differentiable Smith–Waterman** *Bioinformatics* **39** https://doi.org/10.1093/bioinformatics/btac724

Richardson L, Allen B, Baldi G, Beracochea M, Bileschi ML, Burdett T, Burgin J, Caballero-Pérez J, Cochrane G, Colwell LJ, Curtis T, Escobar-Zepeda A, Gurbich TA, Kale V, Korobeynikov A, Raj S, Rogers AB, Sakharova E, Sanchez S, Wilkinson DJ, Finn RD (2023) **MGnify: the microbiome sequence data analysis resource in 2023** *Nucleic Acids Research* **51**:D753–D759 https://doi.org/10.1093/nar/gkac1080

Rost B. (1999) **Twilight zone of protein sequence alignments** *Protein Engineering, Design and Selection* **12**:85–94 https://doi.org/10.1093/protein/12.2.85

Sanderson T, Bileschi ML, Belanger D, Colwell LJ (2023) **ProteInfer, deep neural networks for protein functional inference** *eLife* **12** https://doi.org/10.7554/eLife.80942

Schütze K, Heinzinger M, Steinegger M, Rost B. (2022) **Nearest neighbor search on embeddings rapidly identifies distant protein relations** *Frontiers in Bioinformatics* **2**

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. (2019) **HH-suite3 for fast remote homology detection and deep protein annotation** *BMC Bioinformatics* **20** https://doi.org/10.1186/s12859-019-3019-7

Steinegger M, Söding J. (2017) **MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets** *Nat Biotechnol* **35**:1026–1028 https://doi.org/10.1038/nbt.3988

van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, Söding J, Steinegger M. (2023) **Fast and accurate protein structure search with Foldseek** *Nat Biotechnol 1–4* 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, Velankar S. (2022) **AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models** *Nucleic Acids Research* **50**:D439–D444 https://doi.org/10.1093/nar/gkab1061

Wheeler TJ, Clements J, Finn RD (2014) **Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models** *BMC Bioinformatics* **15** https://doi.org/10.1186/1471-2105-15-7

Ye Y, Iovino BG (2023) **Ye Y, Iovino BG. 2023. Protein Embedding based Alignment (preprint). Preprints. doi:10.22541/au.168534397.72964200/v1** https://doi.org/10.22541/au.168534397.72964200/v1

Article and author information

Sean R. Johnson
New England Biolabs Inc., 240 County Road, Ipswich, MA 01938, United States
For correspondence: sjohnson@neb.com
ORCID iD: 0000-0001-8261-9015

Meghana Peshwa
New England Biolabs Inc., 240 County Road, Ipswich, MA 01938, United States

Zhiyi Sun
New England Biolabs Inc., 240 County Road, Ipswich, MA 01938, United States
For correspondence: sunz@neb.com
ORCID iD: 0000-0002-6106-5356

Copyright

Editors

Reviewing Editor
Ignacio Sanchez
CONICET / Universidad de Buenos Aires, Argentina

Senior Editor
Christian Landry
Université Laval, Canada

Reviewer #1 (Public Review):

Summary:

This work describes a new method for sequence-based remote homology detection. Such methods are essential for the annotation of uncharacterized proteins and for studies of protein evolution.

Strengths:

The main strength and novelty of the proposed approach lies in the idea of combining state-of-the-art sequence-based (HHpred and HMMER) and structure-based (Foldseek) homology detection methods with recent developments in the field of protein language models (the ESM2 model was used). The authors show that features extracted from high-dimensional, information-rich ESM2 sequence embeddings can be suitable for efficient use with the aforementioned tools.

The reduced features take the form of amino acid occurrence probability matrices estimated from ESM2 masked-token predictions, or structural descriptors predicted by a modified variant of the ESM2 model. However, we believe that these should not be called "embeddings" or "representations". This is because they don't come directly from any layer of these networks, but rather from their final predictions.

The benchmarks presented suggest that the approach improves sensitivity even at very low sequence identities <20%. The method is also expected to be faster because it does not require the computation of multiple sequence alignments (MSAs) for profile calculation or structure prediction.

Weaknesses:

The benchmarking of the method is very limited and lacks comparison with other methods. Without additional benchmarks, it is impossible to say whether the proposed approach really allows remote homology detection and how much improvement the discussed method brings over tools that are currently considered state-of-the-art.

Reviewer #2 (Public Review):

Summary:

The authors present a number of exploratory applications of current protein representations for remote homology search. They first fine-tune a language model to predict structural alphabets from sequence and demonstrate using these predicted structural alphabets for fast remote homology search both on their own and by building HMM profiles from them. They also demonstrate the use of residue-level language model amino acid predicted probabilities to build HMM profiles. These three implementations are compared to traditional profile-based remote homology search.

Strengths:

- Predicting structural alphabets from a sequence is novel and valuable, with another approach (ProstT5) also released in the same time frame further demonstrating its application for the remote homology search task.
- Using these new representations in established and battle-tested workflows such as MMSeqs, HMMER, and HHBlits is a great way to allow researchers to have access to the state-of-the-art methods for their task.
- Given the exponential growth of data in a number of protein resources, approaches that allow for the preparation of searchable datasets and enable fast search is of high relevance.

Weaknesses:

- The authors fine-tuned ESM-2 3B to predict 3Di sequences and presented the fine-tuned model ESM-2 3B 3Di with a claimed accuracy of 64% compared to a test set of 3Di sequences derived from AlphaFold2 predicted structures. However, the description of this test set is missing, and I would expect repeating some of the benchmarking efforts described in the Foldseek manuscript as this accuracy value is hard to interpret on its own.
- Given the availability of predicted structure data in AFDB, I would expect to see a comparison between the searches of predicted 3Di sequences and the "true" 3Di sequences derived from these predicted structures. This comparison would substantiate the innovation claimed in the manuscript, demonstrating the potential of conducting new searches solely based on sequence data on a structural database.
- The profile HMMs built from predicted 3Di appear to perform sub-optimally, and those from the ESM-2 3B predicted probabilities also don't seem to improve traditional HMM results significantly. The HHBlits results depicted in lines 5 and 6 in the figure are not discussed at all, and a comparison with traditional HHBlits is missing. With these results and presentation, the advantages of pLM profile-based searches are not clear, and more justification over traditional methods is needed.
- Figure 3 and its associated text are hard to follow due to the abundance of colors and abbreviations used. One figure attempting to explain multiple distinct points adds to the confusion. Suggestion: Splitting the figure into two panels comparing (A) Foldseek-derived searches (lines 7-10) and (B) language-model derived searches (line 3-6) to traditional methods could enhance clarity. Different scatter markers could also help follow the plots more easily.
- The justification for using Foldseek without amino acids (3Di-only mode) is not clear. Its

utility should be described, or it should be omitted for clarity.
- Figure 2 is not described, unclear what to read from it.