FlashFold: a standalone command-line tool for accelerated protein structure prediction

Chayan Kumar Saha [1,2*], Mohammad Roghanian [4], Susanne Häussler [2,3,4,5†], Lionel Guy [1†]

[1] Department of Medical Biochemistry and Microbiology, SciLifeLab, Uppsala University, Uppsala, Sweden

[2] Department of Molecular Bacteriology, Helmholtz Centre for Infection Research, Braunschweig, Germany

[3] Institute for Molecular Bacteriology, TWINCORE, Centre for Experimental and Clinical Infection Research, Hannover, Germany

[4] Department of Clinical Microbiology, Copenhagen University Hospital - Rigshospitalet, Copenhagen, Denmark

[5] Cluster of Excellence RESIST (EXC 2155), Hannover Medical School, Hannover, Germany

† equal contribution

[*] corresponding author:

Chayan Kumar Saha: chayan.saha@imbim.uu.se

ORCIDs:

Chayan Kumar Saha: 0000-0002-5357-517X

Mohammad Roghanian: 0000-0003-3116-3817

Susanne Häussler: 0000-0001-6141-9102

Lionel Guy: 0000-0001-8354-2398

Abstract

FlashFold is a fast, user-friendly tool for predicting protein structures and complexes. It can be installed locally on Linux or Mac operating systems, alleviating the computational needs, and eliminating reliance on third-party servers. FlashFold relies on a highly compact and adaptable built-in database to speed up the homology searching. Overall structure predictions are up to 3-fold faster than AlphaFold3 with sacrificing little or no accuracy. FlashFold is available at https://github.com/chayan7/flashfold.

Main

Proteins are responsible for all essential cellular functions and their tertiary structure is key to understanding their role. However, resolving protein structures using conventional experimental approaches like X-ray crystallography or Cryo-EM requires significant amount of time and effort, and thus researchers have long attempted to use *in silico* solutions instead. In 2020, DeepMind released AlphaFold2, an advanced machine-learning tool that predicts three-dimensional structures of single-chain proteins with unmatched accuracy, outperforming all competitors at 14th Critical Assessment of protein Structure Prediction competition (CASP)[1]. AlphaFold-Multimer, a later extension of AlphaFold2, could accurately predict the structure of protein complexes in many cases, allowing researchers to explore protein-protein interactions (PPI)[2]. To predict protein structures, AlphaFold relies on a novel end-to-end deep neural network that utilizes evolutionary information from multiple sequence alignments (MSA) of homologous sequences as its main input. Therefore, building reliable MSA is crucial, as the diversity of MSA influences the prediction accuracy[1,3].

For finding homologues as input, AlphaFold2 utilizes highly sensitive algorithms, HMMer[4] and HHblits[5], to search a comprehensive collection of protein sequences from publicly available resources and environmental databases. Due to the large data volume, researchers require about 3 TB storage on a local computer, and encounter extended search times as MSA construction often takes hours, significantly impacting overall prediction time[6]. Additionally, for executing the deep neural network algorithm on query proteins usually having size of about 1000 residues, AlphaFold2 requires graphics processing units (GPUs) with a considerable quantity of GPU RAM (Random Access Memory) of 16 GB[1,6].

To make protein structure prediction computationally more accessible and less expensive, ColabFold was released, that achieves similar accuracy as AlphFold2 and AlphaFold-Multimer, but in a fraction of the time[6,7]. Apart from the speed, ColabFold also stands out for its lack of GPU limitations. However, the presence of a parallel computing platform-compatible GPU (e.g. CUDA 12.1 or later for ColabFold version 1.5.5) speeds up structure inference and energy minimisation. ColabFold reduces database searching time by using the more efficient, server-based MMseqs2 homology detection software to produce MSA[6,8]. However, accessing the dedicated homology search server requires continuous internet connection. Tasks are also often standing in queue to start, due to the server offering only limited access for each user. Additionally, it is not recommended to use multiple computers to

access the server at the same time. To operate independently of the MMseqs2-based homology search server, users can download the smaller (1 TB) built-in ColabFold sequence database. However, the local ColabFold still requires high-performance computing, as the entire database must be loaded into RAM before homology detection begins, making the process time-consuming for single or few queries.

Recently, DeepMind released AlphaFold3[9], the third major version of AlphaFold, which introduces the ability to predict complex structure of proteins, nucleic acids, small molecules, ions and modified residues. Initially AlphaFold3 was only accessible via a dedicated commercial server that limited users to a maximum of 20 job submissions per day[9], making it unsuitable for high throughput analyses and the pseudocode raised controversy[10]. Eventually, the code was released, though not strictly in open-source fashion, since model weights – the core of the prediction software – are accessible only upon request. Nonetheless, users may now download and use the software for non-commercial purposes[11]. It can only be installed on a Linux-based computer having at least 64 GB of RAM, an expensive NVIDIA GPU with Compute Capability 8.0, and about 1TB disk space (preferably SSD) for storing the genetic database. Moreover, the built-in databases offered by AlphaFold3, or its previous version AlphaFold2 or ColabFold cannot be modified, for example by adding newly sequenced or private data.

Given the limitations of the currently available protein structure prediction tools discussed above, our aim was to develop a tool that can be run locally, using a compact database, allowing users to extend it with additional sequence from public or private resources, which might enrich the MSA information content. MSA enrichment has been shown to be a successful strategy, significantly improving the prediction accuracy of ColabFold[12]. Here, we present FlashFold, a python-based tool that; i) predicts the structure of protein monomers and multimers with an accuracy comparable to AlphaFold3, ii) provides a database at least an order of magnitude smaller than the ones offered by AlphaFold2, AlphaFold3, and ColabFold, and allows to build custom databases from public or private sequences, iii) can be installed on local computer or cluster running Linux or Mac operating system, and is independent of any third-party server, and iv) reduces GPU limitations and ensures proper utilization of computer resources by keeping CPU and GPU based calculation separated.

FlashFold consists of three main steps (Fig. 1). The first includes the sensitive homology detection using the Hidden Markov Model-based method Jackhmmer, part of the HMMer distribution[4], to construct a MSA including non-redundant sequences (Methods 3). Second, the MSA is used as an input for the structure prediction using the colabfold_batch script, part of the ColabFold package[6]. Third, a table is generated containing the scoring matrices for the predicted structures, including pDockQ2 score for protein complexes which estimates the quality of each interface in a multimer model[13].

Three compact sequence databases are included with FlashFold to facilitate quicker homology detection (Fig. 2a, see Method 2). To compare the structure prediction performance with AlphaFold2 and AlphaFold3, we used FlashFold to predict 15 monomeric and 10 dimeric proteins from *Pseudomonas aeruginosa* (Supplement Table 1, Methods 6.1). The predictions could be then compared to the structures already experimentally determined and published in Protein Data Bank (PDB)[14]. On average, FlashFold generated MSA considerably quicker than AlphaFold3 (Fig. 2b) making the overall prediction up to three-times faster (Fig. 2c) while maintaining reliable accuracy (Fig. 2d). FlashFold achieved a mean TM-score[15] of 0.90, compared to AlphaFold2 and AlphaFold3's 0.91, indicating a "very high" similarity between the predicted and experimentally resolved structures (see predicted hetero-complexes in Supplementary Fig. 1a, superimposed with the native structures using PyMol v2.5.4). Additionally, we used a dataset from ClusPro[16] – a web server dedicated to predicting models of the complex using the structures of interacting proteins – for benchmarking the multimeric prediction accuracy of FlashFold (Methods 6.2). FlashFold nearly always achieved TM-scores that were equivalent to those of AlphaFold2 and AlphaFold3 (Fig. 2e). Moreover, the predicted DockQ scores, a reliability metric for protein-protein docking models[17], were nearly always better with FlashFold when compared to AlphaFold-Multimer (Fig. 2f)[2,16]. For analysing the ClusPro dataset, we only used the FlashFold *rep_uni_db*, because the dataset contains query protein sequence from eukaryotes.

Another unique feature of the FlashFold databases is that they can be complemented by sequences from both public and private resources, e.g. from a library of genomes from a species of interest. We found this additional evolutionary information from both public and a minimal amount of privately sequence data, increased the TM-score on average by 0.074% (Methods 6.3, Supplementary Table 2).

Lastly, we evaluated the potential of FlashFold to predict novel PPIs. Using the criteria we defined (Methods 5), we validated seven previously identified interactions, six of which have resolved structures (Supplementary Fig. 1a-c). Furthermore, FlashFold could effectively distinguish between interacting and non-interacting protein partners (Supplementary Fig. 1d and 1e), highlighting its potential for large-scale identification of novel PPIs.

In summary, FlashFold offers a comprehensive, stand-alone solution for rapid and accurate structure prediction on a high-throughput scale, with great potential for uncovering novel PPIs. For researchers with access to private sequencing data, it serves as a valuable resource for exploring how additional sequence information may influence the structure of their protein of interest or its possible interactions.

Methods

1. Implementing FlashFold

FlashFold is a standalone program, with an emphasis on being computationally less demanding so that researchers can predict structures of proteins of interest faster on a high-throughput manner using their local machine. FlashFold is a user-friendly, command line tool which can be installed locally or on a server running Linux or Mac operating system. It can be installed from its source at https://github.com/chayan7/flashfold. It supports FASTA file as the primary input for protein sequences. It allows user to predict structures using the FlashFold-generated MSA (using "--only_msa" option) or user-specified custom MSA file in A3M format. In the latter case, the CPU-dependent homology detection step will be skipped. FlashFold also offers a batch prediction mode ("--batch" option), with which the user can input a directory that contains files formatted as FASTA or A3M.

2. FlashFold compact databases

To execute AlphaFold2, AlphaFold3 or ColabFold in a local machine, the user needs to download the databases which require a very large storage space (Fig 2a). Unlike these, FlashFold built-in databases are extremely compact. Using publicly available sequence information obtained from NCBI RefSeq[18], we created three unique databases for different application.

First, we built a database specific for bacterial proteins. The database is called *rep_bac_db*, ("representative bacterial database"). To build it, we downloaded GenBank files of all the representative bacterial genomes using FlashFold subcommand *ncbi_data*. Then, to create a database from the downloaded GenBank files, we used the FlashFold subcommand *create_db*. It generates a FASTA file named "sequence_db.fasta" containing unique protein sequences extracted from all GenBank files, which is then used as a database for homology detection. Each protein sequence is assigned to a unique hash value, which is used as a key for accession sequence related information. To determine which taxa the protein belongs to, a file called "protein_to_gbks.json" is being generated, which contains a list of taxa (as hash value) for each sequence hash key. It also generates another file called "prot_hash_to_accession.json" to store a list of protein accessions against each protein hash key. The *rep_bac_db* database is 42 GB in size and contains 76,005,099 protein sequences from 19,695 bacterial species.

Second, we created a *rep_prokvir_db*, short for "representative prokaryotic and viral database" which is an extended version of *rep_bac_db* with additional sequence information retrieved from archaea and viruses. We downloaded GenBank files of all the reference archaeal genomes. For viruses, we downloaded GenBank files from all the RefSeq annotated reference genomes. Then we extended the *rep_bac_db* using FlashFold subcommand *extend_db*. The rep_prok_db is 43 GB, contains 78,630,657 protein sequences from 35,220 taxa.

Third, we extended *rep_prokvir_db* with the sequence information from eukaryotes and named it as *rep_uni_db,* short for "representative universal database". As previously, we downloaded GenBank files of all the RefSeq annotated reference eukaryotic genomes and later used FlashFold subcommand *extend_db* for extension. This is the largest database among all three (82 GB; Fig. 2a), contains 124,790,329 protein sequences from 37,244 taxa.

The user can download these abovementioned built-in databases using FlashFold subcommand *download_db* and apply any of these based on their research interest. The initial databases correspond to NCBI RefSeq database release 226 (September 13, 2024). They will be regularly updated, following the bimonthly updating schedule of RefSeq; however, user can also independently create and maintain these databases using FlashFold database-related subcommands.

196 The ability for the user to update the database as needed is a unique characteristic of FlashFold
197 databases. Adding more sequences can influence the homology detection and the processing
198 time of the MSA (Fig. 2b), but it can be useful to discover novel structural folds or PPI.
199 FlashFold built-in databases are designed with an aim to make faster detection of distant
200 homologues of the raw input sequence. It is also possible to enrich the MSA with closely related
201 homologues by adding more information to the database. Below (Methods 6.3), we explain
202 how we extended the *rep_bac_db* with additional sequence information from both public and
203 private resources and checked its influence on prediction accuracy (Supplementary Table 2).

204 **3. Faster MSA construction with Jackhmmer**

205 The compact nature of FlashFold databases enables a much faster, yet sensitive homology
206 search. For this, FlashFold relies on Jackhmmer with the following parameters: --F1 0.0005 --
207 F2 0.00005 --F3 0.0000005 --incE 0.0001 -E 0.0001 -N 1. For each input sequence, it performs
208 a single iteration of alignment against the database (sequence_db.fasta file) by applying three
209 filtering stages (--F1, --F2, and --F3) to progressively refine the search and finally accepts hits
210 with an E-value ≤ 0.0001.
211
212 For single chain or monomer structure prediction, only non-redundant sequences are included
213 in the final MSA file, to ensure that a diversity of sequences are included. The MSA for homo-
214 oligomeric complexes is constructed in the same way as a monomer, but the MSA is duplicated
215 repeatedly for each component in the complex. For hetero-oligomeric complex structure
216 prediction, FlashFold executes parallel homology searches for each component, using the same
217 parameters mentioned before. It then creates one paired alignment and one unpaired alignment.
218 To create paired MSA, firstly, for each species, it collects the best hits from each component
219 and pairs them if the paired alignment covers at least 50% of the concatenated query sequence.
220 Finally, from all paired alignment, it selects only the non-redundant ones. Unpaired MSA is
221 created in the same way as described for monomer, with gaps in place of missing homologues
222 for component.

223 **4. Structure prediction using colabfold_batch**

224 For structure prediction, FlashFold relies on the colabfold_batch script that comes along with
225 ColabFold python package[6]. The custom alignment (A3M format) produced by FlashFold is
226 directly used as an input for colabfold_batch script. For single chain structure prediction, the
227 pLDDT score is used to rank the models. It is a per residue measure of local confidence that
228 spans from 1 to 100, with a higher score indicating a more accurate prediction[1]. For complexes,

229 it uses the predicted TM-score which is an integrated assessment of the overall predicted
230 structure of the complex[1]. To further evaluate the prediction quality, the output also includes
231 graphics – generated by colabfold_batch – that show the sequence coverage in the MSA and
232 AlphaFold2 confidence measures as in PAE that is also useful to rank and judge the quality of
233 the predicted structures[1,6].

234 **5. Predict PPI using the FlashFold scoring matrices**

235 FlashFold generates a scoring table for all the predicted models, ranked by prediction accuracy.
236 In addition, for complexes, it calculates pDockQ2 score (script available from
237 https://gitlab.com/ElofssonLab/huintaf2/-/blob/main/bin/pDockQ2.py) that estimates the
238 quality of each interface in a multimer model[13]. Protein complex models with a predicted TM-
239 score > 0.6 are generally considered of high quality[19]. Researchers have used predicted TM-
240 score alone, or along with pDockQ2 score to predict interaction between subunits within a
241 predicted multimer model[13,19,20]. Protein complexes are deemed reliable when the pDockQ2
242 score exceeds 0.23[13]. FlashFold calculates all these scores and summarizes the output in a
243 single table, offering the opportunity to confirm known, and discover novel, protein-protein
244 interaction. We used the following criteria to identify seven previously known interactions
245 (Supplementary Fig. 1), using:

246 (1) For heterodimers, the predicted model should have a predicted TM-score > 0.6 and the
247 minimum of the pDockQ2 scores calculated for two chains should be > 0.23
248 (2) For other complexes, the predicted model should have a predicted TM-score > 0.6 and the
249 mean of the pDockQ2 scores calculated for each chain should be > 0.23.

250 **6. Benchmark**

251 We utilised multiple datasets to benchmark FlashFold with AlphaFold2 and AlphaFold3,
252 measuring both computing resources and levels of prediction accuracy. We used a Linux
253 computer with Intel® Xeon® W-2295 18-core CPU, 64GB RAM and a NVIDIA GeForce
254 RTX 3060 GPU with 12GB RAM. No template was used for FlashFold-based structure
255 prediction. We have not considered benchmarking with ColabFold. In the case of ColabFold,
256 the sequence search is server-based and should be run in a local machine with at least 768GB
257 to 1TB RAM[6,9].
258

6.1 Benchmark with twenty-five *Pseudomonas aeruginosa* proteins and complexes

We randomly selected 25 proteins from *Pseudomonas aeruginosa* (Supplementary Table 1) whose structures were already experimentally determined and published in Protein Data Bank (PDB)[14]. We selected five homodimers, five heterodimers and 15 monomers. We extracted the sequences from UniProtKB Database and used them as input for structure prediction[21]. For a fair comparison of MSA construction and overall run time, only monomer structure prediction data were included (Fig 2b and 2c), and both FlashFold and AlphaFold used the same number of CPUs. We calculated TM-score, a sequence independent measure of structural similarity, for each predicted models against the experimental structures using TM-align[15] (https://zhanggroup.org/TM-align/) and used it for comparing the prediction accuracy (Fig 2d).

6.2 Benchmark with ClusPro dataset

We used both TM-align[15] and DockQ[17] to compare 17 ClusPro target predictions to their native structures[2,16]. The predicted models from FlashFold were ranked by predicted interface TM-score (default) as returned by AlphaFold-Multimer. TM-scores were calculated for each highest ranked predicted model as mentioned in Methods 6.1 and was utilized to compare the prediction accuracy between AlphaFold and FlashFold (Fig 2e). The reference DockQ scores for the AlphaFold-Multimer were gathered from the Mirdita, et al. [6] (Fig 2f) and for rest were calculated with the DockQ tool using the native and highest ranked predicted models as input.

6.3 Benchmark with extended FlashFold database

For this, we used the 25 previously selected proteins (Supplementary Table 1) from *P. aeruginosa* for structure prediction. We extended the *rep-bac-db* with publicly and privately available *P. aeruginosa* specific sequence information to investigate whether the addition of new pathogen specific data influences the prediction accuracy. Consequently, we started by adding all the *P. aeruginosa* GenBank files that are publicly accessible in the NCBI RefSeq database (a total of 100,53 as of October 21, 2024) to the *rep-bac-db* and renamed the database to *rep-bac-db+* which is 46 GB in size and contains 1,822,441 additional unique protein sequences. Furthermore, we extended *rep-bac-db+* to *rep-bac-db** by adding privately accessible GenBank files from a huge collection of clinical *P. aeruginosa* isolates (a total of 2,500 as of June 12, 2024, from Copenhagen Hospital). The *rep-bac-db** has 215,438 additional unique sequences than that of *rep-bac-db+,* and a size of 47 GB. We later used these extended databases to predict structure. For each predicted model, we calculated the TM-score as mentioned in Methods 6.1 and compared the scores with that of *rep-bac-db* (Supplementary Table 2).

Data availability

The three in-built databases of FlashFold, namely *rep-bac-db, rep_prokvir_db* and *rep_uni_db* are readily available for download from: https://doi.org/10.5281/zenodo.14717354, https://doi.org/10.5281/zenodo.14720530, and https://doi.org/10.5281/zenodo.14720238 respectively. All the structures predicted during benchmarking are available at: https://doi.org/10.5281/zenodo.14722213. The data used for generating the plots are available at: https://doi.org/10.5281/zenodo.14722254.

Code availability

FlashFold is freely available under the MIT license at https://github.com/chayan7/flashfold.

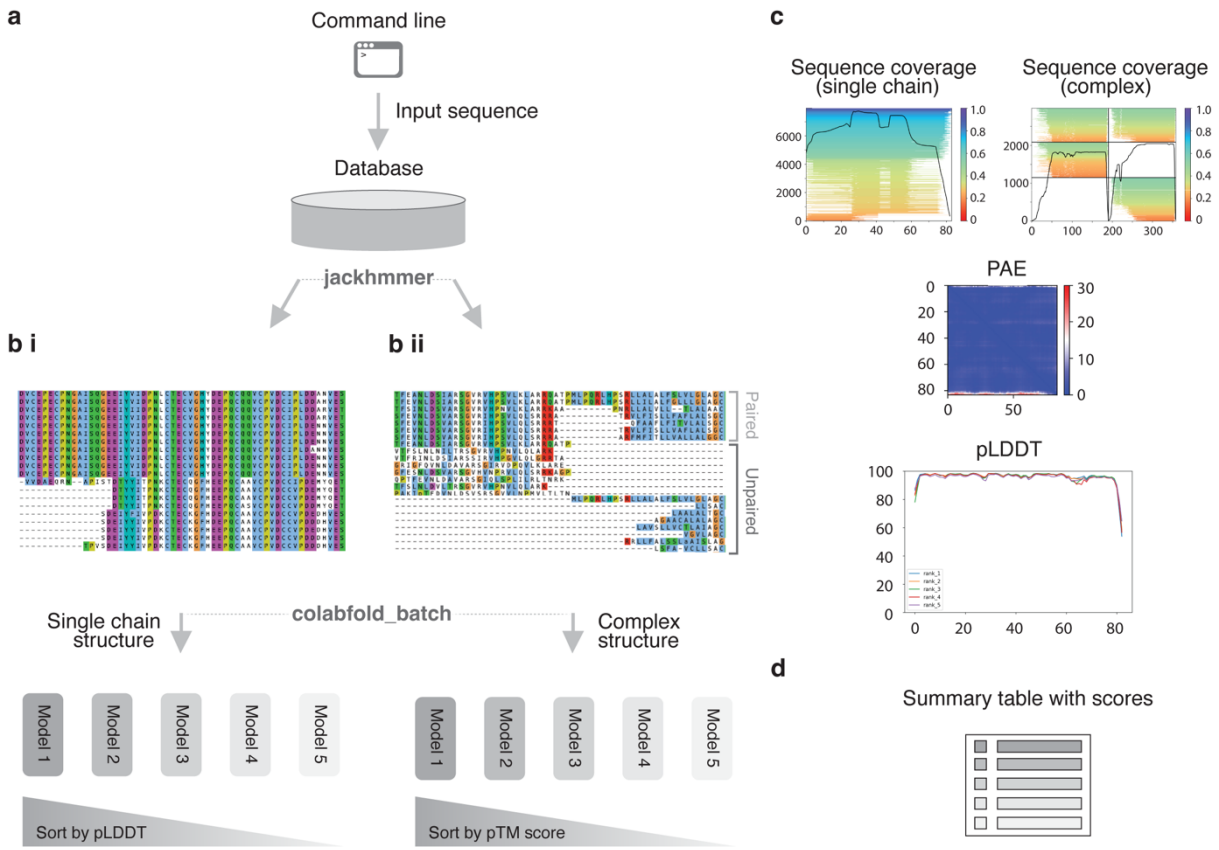


Figure 1 | Overview of FlashFold (adapted from Fig. 1 in Mirdita, et al. [6]). **(a)** FlashFold uses FASTA sequences as input to search for related sequences against a very compact built-in sequence database using Jackhmmer [4]. (**b**) Homologous sequences are filtered and processed as an MSA file in A3M format, and then sent to colabfold_batch for structure prediction: for single chain prediction (**i**), only non-redundant sequences from among all those detected are chosen while for complex structure prediction (**ii**), the final MSA is the combination of paired and unpaired MSA. Top hits (lowest e-values) from the same species are paired and only unique paired sequences are selected. Unpaired MSA is produced in the same way as in (i), with gaps in place of missing homologues. Single chain predictions are ranked by pLDDT score, while complexes are ranked by TM-score. (**c**) FlashFold output includes graphics that illustrate MSA depth and coverage, and AlphaFold2 confidence metrics such as PAE and pLDDT, to evaluate prediction quality. (**d**) FlashFold compiles a table of model quality scores and interface attributes at the end of the prediction phase.

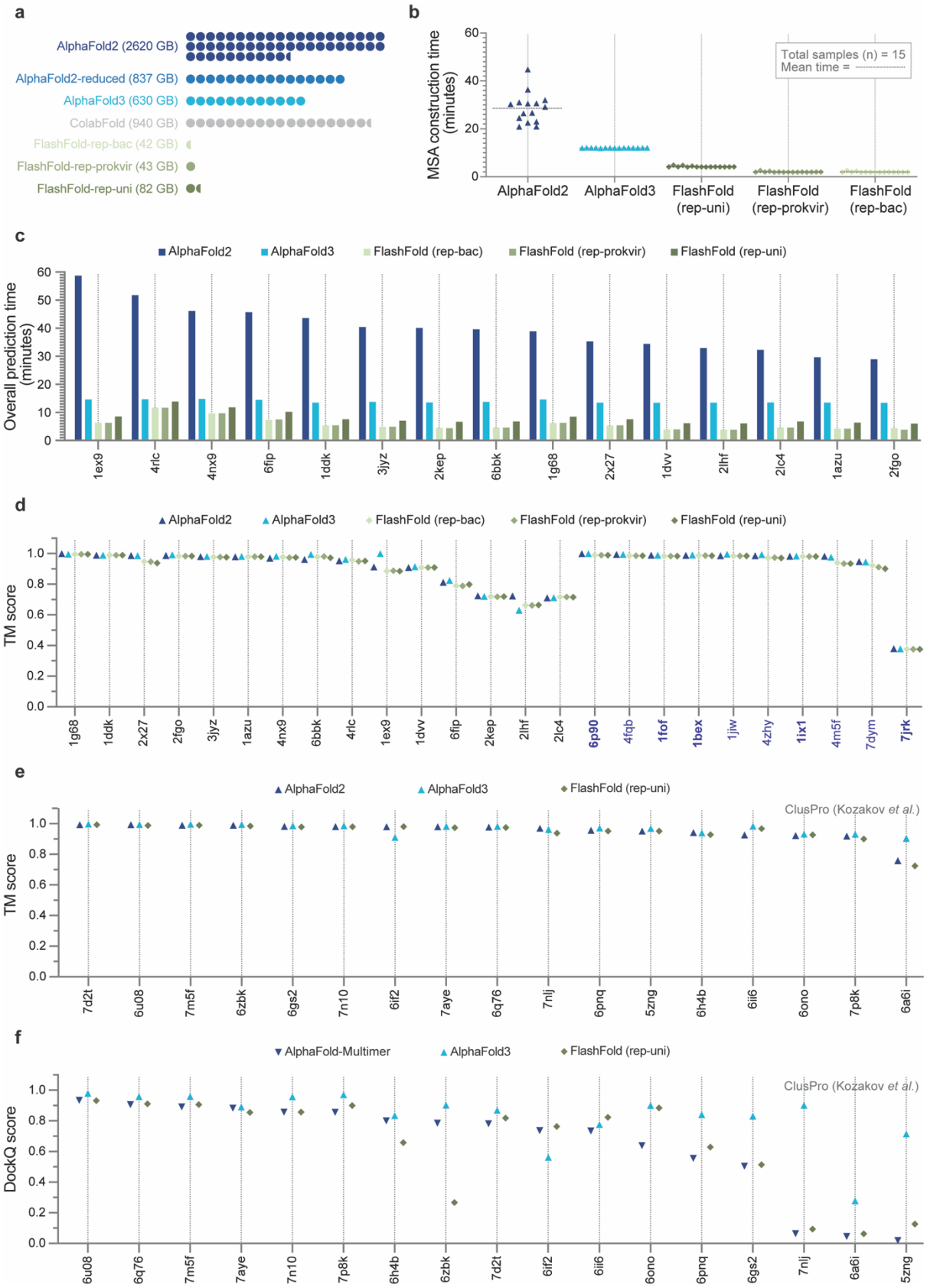


Figure 2 | Performance of FlashFold. (**a**) FlashFold offers three built-in databases which are extremely compact compared to the databases offered by AlphaFold2, AlphaFold3 and ColabFold. The largest of these three FlashFold databases, *rep_uni_db*, is generally applicable,

while *rep_prokvir_db* is designed for prokaryotes and viruses, and *rep_bac_db* for bacteria only. (**b**) A comparison of MSA construction times for AlphaFold2, Alphafold3 and FlashFold. (**c**) A comparison of monomer structures prediction times for AlphaFold2, Alphafold3 and FlashFold. (**d**) Comparison of AlphaFold2, AlphaFold3 and FlashFold prediction accuracy for monomers and dimer complexes (coloured in blue and homodimers in bold). (**e**) Evaluating the accuracy of multimeric structure prediction of AlphaFold2, AlphaFold3 and FlashFold using ClusPro dataset [2,16]. (**f**) Comparison of AlphaFold-Multimer, AlphaFold3 and FlashFold multimeric prediction modes on the ClusPro dataset [2,16]. A reliability metric for protein-protein docking models known as DockQ was calculated and compared.

References

1 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

2 Evans, R. *et al.* Protein complex prediction with AlphaFold-Multimer. *bioRxiv*, 2021.2010.2004.463034, doi:10.1101/2021.10.04.463034 (2022).

3 Zheng, W. *et al.* Improving deep learning protein monomer and complex structure prediction using DeepMSA2 with huge metagenomics data. *Nature Methods* **21**, 279-289, doi:10.1038/s41592-023-02130-4 (2024).

4 Eddy, S. R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195, doi:10.1371/journal.pcbi.1002195 (2011).

5 Steinegger, M. *et al.* HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics* **20**, 473, doi:10.1186/s12859-019-3019-7 (2019).

6 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nat Methods* **19**, 679-682, doi:10.1038/s41592-022-01488-1 (2022).

7 Kim, G. *et al.* Easy and accurate protein structure prediction using ColabFold. *Nature Protocols*, doi:10.1038/s41596-024-01060-5 (2024).

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

9 Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493-500, doi:10.1038/s41586-024-07487-w (2024).

10 Why did Nature publish AlphaFold3 without its code? *Nature* **629**, doi:https://doi.org/10.1038/d41586-024-01463-0 (2024).

11 Callaway, E. AI protein-prediction tool AlphaFold3 is now more open. *Nature* **635**, 531-532, doi:https://doi.org/10.1038/d41586-024-03708-4 (2024).

12 Lee, S. *et al.* Petascale Homology Search for Structure Prediction. *bioRxiv*, doi:10.1101/2023.07.10.548308 (2023).

13 Zhu, W., Shenoy, A., Kundrotas, P. & Elofsson, A. Evaluation of AlphaFold-Multimer prediction on multi-chain protein complexes. *Bioinformatics* **39**, doi:10.1093/bioinformatics/btad424 (2023).

14 Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res* **28**, 235-242, doi:10.1093/nar/28.1.235 (2000).

15 Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Research* **33**, 2302-2309, doi:10.1093/nar/gki524 (2005).

16 Kozakov, D. *et al.* The ClusPro web server for protein-protein docking. *Nat Protoc* **12**, 255-278, doi:10.1038/nprot.2016.169 (2017).

17 Basu, S. & Wallner, B. DockQ: A Quality Measure for Protein-Protein Docking Models. *PLoS One* **11**, e0161879, doi:10.1371/journal.pone.0161879 (2016).

18 O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**, D733-745, doi:10.1093/nar/gkv1189 (2016).

19 Gomez Borrego, J. & Torrent Burgas, M. Structural assembly of the bacterial essential interactome. *Elife* **13**, doi:10.7554/eLife.94919 (2024).

20 Sawa, T. *et al.* Comprehensive computational analysis of the SRK-SP11 molecular interaction underlying self-incompatibility in Brassicaceae using improved structure prediction for cysteine-rich proteins. *Comput Struct Biotechnol J* **21**, 5228-5239, doi:10.1016/j.csbj.2023.10.026 (2023).

21 UniProt, C. The universal protein resource (UniProt). *Nucleic Acids Res* **36**, D190-195, doi:10.1093/nar/gkm895 (2008).
22 Quinaud, M. *et al.* The PscE-PscF-PscG complex controls type III secretion needle biogenesis in Pseudomonas aeruginosa. *J Biol Chem* **280**, 36293-36300, doi:10.1074/jbc.M508089200 (2005).

Funding

C.K.S. is funded by the Swedish Research Council [Grant No. 2023-06633]. L.G. is funded by the Swedish Research Council [Grant no. 2023-04359] and the Carl Trygger Foundation [Grant no. CTS 21:1235]. S.H. is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2155 "RESIST" – Project ID 390874280, within the SFB/TRR-298-SIIRI – Project-ID 426335750 and in the SPP 2389 (HA 3299/9-1, AOBJ: 687646), from the Ministry of Science and Culture of Lower Saxony (Niedersächsisches Ministerium für Wissenschaft und Kultur) BacData, ZN3428, and by the Novo Nordisk Foundation (NNF 18OC0033946). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

All authors conceptualised and coordinated the study. C.K.S. developed the software and drafted the manuscript with contributions from all authors.

Ethics declarations

The authors have no conflicting interests.

401 **Supplementary information**

402

403 **Supplementary Table 1. Twenty-five randomly selected *Pseudomonas aeruginosa***
404 **proteins and complexes used for benchmarking.** The first 15 single chain proteins were
405 considered for run time benchmarking but all of them were included for the prediction accuracy
406 benchmark.

407

Serial	PDB ID	UniProt Accession	Gene Name	State
1	1AZU	P00282	*azu*	Monomer
2	1DDK	Q79MP6	*blaIMP-1*	Monomer
3	1DVV	P00099	*nirM*	Monomer
4	1EX9	P26876	*lip*	Monomer
5	1G68	P16897	*pse4*	Monomer
6	2FGO	Q9I6D2	*fdx1*	Monomer
7	2KEP	Q00514	*xcpT*	Monomer
8	2LC4	Q51354	*pilP*	Monomer
9	2LHF	Q51486	*oprH*	Monomer
10	2X27	Q9HWW1	*oprG*	Monomer
11	3JYZ	Q8KQ36	*pilA* (Pilin)	Monomer
12	4NX9	O86266	*fliC*	Monomer
13	4RLC	P13794	*oprF*	Monomer
14	6BBK	P18774	*pilA* (Fimbrial protein)	Monomer
15	6FIP	Q51368	*tonB*	Monomer
16	1BEX	P00282	*azu*	Homodimer
17	1FOF	P14489	*bla*	Homodimer
18	1IX1	Q9I7A8	*def*	Homodimer
19	6P90	Q9I4W3	*dapA*	Homodimer
20	7JRK	O68562	*bamE*	Homodimer
21	1JIW	Q03023 - Q03026	*aprA - inh*	Heterodimer
22	4FQB	Q9I2Q - Q9I2Q0	*tse1 - tsi1*	Heterodimer
23	4M5F	Q9HYC5 - Q9HYC4	*tse3 - tsi3*	Heterodimer
24	4ZHY	Q9I4L4 - Q9I4L6	*yfiR - yfiB*	Heterodimer
25	7DYM	Q9HXA6 - Q9HXA5	*PA3907 - PA3908*	Heterodimer

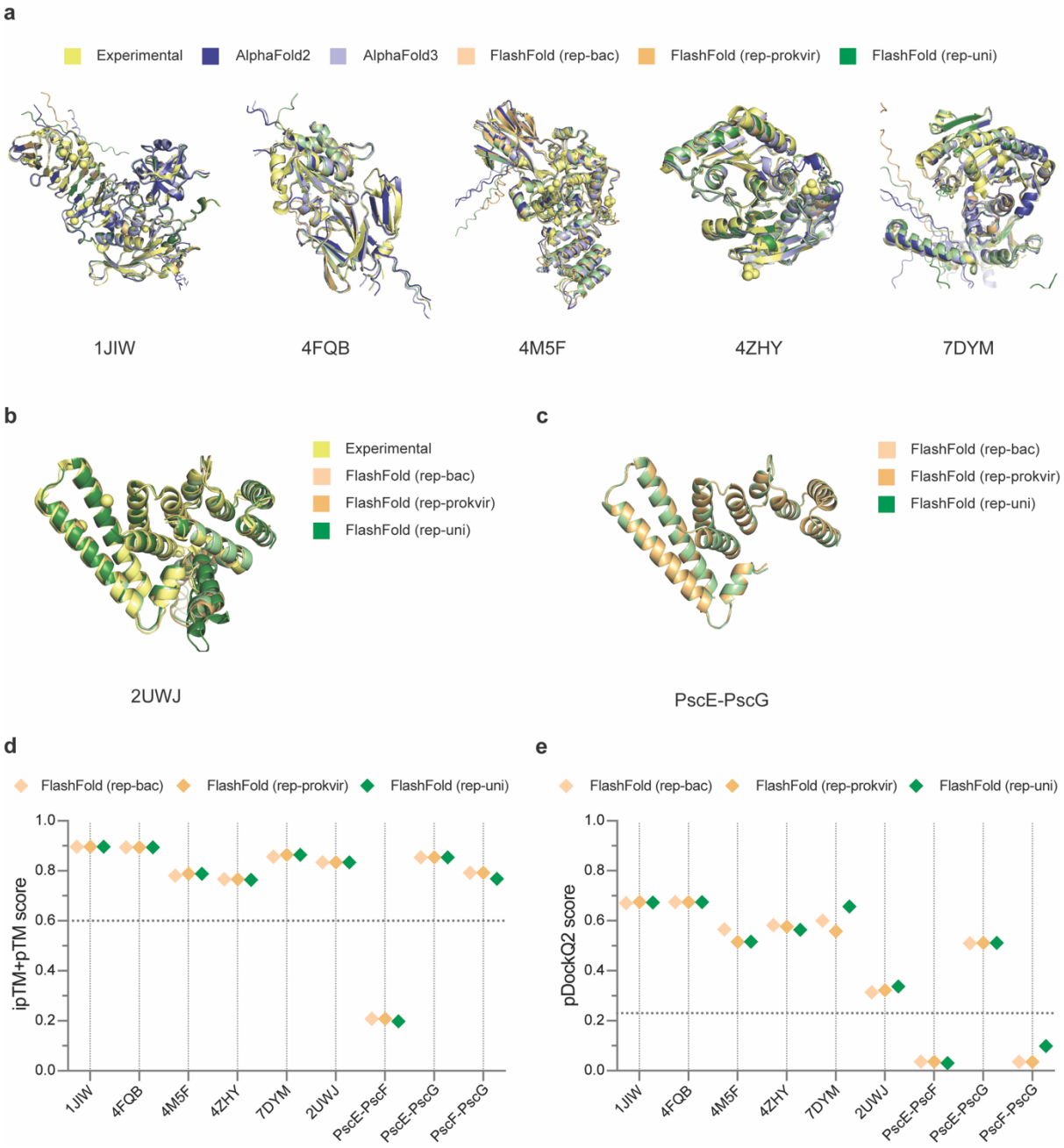
408

409

410 **Supplementary Table 2. FlashFold database extension with sequence from public and**
411 **private resources impacts prediction accuracy.** We used twenty-five *P. aeruginosa* proteins
412 and complexes (Supplementary Table 1) for this analysis. The TM scores were influenced in
413 many cases (increasing scores are shown as bold and decreasing scores are shaded grey) when
414 we used FlashFold with *rep-bac-db+* or *rep-bac-db** instead of *rep-bac-db*.
415

PDB ID	State	FlashFold		
		rep-bac-db	rep-bac-db+	rep-bac-db*
1azu	Monomer	0,97903	**0,97945**	**0,97917**
1ddk	Monomer	0,98915	0,98904	0,98904
1dvv	Monomer	0,9089	0,9072	0,90792
1ex9	Monomer	0,88507	**0,89325**	**0,89285**
1g68	Monomer	0,99577	0,99554	0,99545
2fgo	Monomer	0,98377	0,98336	0,98338
2kep	Monomer	0,7196	**0,72065**	**0,72011**
2lc4	Monomer	0,71684	0,71385	**0,71766**
2lhf	Monomer	0,66292	0,66252	**0,66439**
2x27	Monomer	0,94799	0,93437	0,93393
3jyz	Monomer	0,97693	**0,9772**	0,97688
4nx9	Monomer	0,97726	0,97693	0,97709
4rlc	Monomer	0,95727	**0,95867**	0,95631
6bbk	Monomer	0,97844	0,97001	0,97351
6fip	Monomer	0,79069	**0,79516**	**0,79288**
1bex	Homodimer	0,98764	**0,98802**	**0,98788**
1fof	Homodimer	0,98323	0,98309	**0,98411**
1ix1	Homodimer	0,98112	0,98075	0,98091
6p90	Homodimer	0,98967	0,98945	**0,99125**
7jrk	Homodimer	0,37599	**0,3895**	**0,39946**
1jiw	Heterodimer	0,98404	**0,98752**	**0,98408**
4fqb	Heterodimer	0,98627	**0,98686**	**0,99385**
4m5f	Heterodimer	0,94048	**0,96352**	**0,94547**
4zhy	Heterodimer	0,97168	0,97123	**0,97264**
7dym	Heterodimer	0,92332	0,87836	0,90976

416



Supplementary Figure 1. FlashFold scoring can be used to differentiate between interacting protein partners from the noninteracting ones. Except 2UWJ, all the other predicted models are heterodimers. 1JIW, 4FQB, 4M5F, 7DYM and 4ZHY are known heterodimer complex PDB IDs from *Pseudomonas aeruginosa* (Supplementary Table 1). 2UWJ is a PDB ID of a ternary (1:1:1) complex that includes PscE, PscF and PscG proteins. PscF is a primary part of the *P. aeruginosa* type three secretion needle, and researchers found that using gel-filtration techniques, a complex of its other two cytoplasmic partners, PscE and PscG, could be separated, suggesting a direct interaction between these two macromolecules [22]. Based on FlashFold prediction across its three built-in databases, (**a**) predicted models of

1JIW, 4FQB, 4M5F, 7DYM, 4ZHY, (**b**) PscE-PscF-PscG complex (2UWJ), and surprisingly (**c**) PscE-PscG complex (which lacks an experimentally resolved structure) all cleared the FlashFold PPI threshold (Methods 5) with the required predicted pTM score (**d**) and pDockQ2 score (**e**). It suggests that the FlashFold PPI threshold could identify protein pairs that do not directly interact such as PscE-PscF and PscF-PscG.