Deep Learning-based *Ab Initio* Protein Structure Prediction and Structure-based Protein Function Annotation

by

Chengxin Zhang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in the University of Michigan
2020

Doctoral Committee:

Professor Gilbert S. Omenn, Chair
Professor Heather Carlson
Assistant Professor Peter L. Freddolino
Associate Professor Yuanfang Guan
Associate Professor Melanie Ohi
Professor Emeritus Rudy J. Richardson

Chengxin Zhang

zcx@umich.edu

ORCID iD: 0000-0001-7290-1324

Dedication

This dissertation is dedicated to my wife, Dr. Xiaoqiong Wei, and our unborn child.

Acknowledgements

Firstly, I thank Dr. Yang Zhang for admitting me into his lab. His meticulous guidance in my early years of graduate study in terms of project design, server construction, and manuscript writing have all been invaluable. His world-renowned expertise in protein structure modeling has attracted many outstanding students and researchers to join his research efforts, including my wife, whom I met in our first month at the Yang Zhang lab.

I would extend my gratitude to all my committee members Dr. Gilbert S. Omenn, Dr. Peter L. Freddolino, Dr. Yuanfang Guan, Dr. Melanie Ohi, Dr. Heather Carlson, and Dr. Rudy J. Richardson. In particular, my thesis would have been reduced by a half if I were not able to work with Dr. Gilbert S. Omenn and Dr. Peter L. Freddolino. Through their co-mentorship, I have the privilege to work on protein function annotation and extend my algorithms to address biological problems in microbes and human.

As a member of the Yang Zhang lab, I have the honor of working with many talented colleagues. The development of deep learning based protein distance and contact prediction by Yang Li has laid the foundation of this thesis. The remote training by Dr. Jianyi Yang enabled me to master the webserver in my first semester at the University of Michigan. Dr. S. M. Mortuza polished many drafts of my manuscripts and grant proposals. Dr. Wei Zheng perfected the DeepMSA algorithm I initially developed and turned it into a generally useful program. Dr. Yang Cao contributed a substantial amount of geometry related code. Dr. Sha Gong introduced me to the field RNA structure modeling. Jonathan Poisson has been a tireless life saver for any hardware issue. Dr. Xiaogen Zhou, Dr. Xiaoqiang Huang, Robin Pearce, Dr. Peng Xiong, Syed

M. Rizvi, Eric W Bell, Dr. Jiong Li, Dr. Wallace Chan, Dr. Jarrett Johnson, Dr. Brandon Govindarajoo, Dr. Ambrish Roy, Fatima Z. Smali, Dr. Lijun Quan, Dr. Hongjie Wu, Dr. Wenyi Zhang, Biao Zhang, Zi Liu, Qidi Zhang, Dr. Xuantin Liu, Yiheng Zhu, Dr. Jaie Woordard and last, but by no means the least, Dr. Xiaoqiong Wei have either assisted in my projects or have me as a major co-author of their projects.

My graduate study would have been incomplete had I not met my life-long collaborator and partner, Dr. Xiaoqiong Wei. For the most part since our marriage, we were physically separated by the Pacific Ocean; yet the family letters and manuscript drafts we exchanged reminded us that love traverses time and space. My gratitude also goes to my parents and in-laws for enduring and supporting a pair of troublesome young couples.

This work was supported in part by the Extreme Science and Engineering Discovery Environment (XSEDE) (MCB160124, MCB160101, MCB200078 to C.Z. and Y.Z.); the National Institute of General Medical Sciences (GM083107, GM116960, GM136422 to Y.Z.); the National Institute of Allergy and Infectious Diseases (AI134678 to Y.Z.); the National Institute of Health Office of The Director (OD026825 to Y.Z.); the National Science Foundation (DBI1564756, DBI2030790, IIS1901191 to Y.Z., and MTM2025426 to P.L.F. and Y.Z.); the National Institute of Environmental Health Sciences (P30ES017885 to G.S.O.); the National Cancer Institute (U24CA210967 to G.S.O.), and the MCubed grant from the University of Michigan (U064195 to R.J.R.).

Table of Contents

List of Tables

List of Figures

Abstract

Predicting protein structure from its sequence (especially in the absence of structure templates) and deduction of biological function from structure remains a significant and unsolved problem. Much progress in *ab initio* (i.e. template-free) modeling of protein structure in recent years is due to the introduction of deep learning predicted inter-residue contacts and, even more recently, inter-residue distances.

We present D-QUARK, an *ab initio* protein folding algorithm guided by residue-residue distances and orientations predicted by deep learning. The D-QUARK pipeline is distinct from existing protein folding programs in the following aspects. Firstly, for a target sequence, it generates a high quality multiple sequence alignment (MSA) with deep and diverse sequence homolog alignment using the in-house DeepMSA algorithm. Secondly, to generate input features for deep learning prediction of distances and orientations from the MSA, raw coevolution features are extracted in the form of a covariance matrix and pseudo-likelihood maximization parameters, rather than traditional post-process coevolutionary features. Thirdly, the distance and orientation potentials are incorporated into a comprehensive replica-exchange Monte Carlo (REMC) simulation with a uniquely designed flat well potential for *ab initio* protein folding. The high quality MSA, accurate deep learning prediction, and REMC simulation with carefully designed energy terms all contribute to the high performance of D-QUARK. In terms of the first model TM-score, D-QUARK outperforms our previous *ab initio* protein folding algorithm by QUARK by 108.8% and two state-of-the-art distance-based structure prediction programs, DMPfold and trRosetta, by 22.9% and 11.4 %, respectively. In a post-CASP experiment, D-

QUARK achieves 8.1% higher first model TM-score on CASP13 FM target proteins than AlphaFold.

To annotate protein functions, including Gene Ontology (GO) terms, Enzyme Commission (EC) numbers, and ligand binding sites, from a predicted structure model, we developed COFACTOR. COFACTOR combines functional templates identified by structure alignment against the target structure model as well as sequence homologs and protein-protein interaction partners to derive consensus function annotations. COFACTOR was blindly tested in the community-wide CAFA3 function annotation challenge and was ranked among the top groups.

The structure and function prediction pipeline developed in this thesis was applied to proteome-wide annotation projects for several model organisms, including human and the JCVI-syn3.0 minimal bacterial genome, where our pipeline reveals previous uncharacterized proteins with important functions. Overall, we showed the impact of deep learning on protein structure and function prediction, and demonstrated its utility for reliable and scalable modeling.

Chapter 1 Introduction to Protein Structure and Function Prediction[1]

1.1 Introduction to Protein Structure Prediction

Proteins are the direct carrier of most biological functions necessary to sustain life. These diverse functions, ranging from enzymatic catalysis to biological pathway regulation and constitution of cellular structural component, are dictated by the unique 3D structures adopted by different protein molecules. Ever since the Anfinsen experiment in the 1970s showing that the protein tertiary structure is determined by its amino acid sequence[1], the protein sequence-structure-function paradigm has become one of the central theme in protein bioinformatics. Due to the extensive genome- and metagenome-wide sequencing efforts fueled by continuous development of new sequencing over the past decades, the number of protein sequences, most of which are translated from their nucleotide sequences, has exceeded 180 million as of UniProt database version 2020_04[2]. However, the rapid accumulation of sequences of proteins does not immediately translate into our deepen understanding of their biological functions, which are essentially encoded by their 3D structures.

The three most common approaches for protein structure experimental determination are X-ray crystallography[3], NMR spectroscopy[4], and cryo-electron microscopy[5]. Each of these methods is associated with significant human effort and expenses. Consequently, the growth in the number of solved protein structures is nowhere close to the explosion of protein sequences. As of September 16, 2020 for example, the Protein Data Bank[6] (PDB) database only host only

[1] The first part of this chapter is adapted from a review article under submission, entitled "Toward the solution of the protein structure prediction problem" by R Pearce, C Zhang, and Y Zhang.

0.17 million structures, many of which are redundant. This accounts for less than 0.1% of the total sequences in the UniProt[7]. This percentage was 0.7% in 2010 and 2% in 2004. Figure 1 illustrates the ever increasing gap between the number of known protein sequences and experimentally solved protein structures in the past decade.

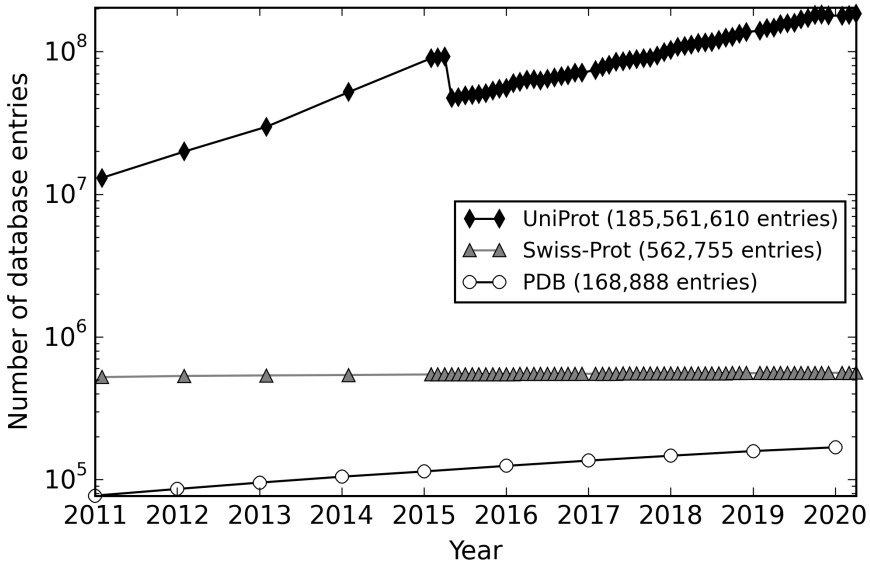


Figure 1. Number of protein sequences (in log-scale) in UniProt and Swiss-Prot (the subset of UniProt with manual function annotation), and the number of structures in PDB. The drop in the number of UniProt sequence in 2015 is due to the UniProt proteome redundancy reduction efforts in 2015 (https://www.uniprot.org/help/proteome_redundancy), where UniProt entries for the same protein from different strains of the same species are combined into a single UniProt entry.

Fortunately, thanks to the collective efforts from the bioinformatics community over the last few decades[8-24], an increasing portion of proteins in organisms are able to have their 3D structures reliably modeled by computational approaches[25-32]. Numerous high-quality structural models are being created every day by online structure prediction systems[23,33-45], some of which are developed by our lab[46,47]. These computational models are routinely used in various biomedical studies, such as structure-based protein function annotation[48-59], mutation analysis[60-66], ligand screening[67-72], and drug discovery[73-83]. Thus, the development of high-accuracy protein

structure prediction algorithms represents perhaps the most promising, yet challenging, approach to enclose the gap between the number of known protein sequences and determined structures.

Approaches for protein structure prediction, also referred to as protein folding, can be roughly classified into template-based modeling (TBM) and template-free modeling (FM). TBM methods refine initial models constructed by copying coordinates and spatial restraints from structurally determined proteins, called templates, identified from the PDB. The accuracy of TBM is therefore contingent on the identification of templates with similar topology to the target protein and the correct template-target alignment. The alignment is often dependent on the evolutionary distances between the query and templates. For proteins with sequence identities (seqID) >50% to the templates, for example, models produced by TBM can have up to 1 Å RMSD for the backbone atoms. For proteins with 30-50% seqID, the models often have ~85% of the core regions within an RMSD of 2-5 Å to the native structure. However, when seqID drops below 30% (the Twilight Zone[84]), modeling accuracy sharply decreases due to alignment errors and the lack of significant template hits[85-87]. Despite this drop off in accuracy, in theory, the protein structure prediction problem could be solved using TBM if algorithms were able to identify and correct align the best templates from the PDB library[88]. Nevertheless, this has yet to be achieved in practice[89].

Unlike TBM methods, FM methods, also called *ab initio* or *de novo* modeling, have been traditionally used to model proteins for which homologous templates cannot be identified from the PDB library. Note although the phrase "*ab initio*" usually refers to quantum chemistry calculation in cheminformatics, in protein structure prediction, any approach not dependent on full length template can be considered "*ab initio*". Since FM methods do not use global template information, they traditionally rely on physics- and/or knowledge-based energy functions and

extensive sampling procedures to construct protein structure models[14,90]. Due to the inherent inaccuracies associated with these procedures, FM has not yet achieved the same accuracy as TBM. However, recently the field has witnessed a remarkable achievement in that, for the first time, the performance gap between TBM and FM has been greatly narrowed by incorporation of deep learning predicted residue-residue contacts and distances in FM folding simulations. This approach resulted in the successful folding of 80% of "hard" proteins that lacked significant homologous templates in the PDB in the most recent Critical Assessment of Structure Prediction (CASP), a community-wide challenge of protein structure prediction techniques[91], compared to an average of just 26.7% in the previous CASP rounds[92-95].

The remaining sections of this chapter will review the history of TBM and FM approaches, with particular emphasis on the impact of contact/distance prediction on structure prediction. We will also introduce several representative state-of-the-art protein structure predictors.

1.1.1 A brief history of protein structure prediction before the introduction of contact and distance map prediction

TBM History

The first published work on protein structure prediction was by Browne et al., who built in 1969 a model for bovine α-lactalbumin by manual sequence alignment between the target protein and the experimentally determined chicken egg-white lysozyme[96]. The hypothesis underlying the study, which has since become an important idea in the majority of TBM methods, was that two proteins sharing high sequence similarity should be structurally similar. Although this early attempt implemented a rudimentary approach, it illustrated the four key steps

of TBM methods: (1) detection of experimentally solved protein structures (templates) related to the target protein to be modeled, (2) alignment of the target to the templates, (3) construction of the initial structure by copying the aligned regions, and (4) filling of unaligned regions and refinement of the overall structure.

The case discussed above for bovine α-lactalbumin is a specific case of TBM called homology modeling or comparative modeling, which typically can be used when the sequence identity between the template and protein of interest is ≥30%. This makes it significantly easier to identify high quality templates and produce reliable alignments using simple sequence-sequence alignment algorithms. Such algorithms include well-established methods developed in the 1970s and 1980s that utilize dynamic programming, such as the Needleman-Wunsch global alignment[97] and the Smith-Waterman local alignment[98] algorithms. Given the target-template alignment, a structure model can be constructed by copying coordinates from the aligned region and refining the overall structure. One of the most often used program for this purpose is MODELLER developed by Sali and Blundell[13]. MODELLER builds structure models by optimal satisfaction of template-derived spatial restraints together whether generic structural constraints such as ideal bond lengths and bond angles.

Since the accuracy of TBM sharply declines when the target-template seqID falls below 30%, a more sophisticated alignment approach called “threading”, which goes beyond simple pairwise sequence alignment, is needed. The concept of threading, or “fold recognition”, was first proposed by Bowie et al. in 1991[11]. In this work, the 3D structure of a template was represented by a 1D profile of local structural features. Since local structure features are more conserved than the sequence itself, they could be detected and aligned by distant homology target proteins with similar local structure but divergent sequences using dynamic programming.

The local structure features used by Bowie et al. were mainly computed from solvent accessibility and secondary structures. Later studies also include sequence profiles in addition to local structure features to further improve threading[99-101]. A sequence profile is typically built by PSI-BLAST[102] or HHblits[103], which searches the target protein sequence through a protein sequence database such as NR or UniRef to construct a multiple sequence alignment (MSA) of target protein and its sequence homologs; the sequence profile therefore records the composition of different types of amino acids at each position and is usually in the form of a position specific scoring matrix (PSSM) in PSI-BLAST or a hidden Markov Model (HMM) in HHblits[103].

As a side note, although "fold recognition" originally only refers to alignment-based threading methods, the concept has since been expanded to also include an unrelated branch of alignment-free template detection methods. These non-threading-based fold recognition methods typically performs machine learning to identify templates similar to the target without an explicit pairwise target-template alignment[104-106]. Therefore, they cannot be directly used to construct the 3D coordinates of the target protein, but are instead used for assignment of structure families to target proteins. Therefore, they will not be further discussed here.

The use of less reliable template alignment for threading of distant- and non-homology modeling targets necessitate the development of more effective template assembly and refinement methods that can tolerate alignment errors and large gaps. An early successful program for this purpose is TASSER[27]. Developed in the early 2000's by Zhang et al, TASSER extracts contiguous fragments from threading aligned regions of multiple threading templates to re-assemble them by structure assembly simulations. For computational efficiency, the unaligned regions are assembled using a lattice-based FM approach. In addition to template restraints, TASSER also includes several statistical energy terms such as hydrogen bonding and side-chain

interactions to guide its parallel hyperbolic Monte Carlo simulations. The simulation generates thousands of conformations, called "decoys", which are clustered by structure similarity to select the centroid of the largest cluster for additional refinement. A critical factor for the success of TASSER is its use of multiple templates rather than a single best scoring template. While rigorous theoretical studies to explain the consistent improvement resulting from combination of multiple structures was not available until many years later[107], its intuition can draw parallel from the famous "Anna Karenina principle": "All happy families are alike; each unhappy family is unhappy in its own way.". Indeed, all correct templates should be structurally similar; while each incorrect template is incorrect in its own way. Since there are many more ways for threading to go wrong than to get a correct answer, it is much more common to get a consensus correct alignment than multiple consistent but incorrect alignments[108]. More recently developed TBM approaches such as RosettaCM[109], Phyre2[110], and our I-TASSER[46,111-114] algorithm, also combine constraints from multiple templates. For example, I-TASSER extends TASSER[15,16,115] by performing an additional structure re-assembly simulation on cluster centroids using constraints from templates and cluster models combined with the inherent knowledge-based potential.

FM History

Unlike TBM, FM approaches fold protein without the use of global template. The earliest attempts at FM focus on refinement of experimental structures to improve their physical characteristics. For example, in 1989, Levitt et al. applied steepest descent to energy-minimize crystallography structures of myoglobin and lysozyme, using an empirical energy function for typical bond length, bond angle, dihedral angle, and van der Waals interactions together with

restraints from the initial experimental structures[116]. A similar energy function was used in 1977 by Karplus' group to study the motion of the bovine pancreatic trypsin inhibitor using molecular dynamics (MD) simulation[10]. Since then, various MD force fields and packages have been developed including AMBER[117-119], CHARMM[120-122], OPLS[123,124], and GROMOS96[125]. Despite their different parameterizations, most of these potentials share similar functional form to the original potential developed by Levitt et al. in 1969. Although empirical force field-based MD is useful for full atomic structure refinement, it is not yet able to consistently fold protein structure starting from sequence, apart from isolated cases of short and fast folding proteins. One reason for the limitation is that, since MD solves Newton's second law for all atoms in every simulation step to determine their motion, it is very computational expensive. This is perhaps best illustrated by the first successful MD-based protein structure prediction in 1998 by Duan and Kollman for a small peptide of only 36 amino acids, which took 2 months CPU time to achieve a resolution of 4.5 Å[119]. Sampling efficiency is not the only limitation of MD. For example, in 2012, DE Shaw's group applied Anton, the most powerful supercomputer for MD, and the latest CHARMM empirical force field to refine structure models of 25 CASP targets using ultra-long MD simulation (>100 μs for each molecule)[126]. The majority of their "refinement" runs simply cause the initial structure to drift further away from native structure at the absence of additional distance restraints. This shows that, despite many years of continuous development[127-133], the empirical force field alone may not be able to accurately describe the real energy landscape during protein folding. It is still not entirely clear whether this is merely a limitation caused by insufficient experimental and quantum chemical data for force field parameterization, or it is an inherent failure of the simple functional forms of empirical force field to capture high order interaction.

A more popular alternative to MD is fragment assembly. Proposed by Bowie and Eisenberg in 1994[134], fragment assembly can be considered a pseudo-*ab initio* method. Although global structure templates were not used, it nevertheless uses local structure fragments with fixed (9 residues) and variable lengths (15-25 residues). These fragments were identified from the PDB by sequence profile-based threading and assembled into full-length structural models. The use of fragments greatly reduced the conformational search space, while ensuring the local structures of the assembled fragments were well formed. Following this idea, Baker's group developed the Rosetta *ab initio* protocol in 1997[135], which has remained one of the widely used fragment assembly methods to this day. In Rosetta[136], 3 and 9 residues fragments are identified by gapless threading using the profile-profile and secondary structure matches. To perform fragment assembly in a simulated annealing Monte Carlo (SAMC) simulation, the backbone torsion angles of the predicted conformation are swapped for those of a selected fragment during each SAMC step. The Rosetta energy function includes terms for helix-strand packing, strand pairing, solvation, van der Waals interactions, radius of gyration, strand arrangement into sheets, and residue pair interactions. Conformations generated from SAMC are clustered to derive the final model. Apart from the Rosetta *ab initio* protocol, additional FM predictors, such as David Jones' FragFold[137] and our QUARK[90] algorithm, were developed based on a similar idea of fragment assembly using variants of Monte Carlo simulations, but with different approaches for fragment generation and energy function design. For example, QUARK includes a distance profile-based energy term, which constrains the distance between two residues based on the inter-residue distances of fragment pairs from the same template. QUARK also includes a more diverse set of 11 different conformation moves in addition to the fragment replacement move, making the conformational sampling procedure more efficient.

1.1.2 Contact and distance map prediction

The tertiary structures of proteins are stabilized by pairwise inter-residue interaction. Prediction of these interactions and the distances between the interacting atoms has become an important area of study in the protein structure prediction field. By the convention proposed by CASP, a residue pair is considered to form a contact if the distance between their C_β atoms (C_α for glycine) is <8 Å. Therefore, a simple representation of inter-residue interactions for a protein with length L is its contact map, which is a symmetric, binary $L \times L$ matrix, where each element of the matrix is a binary value that indicates if the residues form a contact or not.

Although deep learning-predicted contact maps transformed the field of protein folding only in recent years, the idea of contact prediction is not new. In the early 1990's, it was already proposed that contacts can be inferred from coevolution, i.e. correlated mutations, in MSAs[138,139]. The hypothesis was that the choice of amino acid types for a pair of interacting residues is usually under strong evolutionary pressure to maintain physical compatibility, resulting in mutations that are correlated. For example, if the first residue in the pair mutates from a positively-charge to a negatively-charged amino acid type, the second residue also need a negative-to-positive charge mutation to maintain electrostatic interaction. In practice, however, the accuracy of such early co-variation-based approaches was limited by the inability to distinguish between direct and indirect interactions. An indirect interaction occurs when position pairs *A* directly interacts with position *B,* and *B* directly interact with position *C.* Even if *A* does not directly contact *C*, co-evolution may still be observed between *A* and *C*. Further limitations were imposed by the limited size of the sequence databases and immature MSA construction methods at the time.

Improving contact prediction through global statistical models

Progress in contact prediction remained stagnant for two decades until a significant leap is brought about by global prediction approaches. Global statistical models, referred to as direct coupling analysis (DCA), were much more successfully in detangling direct from indirect interactions[140,141]. Unlike earlier "local" approaches such as mutual information[142], DCA is "global" because it determines the set of direct interactions that accounts for the observed sequence co-variation by simultaneously considering the entire set of pairwise interactions.

Many DCA methods fit a Markov random field (MRF), or more specifically a Potts model, to an MSA. MRF is a graphical model that represents each column of an MSA as a node that describes the distribution of amino acids at a given position, where the edges between nodes indicate the joint distributions of amino acids between each position pair. The couplings or co-evolutionary parameters can be determined from the edge weights. Since fitting an MRF model using its actual likelihood function is computationally intractable due to the need to calculate the partition function, various approximations have been developed including those based on message passing[140], Gaussian approximation[143], mean-field approximation[141], and pseudo-likelihood maximization[144-146]. Another popular method was introduced by PSICOV[147], which determines the coupling parameters by estimating the inverse covariance matrix or precision matrix under L1 regularization instead of directly fitting an MRF model to an MSA. This was later extended by our ResPRE[148] predictor, where the inverse covariance matrix is estimated using L2 regularization instead of L1 regularization, allowing for faster inference without apparent compromise in contact prediction accuracy.

Deep learning-based contact prediction

While the use of DCA represents one promising avenue to improve contact prediction, another breakthrough is made by deep convolutional neural network (CNN) that significantly improves DCA features. While CNN-based contact prediction is a new trend, the use of machine learning (ML) in general and neural network (NN) in particular in contact prediction dates back as far as simple co-variation-based techniques. Early ML methods utilized shallow (i.e. less than three hidden layers), fully connected NNs, whose inputs features are from coevolution, secondary structure, and sequence conservation[149,150]. These early machine learning-based predictors achieved comparable or slightly better accuracies than the contemporaneous methods based solely on coevolution. Following the first iteration of ML-based predictors, more complex NN architectures were developed[151-154]. Furthermore, contact predictors based on other ML techniques such as support vector machines (SVMs), including Jianlin Cheng's SVMcon[155] by group and our SVMSEQ[156] algorithm, or Random Forest models, including PconsC[157], also achieved similar performance as NN. Representative meta-methods in this era includes David Jones' MetaPSICOV[158] and our NeBcon[159], which combined the output of multiple DCA methods using shallow neural networks. Despite all of these advancements, the accuracy of contact prediction still remained unsatisfactory.

In the early 2010's, predictors began to incorporate deep learning into contact prediction. Early attempts included CMAPpro[160], which used a 2D recursive neural network, and DNCON[161], which used a deep belief network. Their simple increase of NN depth hardly substantiates improvement; in fact, neither of these two methods could outperform MetaPSICOV, a traditional shallow NN approach. An important reason for the disappointing performance of these early deep learning approaches was that contacts for a residue pair were

predicted using features extracted only from a small window of residues around the target residue pair. This sliding window approach ignores the global context of the residue pair, therefore not realizing the true potential of deep learning. The first deep learning predictor that consistently improves contact prediction over traditional NN was proposed in 2017 when Jinbo Xu's group proposed RaptorX-Contact[162]. RaptorX-Contact reformulated the contact prediction problem as an image segmentation (i.e. pixel labeling) problem in computer vision, where the whole contact map is consider an image and each residue pair is a pixel. The goal of this image segmentation is to label pixels that are in contact (represented as "1") or non-contact (represented as "0"). This pixel labeling problem is naturally suitable for deep convolutional neural network (CNN) in general, and residual neural network (ResNets[163]) in particular. While RaptorX-Contact uses similar coevolution, secondary structures, and sequence profile features as other predictors, the reformulation and introduction of ResNets with approximately 60 hidden layers enabled RaptorX-Contact to dramatically outperform other state-of-the-art methods. The demonstrated power of ResNets has inspired the vast majority of top ranked methods[164-166] developed since CASP12. Another successful method in CASP13 is our TripletRes[167] algorithm, which uses a similar ResNet basic block but with a very different design of features Instead of using the post-processed $L \times L$ evolutionary coupling information used by almost all other predictors at the time, TripletRes directly uses the $21 \times 21 \times L \times L$ raw coupling parameters as an input feature to its network, where 21 is the number of amino acid types (plus one type for gap).

Distance prediction

A natural extension of contact map prediction is distance map prediction. While similar in concept to the binary (contact or non-contact) contact map, a distance map provides more

detailed information on the distance between interacting residues. In practice, most distance map predictors do not predict the exact distance between residues, but the probability that the distance falls within one of the many distance bins. In other words, distance map prediction extends the binary classification problem of contact prediction into a multi-class classification problem (although recent attempts[168,169] have been made to directly predict the real-value distances). The idea of distance prediction is not new; as mentioned above, QUARK[170], for example, includes distance predictions derived from fragments detected from templates. Yet, the implementation of distance prediction in deep-learning frameworks is a recent advancement and makes the prediction much more robust and successful even in the absence of analogous structural templates. Three different CASP13 groups (RaptorX-Contact[171], DMPfold[172], and AlphaFold[173]), have extended the use of deep ResNets for contact prediction to distance prediction. The advantage of distance-based folding was most clearly demonstrated by AlphaFold in CASP13. Starting from the co-evolutionary features obtained from an MSA, AlphaFold trained an ultra-deep ResNet with 220 residual blocks to predict the distance map, which was then used to guide protein folding by either fragment-assembly SAMC simulation[173] or fragment-free gradient descent.

Orientation prediction

A further extension of distance prediction is orientation prediction. It has been known for years that knowledge-based energy functions that are dependent only on distances are often less accurate than those that use both distances and orientations for protein structure prediction[174-176]. The importance of orientation-dependent energy functions is twofold. Mathematically, it is impossible to uniquely determine the geometry of a structure without dihedral angle information,

as distance information alone cannot differentiate a pair of mirrored structures. Biologically, certain types of residue-residue interactions require not only distance proximity but also specific orientations between the residue pairs: for example, a pair of residues involved in beta strand pairing is required to be in an approximately parallel orientation. Recently, trRosetta[177] has implemented this idea by simultaneously predicting both pairwise residue distances and inter-residue orientations from co-evolutionary features using a unified deep ResNet, thereby outperforming AlphaFold in a post-CASP13 experiment.

Incorporating metagenomic sequences into prediction

As noted previously, another limitation of early contact prediction approaches was the small number of homologous sequences that could be used to construct MSAs for a target sequence. DCA methods in particular, and deep-learning approaches to a lesser extent, rely on collecting a sufficient number of sequence homologous in an MSA, as the more sequence homologs there are, the more reliable the co-evolutionary information is. Fortunately, the implementation of DCA and deep-learning contact/distance prediction has coincided with the expansion of sequence databases, in particular metagenomics sequence databases. Metagenomics is the application of next generation sequencing to sequence the DNA collected from environmental samples. These DNA sequences can be translated to protein sequences automatically, thereby producing large databases with billions of protein sequences. The utility of metagenomics sequences in contact-assisted structure prediction was first demonstrated by Baker for GREMLIN/Rosetta[178] by significantly enhancing the number of effective sequences in an MSA, thus producing "deep" MSAs with diverse sequences for DCA. Later MSA construction methods developed by our group and other teams[179,180] confirmed the usefulness of

metagenome-derived MSAs for improving contact prediction[167,179,180], threading results for distantly homologous targets[180,181], and the ability to model proteins that belong to families with unknown structures[178,182].

1.1.3 State-of-the-art protein structure prediction methods

This section reviews four representative structure prediction protocols from three research groups: C-I-TASSER/C-QUARK[183] from the Zhang group, RaptorX-DeepModeller[171,184] from the Xu group, and AlphaFold[173] from DeepMind. The particular selection of these three studies is based solely on their rankings in CASP13 (Appendix Figures A and B), and we by no means suggest their superiority over other methods not reviewed herein. Each of these state-of-the-art structure prediction algorithms utilize constraints taken from contact maps (C-I-TASSER/C-QUARK) or from distance maps (RaptorX-DeepModeller and AlphaFold).

C-I-TASSER/C-QUARK

C-I-TASSER and C-QUARK[183] are the latest iterations of the aforementioned I-TASSER and QUARK pipeline we developed. The main difference between C-I-TASSER/C-QUARK and I-TASSER/QUARK is the inclusion of constraints derived from contact maps predicted using deep-learning into both threading and structure assembly simulations. To predict the contact map, a deep MSA[180] is constructed for the query sequence by iteratively searching various sequence databases, including a large metagenome sequence database[185]. The deep MSA is used to extract co-evolutionary features as well as several other predicted local structural features, which are used as input by various deep-learning-based contact prediction algorithms[148,167] and

the results are combined to form a consensus contact map. In C-I-TASSER, the contact map is used by CEthreader[186], a contact-based threading program in LOMETS2[181] for template identification. Subsequently, the contact map is combined with the inherent I-TASSER/QUARK knowledge-based potentials and, in the case of C-I-TASSER, distance and contact constraints from structural templates, to assembly structural fragments into full-length structures by REMC simulations. Decoys from REMC simulations are clustered[187] and the cluster centroids are refined at the atomic level[188] to produce the final models. Despite using less informative contact map prediction from their deep-learning models, rather than distance map prediction as the other top performing groups, C-I-TASSER and C-QUARK were ranked as the top two automated server groups in CASP13, partly because the inherent and highly optimized template- and knowledge-based force fields and more sophisticated structure assembly simulations. This suggests the importance of using comprehensive conformational sampling simulations to optimally satisfy the constraints predicted by deep-learning. C-I-TASSER is also the first successful demonstration of consistent improvement of TBM using predicted contact maps, as the optimal balance between template and contact map constraints was previously considered to be particularly difficult to achieve for targets with homologous structure templates[189,190].

RaptorX-DeepModeller

RaptorX-DeepModeller[171,184] is another method that combines TBM (threading) and FM (sequence-derived contact/distance prediction) approaches into a single model and was the third ranked automated server in CASP13. Developed by the Xu lab, RaptorX-DeepModeller uses a similar deep ResNet architecture for distance prediction as the aforementioned RaptorX-Contact[171] predictor. Apart from the MSA-derived features used by RaptorX-Contact, RaptorX-

DeepModeller additionally includes query-template similarities and template distance maps calculated from the threading alignments produced by DeepThreader[191]. DeepThreader identifies templates by query-template similarity of sequence profile, secondary structure, and distance map to obtain alignments using ADMM. The predicted distances, together with the secondary structure and backbone torsion angles predicted by another one-dimensional ResNet, are fed into the Crystallography and NMR System (CNS) [192,193], a distance geometry-based protein folding program, to construct tertiary models. In CASP13, even though it was trained on a smaller set of proteins, RaptorX-DeepModeller slightly outperformed RaptorX-Contact, which uses non-template-based distance predictions with CNS. RaptorX-DeepModeller also consistently outperformed RaptorX-TBM, which inputs templates detected by DeepTheader into RosettaCM for comparative modeling. This shows a new approach to improve TBM by refining threading-derived constraints with co-evolutionary features and deep-learning.

AlphaFold

AlphaFold[173,194], which was developed by DeepMind during the latest CASP experiment (CASP13), is a collection of three FM approaches that combine three neural networks to predict the distances between residue pairs, to estimate the accuracy of a predicted protein structure (GDT-net), and to directly generate local structural fragments. Distances are predicted using the neural network described in the preceding section for this group. Apart from the distance predictions, candidate structures may alternatively be scored by directly predicting their accuracy in terms of their GDT_TS scores by GDT-net. The input to GDT-net is the MSA features similar to those used for distance map prediction, the predicted contact map calculated by collapsing the predicted distance map into a binary matrix, and features that describe the predicted structure.

The third network is a generative network trained end-to-end to create fragments by predicting the backbone torsion angles for each residue. Unlike typical fragment assembly approaches, which collect fragments from known structural databases, this approach allows for the *de novo* generation of fragments conditioned on the input features.

In CASP13, AlphaFold used three different folding strategies: SAMC guided by the GDT-net potential, SAMC guided by the distance map potential, and iterative gradient descent guided by the distance map potential. The final iteration of AlphaFold uses a potential that combines the log probability of the distance map predictions with Rosetta's score2 and a torsion angle potential. This potential is directly optimized with respect to the torsion angles using thousands of repeated gradient descent simulations.

While AlphaFold was developed with extensive engineering efforts and computational resources unattainable to most academic labs (e.g. the distance predictor in AlphaFold has 4.6 times more layers and was trained on 3.8 times more proteins than the contact predictors used by C-I-TASSER/C-QUARK), the scientific contribution of AlphaFold is also important. Firstly, it demonstrates the power of distance map compared to more conventional contact map in protein folding. Secondly, it demonstrates that deep-learning distance predictions can be constructed into an accurate and smooth energy landscape, on which conformations can be optimized by relatively simple gradient descent simulations. This idea has probably encouraged the development of several gradient descent based protein folding programs[195,196] after CASP13.

1.2 Introduction to Structure-based Protein Function Annotation

Even though protein function is usually dependent on the tertiary structure of proteins, inference of function from structure is not always straightforward. Firstly, since tertiary structure

is often more conserved than function, proteins with different functions can correspond to similar topology. For example, the TIM barrel fold is a notably promiscuous topology adopted by at least 60 different enzymes as well as several non-enzymes[197]. Secondly, unlike sequences, protein structures can be flexible; there are even proteins that are that performs their function through intrinsically disordered regions. Finally, as shown in Figure 1, at least two thirds of protein sequences with manually annotated protein functions do not have experimental structures. Such incompleteness of structure- function library hinders the training and template-usage of structure-based function annotations. To address these challenges, many algorithms are developed to annotate protein function annotation from structure and to complement structure-based function annotation with other non-structure-based methods.

Analogous to protein structure prediction, protocols for structure-based protein function annotation can also be roughly divided into template-based methods (TBM) and template-free methods (FM). TBM methods[198-200], including our COFACTOR[201] method reported in Chapter 3, draw functional insights from function templates structurally similar to target proteins. Structure similarities usually need to be measured by both global structure similarity and local structure matching to address the structure-function promiscuity issue mentioned above. On the other hand, FM methods[202-204] convert a protein structure to a graph or to a 3D density map and train graph CNN or 3D CNN models to directly output functions without explicit usage of templates.

Protein "function" can refer to a wide variety of biological vocabularies, ranging from protein-level functions such as Gene Ontology (GO) terms, Enzyme Commission (EC) numbers, and Human Phenotype Ontology (HPO), to residue-level functions such as ligand binding sites (LBS). Since LBS prediction is relatively well-addressed by previous works[198,199], this thesis mainly discussed GO terms, one of the most representative protein-level function annotation.

1.3 Questions Discussed by This Thesis

The remaining chapters of this thesis will address issues in structural bioinformatics: how to predict protein structure from sequence (Chapter 2 and 3) and how to predict protein function from predicted structure model (Chapter 4, 5 and 6).

Chapter 2 describes DeepMSA, which constructs for a given protein a deep and high quality MSA for three basic structure prediction tasks: contact prediction, threading and secondary structure prediction. Chapter 3 reports D-QUARK, a distance-based protein folding program that uses the MSA generated by DeepMSA and its variants as input to predict the inter-residue distances and orientations. The distance and orientation maps are then used to guide REMC simulation for *ab initio* protein folding.

Chapter 4 introduces COFACTOR, a meta-server for structure-based protein function prediction based on predicted structure models, sequence- and sequence-profile-based alignment, and protein-protein interaction networks. As indicated by its high number of citations, the COFACTOR pipeline has been applied to many small-scale function studies and genome-wide function annotations projects, including those for *E. coli* (https://epic.sites.uofmhosting.net/) and SARS-CoV-2[205].For the consideration of page limits, we will focus on two representative large-scale applications of COFACTOR to human (Chapter 5) and the JCVI-syn3.0 minimal bacterial genome (Chapter 6). Chapter 7 will conclude the thesis with proposals for future developments.

Chapter 2 DeepMSA: Deep Multiple Sequence Alignment Construction for Protein Structure Modeling[1]

2.1 Introduction

Multiple sequence alignment (MSA), also called "sequence profile", is designed to collect and align multiple homologous sequences of a query protein of interest. Since it contains rich information about the evolutionarily conserved positions and motifs, which cannot be derived from the query sequence alone, it has found fundamental usefulness in various bioinformatics studies. In protein structure prediction, for example, the MSA is the primary component to derive local secondary structure features [206,207], residue-residue contacts [162,166,208,209], and homologous structural templates [99,210]; these are of critical importance for the full-length 3D structure constructions [211,212]. In protein function annotations, the use of MSAs also has major impacts on the accuracy of Gene Ontology [201,213] and ligand-binding site [214,215] predictions.

Due to the critical importance of MSA, much attention has been paid to the development of various MSA and sequence profile construction methods. While PSI-BLAST is one of the most widely used approaches to query-specific sequence profile generation [102], HHblits [103] from the HH-suite [216] recently becomes popular for profile hidden Markov model (HMM) construction. Meanwhile, Jackhmmer and HMMsearch tools from the HMMER suite [217] are common alternatives for the applications. Both lines of programs have been heavily used,

[1] This chapter was adapted from a previously-published work: C Zhang, W Zheng, SM Mortuza, Y Li, and Y Zhang (2020) "DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins." *Bioinformatics*, 36(7), 2105–2112.

especially for the contact predictions that are recently found critical for template-free (or *ab initio*) protein structure prediction [17,218,219]. Most recently, a hybrid MSA generation approach combining HHblits and Jackhmmer searches is shown to improve contact prediction by MetaPSICOV2 [220]. There was also evidence showing that MSAs collected from metagenome protein sequences can increase the coverage of sequence homologies and be useful for contact-assisted de novo structure prediction [17,221].

Despite the importance of MSA construction, few standalone pipelines exist which can efficiently generate sensitive MSAs from a query input sequence, especially when multiple large sequence databases are involved. To address this urgent need, we developed and release DeepMSA, a new open-source program that constructs deep (in the sense of more sequences with a high diversity) and sensitive MSAs by merging sequences from three whole-genome and metagenome databases through a hybrid homology-detection approach. In this approach, HHblits from HH-suite 2.0.16 [216] and Jackhmmer/HMMsearch, which were modified from HMMER 3.1b2 [217] package to make the output format more compact in order to reduce file input/output (I/O), are used to perform sequence search, and the alignments are further refined by a custom HHblits database reconstruction step. Large-scale benchmark experiments have showed that, compared to the widely-used HHblits, PSI-BLAST and Jackhmmer programs, DeepMSA can consistently improve the accuracy of contact and secondary structure predictions, and threading programs, which is particularly important for distant-homology proteins.

2.2 Methods

2.2.1 Counting the number of effective sequences in MSAs

A common approach to quantify the homologous sequence coverage and/or alignment depth of an MSA is by counting the normalized number of effective sequence (Nf):

$$Nf = \frac{1}{\sqrt{L}} \sum_{n=1}^{N} \frac{1}{1 + \sum_{m=1,m\neq n}^{N} I\left[S_{m,n} \geq 0.8\right]} \qquad (2.1)$$

where L is the length of the query protein, N is the number of sequences in the MSA, $S_{m,n}$ is the sequence identity between the mth and nth sequences, and $I[\,]$ is an Iverson bracket, i.e. $I\left[S_{m,n} \geq 0.8\right]$ equals to 1 if $S_{m,n} \geq 0.8$, and to zero otherwise. While current literature lacks consensus in terms of the ideal Nf for contact prediction, we optimize the Nf cutoff as 128 to attain accurate contact prediction, as discussed later. The mathematical meaning of Nf is illustrated at Figure 2.

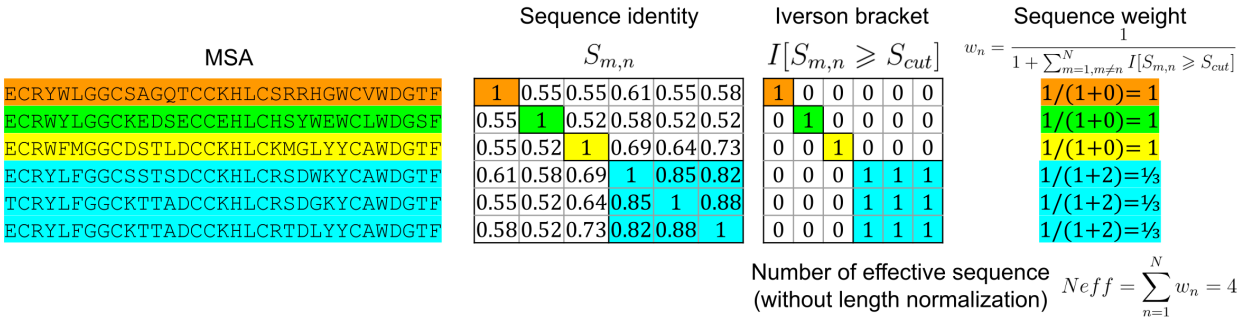


Figure 2. Graphic illustration for the calculation of sequence weights and the number of effective sequence. The MSA used in this example consists of $N = 6$ sequences with length $L = 33$. Using a sequence identity cutoff $S_{cut} =$ 0.8, the first three sequences forms three independent sequence clusters while the last three sequences form a single cluster. The four clusters are indicated by blocks colored in orange, green, yellow, and cyan in the sequence identity matrix. The Iverson bracket operation $I\left[S_{m,n} \geq S_{cut}\right]$determines whether the sequence pair m and n has sequence identity above sequence identity cutoff. In other words, this operation determines whether sequence m and n are neighbors within the same cluster. We can then assign a weight for each sequence, so that the w_n weight for sequence n is inverse proportional to its number of sequence neighbor:

$$w_n = \frac{1}{1 + \sum_{m=1,m\neq n}^{N} I\left[S_{m,n} \geq S_{cut}\right]} \qquad (2.2)$$

Any sequence n is always the sequence neighbor of itself, hence the addition of one in denominator of Equation (2.2). The number of effective sequence (without length normalization) is:

$$Neff = \sum_{n=1}^{N} w_n \qquad (2.3)$$

which equals to (1+1+1+1/3+1/3+1/3)=4 in this case. Therefore, the normalized number of effective sequence expressed in Equation (2.1) can be alternatively written as:

$$Nf = \frac{1}{\sqrt{L}} \cdot Neff \qquad (2.4)$$

which is $1/\sqrt{33} \times 4 = 0.70$ in this case. While our approach to calculate the number of effective sequences and sequence weights are also commonly used in other state-of-the-art contact prediction programs such as CCMpred, MetaPSICOV2 and TripletRes, there are also other programs (such as "plmc" module of the EVcoupling package) that calculates that the sequence weight and the number of effective sequence by first performing a sequence clustering. The weight of sequence n is $w_n = 1/k_n$ where k_n is the number of sequences in the sequence cluster to which sequence n belongs. This approach is essentially equivalent to our approach because Equation (2.3) quantifies the number of sequence clusters, except that our approach can save computation time to perform an explicit sequence clustering.

2.2.2 DeepMSA pipeline for MSA construction

The MSA construction process in DeepMSA can be divided into three stages, which correspond to the searching of three sequence databases (Uniclust30 [222], UniRef90 [223], and Metaclust [185]) through a combination of the HH-suite and HMMER programs (Figure 3).

In Stage 1 (Figure 3 first column), HHblits from HH-suite 2.0.16 is used to search UniClust30 with the parameters "-diff inf -id 99 -cov 50 -n 3". After testing HHblits MSA generated using the last version of UniProt20 (2016_02), latest Uniboost30 (2016_09), and three recent versions of Uniclust30 (2017_04, 2017_07, 2017_10), we found the three versions of Uniclust30 generate MSAs with comparable quality, all with a higher contact prediction accuracy than MSA generated by either UniProt20 or Uniboost30. Therefore, an arbitrary UniClust30 version (2017_10) is used for this study.

If Stage 1 does not generate enough sequences, i.e. *Nf*<128, Stage 2 will be performed (Figure 3 second column), where Jackhmmer is used to search against UniRef90 with parameters "-N 3 -E 10 --incE 1e-3". We choose "-E 10" because lowering this e-value cutoff occasionally results in the inclusion of excessive number of non-homologous multi-domain hits in edge cases, although the final number of significant hits in the Jackhmmer alignment is determined by "--incE". Instead of directly using the alignment generated by Jackhmmer search, esl-sfetch from the HMMER package is used to extract full length sequences according to the list of Jackhmmer

hits. These full-length sequences are converted into a custom HHblits format database by "hhblitdb.pl" script from HH-suite. After the construction of the custom database, HHblits is again applied to search this custom database using the same search parameter as in Stage 1 but jump-starting the search from the Stage 1 sequence MSA. If the MSA from Stage 2 has an *Nf* higher than that from Stage 1 MSA, it will replace the Stage 1 MSA for subsequent computation.

DeepMSA implements two time-saving heuristics to reduce time complexity associated with construction of HHblits format database, which, unlike conventional sequence databases, comprise of sequence profiles. Each profile can be either one sequence or one MSA within a family of protein sequences clustered by sequence identity. The time required to construct a profile database is proportional to the number of profiles and the average number of positions of the profiles. It may take many hours to construct a custom HHblits database if the sequences are very long or if there are too many sequences. To shorten the time for database construction, we trim the Jackhmmer hits and perform sequence clustering. In particular, instead of using the full-length Jackhmmer hit, we trim the Jackhmmer hit to extract the local region aligned to the query in the Jackhmmer alignment, as well the *L* flanking residues at both sides of the aligned regions. Moreover, all trimmed hits from the previous step are further clustered by kClust [224] into sequence clusters by 30% sequence identity cutoff. Next, Clustal Omega [225] is then used to align sequences within each cluster into aligned sequence profiles. These profiles are fed into hhblitsdb.pl to construct the custom HHblits database. As kClust and Clustal Omega usually take only a few minutes, and the number sequences is approximately ten times larger than the number of kClust sequence clusters, it will take less than half an hour to construct the custom database.

If the MSA from previous stages still has $Nf<128$, Stage 3 is performed (Figure 3 third column), where the MSA from the previous stage is converted into a hidden Markov model

(HMM) by HMMbuild from the HMMER package. This HMM is searched against Metaclust metagenome sequence database by HMMsearch, using parameters "-E 10 --incE 1e-3". Similar to Stage 2, sequence hits from HMMsearch are built into a custom HHblits database. The MSA from previous stages is used to jump-start an HHblits search against this new custom HHblits database to derive the final Stage 3 MSA.

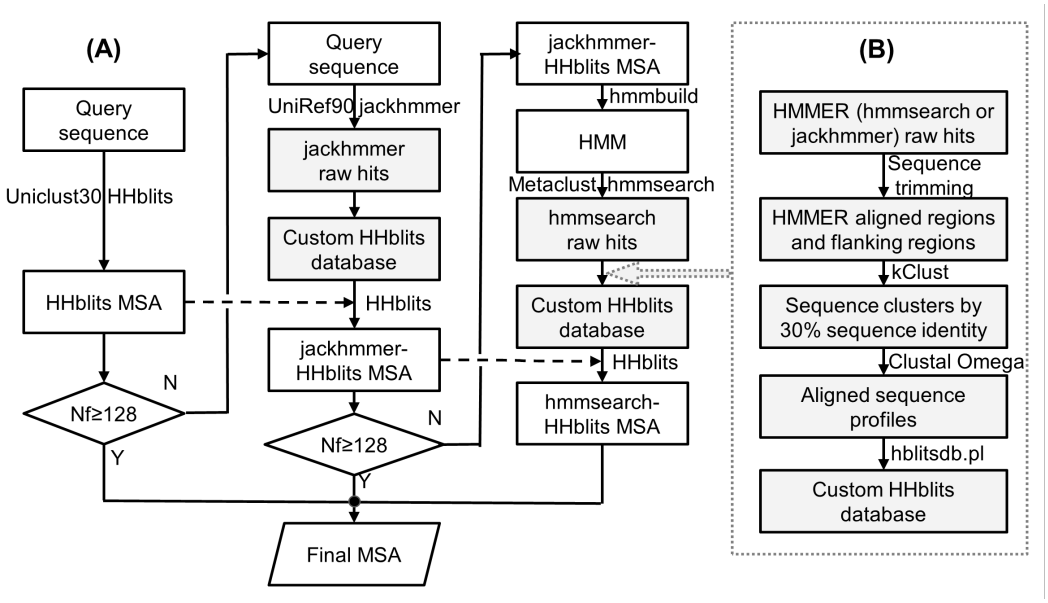


Figure 3. (A) Flowchart of DeepMSA. Three stages of MSA generations are performed consecutively using sequences from HHblits search through Uniclust30 (first column), Jackhmmer through UniRef (second column), and HMMsearch through Metaclust (third column). (B) Details of constructing custom HHblits database from Jackhmmer/HMMsearch hits.

2.3 Results

2.3.1 Dataset

DeepMSA is tested on a set of 614 non-redundant proteins curated from the SCOPe database [226] according to the following criteria: (i) any target coming from a fold with only one superfamily is excluded, because such a target is unlikely to have any remote structure analog; (ii) redundant sequences with a 30% pair-wise sequence identity are removed; (iii) each query should have at least one template structure, detectable by TM-align [227], from the PDB which has a TM-score >0.5 with the sequence identity <0.3 to the query. These resulted in 614 proteins,

which are classified into 403 “Easy” and 211 “Hard” targets by the meta-threading program, LOMETS [228], based on the significance of threading alignments between query and template sequences.

2.3.2 Coverage and depth of MSAs by DeepMSA

Since one of the initial motivations for DeepMSA to combine sequences from different sequence databases is to collect more diverse sequences, it is instrumental to examine the coverage and depth of the MSA brought by DeepMSA. To this end, Table 1 lists the depth results of MSAs generated by six different schemes, including DeepMSA, its three stages, and three baseline methods. Here, to obtain data for different stages, we force DeepMSA to perform all three stages regardless of Nf cutoff. Nevertheless, the final MSA in DeepMSA is calculated as the normal procedure, i.e., having the MSA constructed from Stage 1 if its *Nf* is $\geqslant$128; or from Stage 2 if Stage 1 has *Nf*<128 but Stage 2 has *Nf*$\geqslant$128; or from Stage 3, otherwise. Two of the baseline methods generate MSAs by Jackhmmer or PSI-BLAST search against the same UniRef90 database as used by DeepMSA. For the last baseline method, denoted as “No custom db” in Table 1, the custom HHblits database construction and HHblits search in Stage 2 and 3 are replaced by direct concatenation of HMMER (Jackhmmer and HMMsearch) MSAs to the MSA from the previous stage, similar to the approach reported earlier [17].

Table 1. *Nf* and the number of aligned homologous sequences (*N*) in the MSAs collected by different schemes.

Schemes†	“Hard” targets		“Easy” targets		All targets	
	Nf	*N*	*Nf*	*N*	*Nf*	*N*
DeepMSA	119.67	3046.16	435.52	8869.82	331.20	6868.53
Stage 1	82.22	1698.12	430.49	8765.65	310.81	6336.91
Stage 2	131.30	3158.46	612.83	14816.79	447.35	10810.43
Stage 3	346.02	8098.61	1031.95	24194.26	796.23	18663.02
Jackhmmer	174.64	3720.27	727.95	17818.32	537.81	12973.55
PSI-BLAST	145.02	5032.81	739.06	21195.11	534.92	15640.96
No custom db	516.27	11751.12	1642.74	49326.13	1255.63	36413.55

†Stage 1, 2 and 3 are three stages of DeepMSA. "No custom db" modifies DeepMSA pipeline by directly concatenating HMMER alignments without custom HHblits database construction in Stage 2 and 3. "PSI-BLAST" and "Jackhmmer" search UniRef90 with PSI-BLAST and Jackhmmer, respectively.

As expected, the alignment depth, when measured by Nf and the total number of detected sequences, gradually increases from Stage 1 to Stage 3. The increase is particularly large for "Hard" targets, where the final MSAs from DeepMSA are on average 1.5 and 1.8 times deeper than Stage 1 in terms of *Nf* and number of sequences, respectively. On the other hand, the alignment depth of DeepMSA is significantly smaller than "No custom db" and "Jackhmmer". This is because all HMMER hits are included in the "No custom db" and "Jackhmmer" alignments, while many HMMER hits are discarded by DeepMSA during HHblits search through custom databases.

The full-length MSA constructions often cost more memory and slow down the computing processes. Moreover, due to the composite profile construction and alignment algorithms, MSAs with greater *Nf* and sequence numbers do not necessarily indicate better MSA quality, as shown in later sections. In fact, there is no single index which can directly assess the performance of MSA collection programs. To more objectively assess the quality of MSA builders, below we apply these MSAs to three protein structure modeling experiments, i.e. residue contact prediction, secondary structure prediction, and protein fold-recognition (i.e., threading).

2.3.3 DeepMSA increases contact prediction accuracy

The utility of DeepMSA for contact prediction is assessed using six state-of-the-art programs: CCMpred [229], MetaPSICOV2 [220], DeepContact [165], DeepCov [230], PConsC4 [231], and TripletRes [232]. Here, CCMpred is a representative coevolution-only contact predictor. MetaPSICOV2 is based on traditional (shallow and fully-connected) neural networks. The rest of

the programs are based on deep convolutional neural networks. While other predictors with good performance also exist, we selected the six programs partly because of the availability of standalone packages, which facilitate the large-scale implement and comparison of the results.

In Table 2, we list the results of contact predictions by the six predictors, each having the MSA collected from the six schemes listed in Table 1. Since MetaPSICOV2 and DeepContact have their own built-in MSA generation protocols, both of which combine HHblits and jackhammer, contact precisions from the built-in MSAs are listed as "default" in Table 2. Here, as in community-wide Critical Assessment of protein Structure Prediction (CASP) challenges [218], a contact is defined as Cβ atoms (Cα atoms for Glycine) from a pair of residues, i and j, being close to each other by less than 8 Å. Contact prediction accuracies of different methods are evaluated by precisions of top L, $L/2$, and $L/5$ medium-range ($12 \leq |i-j| \leq 23$) and long-range ($24 \leq |i-j|$) predicted contacts. In accordance with CASP convention, Table 2 only lists the long-range contacts of "Hard" targets, where. For completeness, the results for medium-range contacts for all targets ("Hard" and "Easy") are listed in a spreadsheet file at https://zhanglab.ccmb.med.umich.edu/DeepMSA/assessment.xlsx.

Table 2. Long-range contact prediction precision for 211 "Hard" protein targets. Bold font indicates the highest value in each category.

Predictor	MSA	Top *L*	*P*-value	Top *L*/2	*P*-value	Top *L*/5	*P*-value
CCMpred	DeepMSA	0.268	*	0.375	*	0.483	*
	Stage 1	0.215	3.73E-24	0.307	4.78E-23	0.410	4.21E-15
	Stage 2	0.237	2.49E-13	0.333	1.19E-14	0.430	3.45E-13
	Stage 3	**0.280**	1.00	**0.381**	0.98	**0.486**	0.79
	Jackhmmer	0.227	3.84E-15	0.317	2.37E-15	0.418	1.54E-11
	PSI-BLAST	0.208	3.35E-24	0.289	2.18E-26	0.394	5.81E-16
	No custom db	0.264	0.187	0.366	4.83E-2	0.468	1.86E-2
MetaPSICOV2	DeepMSA	0.410	*	0.532	*	**0.654**	*
	Stage 1	0.373	6.66E-13	0.483	1.32E-12	0.595	1.19E-10
	Stage 2	0.388	1.43E-6	0.501	2.25E-7	0.618	6.56E-6
	Stage 3	**0.412**	0.93	**0.534**	0.74	0.653	0.67
	Default	0.387	4.75E-5	0.500	1.79E-5	0.612	2.11E-5
	Jackhmmer	0.377	2.27E-7	0.490	1.24E-6	0.604	1.07E-5
	PSI-BLAST	0.336	1.46E-19	0.441	6.32E-16	0.546	4.42E-13
	No custom db	0.400	3.29E-2	0.515	1.43E-2	0.629	7/03E-3
DeepContact	DeepMSA	0.485	*	0.630	*	**0.756**	*
	Stage 1	0.445	3.43E-15	0.581	4.00E-13	0.716	3.60E-7
	Stage 2	0.458	5.07E-10	0.598	3.15E-8	0.730	7.63E-5
	Stage 3	**0.488**	0.99	**0.632**	0.92	0.754	0.13
	Default	0.434	1.37E-13	0.562	1.75E-13	0.681	5.35E-10
	Jackhmmer	0.441	1.07E-11	0.576	7.55E-10	0.702	2.76E-6
	PSI-BLAST	0.427	1.99E-16	0.553	6.42E-15	0.681	2.77E-9
	No custom db	0.472	1.84E-3	0.614	1.88E-3	0.732	5.17E-3
DeepCov	DeepMSA	**0.439**	*	**0.588**	*	**0.738**	*
	Stage 1	0.408	6.01E-9	0.553	6.85E-7	0.701	3.36E-5
	Stage 2	0.420	1.03E-5	0.561	3.51E-6	0.712	5.46E-5
	Stage 3	0.439	0.49	0.586	0.35	0.730	9.68E-3
	Jackhmmer	0.392	1.21E-11	0.521	4.80E-11	0.662	2.28E-9
	PSI-BLAST	0.377	2.96E-18	0.505	7.01E-17	0.649	5.16E-12
	No custom db	0.421	7.09E-4	0.563	1.61E-3	0.708	2.21E-3
PConsC4	DeepMSA	0.475	*	0.610	*	0.718	*
	Stage 1	0.420	6.64E-17	0.544	4.10E-13	0.653	1.04E-7
	Stage 2	0.443	1.19E-8	0.572	5.52E-7	0.681	3.24E-4
	Stage 3	**0.478**	0.97	**0.612**	0.75	**0.719**	0.70
	Jackhmmer	0.420	2.08E-11	0.545	1.61E-8	0.652	3.69E-6
	PSI-BLAST	0.364	8.64E-16	0.474	2.89E-14	0.572	4.55E-12
	No custom db	0.462	1.09E-2	0.593	2.38E-2	0.697	3.72E-2
TripletRes	DeepMSA	**0.610**	*	**0.759**	*	**0.860**	*
	Stage 1	0.594	6.37E-6	0.742	5.78E-4	0.849	2.59E-2
	Stage 2	0.601	2.65E-4	0.747	6.65E-4	0.856	0.17
	Stage 3	**0.610**	0.34	0.756	8.34E-2	0.859	0.29
	Jackhmmer	0.565	3.11E-8	0.704	1.00E-7	0.815	9.40E-5
	PSI-BLAST	0.547	1.35E-13	0.684	2.15E-13	0.790	2.50E-9
	No custom db	0.584	1.83E-5	0.728	8.00E-5	0.830	7.85E-4

* Each p-value is calculated by one-tailed paired t-test to test whether DeepMSA has significant better contact prediction accuracy than the respective MSA.

It is shown that the MSA from DeepMSA outperforms the default MSA for contact prediction in all six contact predictors. For instance, the precisions for the top L contacts

generated by TripletRes and CCMpred increased by 2.7% and 24.4%, respectively, when they use the MSA from DeepMSA instead of the default MSA. Furthermore, contact precision improves progressively from Stage 1 to Stage 3 for all the programs, indicating the effectiveness of depth of MSAs in contact prediction. Contact precisions from DeepMSA are also consistently higher than those from HHblits (i.e. Stage 1), Jackhmmer, and PSI-BLAST alone.

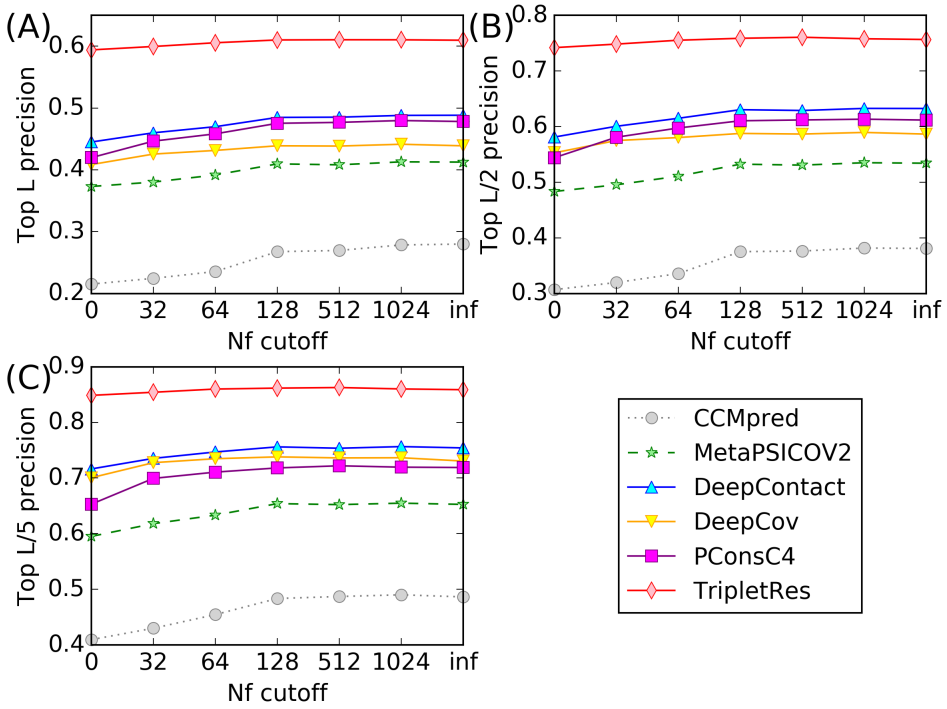


Figure 4. Nf cutoff of DeepMSA versus top *L* (A), top *L*/2 (B) and top *L*/5 (C) long range contact prediction precision. The *Nf* cutoff of "0" and "inf" correspond to always using Stage 3 and Stage 1 MSAs, respectively.

The output MSA of DeepMSA is not always created from Stage 3 if previous two stages achieve $Nf \geq 128$, which helps to save the memory and running time of DeepMSA. Interestingly, this setting does not degrade contact precision significantly for most predictors. In fact, for TripletRes and DeepCov, the MSA from DeepMSA yields slightly better contact precision compared to the MSA from DeepMSA Stage 3. Figure 4 shows the effect of *Nf* cutoff in DeepMSA on the precision of contact prediction, where, for all but one program (CCMpred), increasing the *Nf* cutoff over 128 hardly improves contact precisions. In other words, when the

alignment is already deep ($Nf \geq 128$), further inclusion of more sequences is indeed not beneficial for all five neural network-based contact predictors. This might be because deeper MSAs are more prone to contain alignment errors and false positive hits, where the cutoff of $Nf=128$ might be the result of the tradeoff between the sequence coverage and alignment noises. Moreover, this result may also suggest that the sequence datasets from the standard Uniclust30 utilized in Stage 1 is more reliable than the UniRef90 and metagenomic database, and thus the addition of more sequences from the latter datasets might have the tendency to introduce more noises.

The high quality of MSA from DeepMSA is not merely the result of combining multiple sequence databases. In particular, apart from the lack of custom HHblits database construction and search step, "No custom db" uses identical sequence databases, with the same HHblits and HMMER programs as DeepMSA. Despite far greater alignment depth as shown in Table 1, "No custom db" is worse than DeepMSA by 1.0% (CCMpred) to 4.2% (TripletRes) in terms of top L contact precision. These data suggest again that deeper alignments (with more sequence homologs) do not necessarily guarantee better contact prediction. It also indicates that although diverse sequence databases are contributive to DeepMSA performance, it is also essential to combine multiple sequence search and alignment algorithms, especially the custom HHblits database construction subroutines in our case.

DeepMSA also outperforms the default MSAs in DeepContact and MetaPSICOV. In particular, the Stage 2 MSA yields slightly more precise (0.3%) top L contact prediction by MetaPSICOV than its default MSA, even though both kinds of MSAs come from HHblits search through custom HHblits database constructed from Jackhmmer hits. This show that our time-saving heuristics (HMMER hit trimming and kClust clustering, which result in an overall

average DeepMSA running time of 0.7 hour per protein, Figure 5) introduce little compromise to final alignment quality.

Apart from benchmark data discussed herein, DeepMSA was also blindly tested in CASP13 as the MSA generation pipeline for our TripletRes server [232], whose average top L contact precisions on all 31 FM targets increased from 0.332 with HHblits MSAs to 0.409 with DeepMSA.

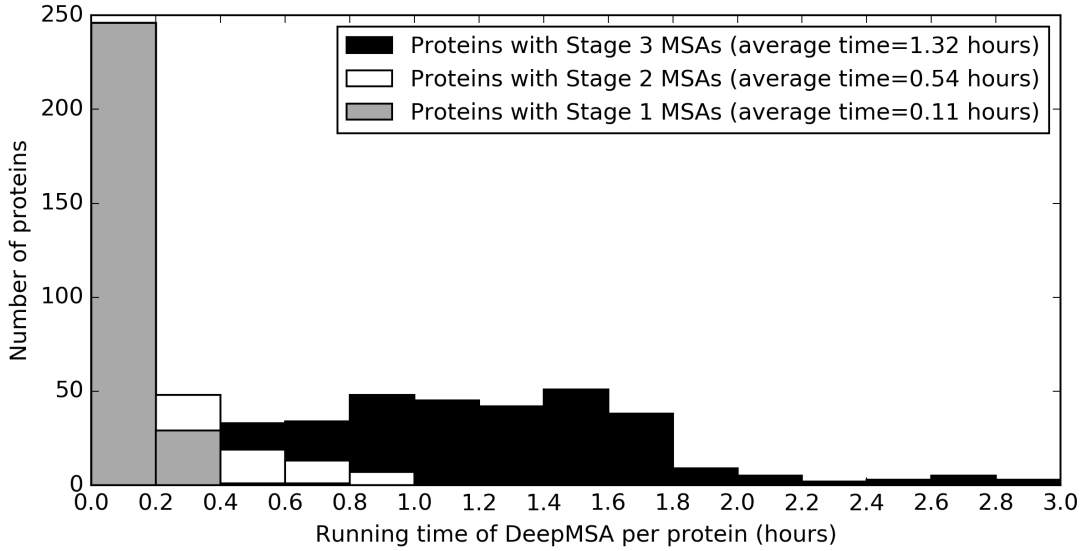


Figure 5. Stacked histogram for per protein running time of DeepMSA, with an average running time of 0.70 hour. DeepMSA does not always run all three stages to generate the final MSA. Grey, white, and black regions correspond to proteins with only Stage 1 MSA, with both Stage 1 and Stage 2 MSAs, and with Stage 1 to 3 MSAs, respectively. The running time for each protein is measured using a single thread using Intel Xeon CPU E5-2680 v2 at 2.80GHz.

2.3.4 DeepMSA enables more accurate threading

Threading is an important approach to template-based protein structure prediction, which recognizes proteins with similar fold to the query proteins. Since most of the state-of-the-art methods use profiles, in the form of either HMM or Position Specific Scoring Matrix (PSSM), to deduce query-template alignments, we examine whether and how DeepMSA can impact the performance of two typical threading programs, HHsearch [210] and MUSTER [233], which by default use HHblits and PSI-BLAST to construct sequence profile, respectively.

The HHsearch and MUSTER template database is constructed from the 71,684 non-redundant (pairwise sequence identity<70%) protein structures from the I-TASSER [112] template library at https://zhanglab.ccmb.med.umich.edu/library/. To generate the HHsearch library with default profile and with our new profiles, we first build MSAs for all templates by HHblits search against Uniclust30 database and DeepMSA, respectively. The hhmake program from HH-suite is then used to convert the MSAs to HHsearch style HMM library.

In MUSTER, the default sequence profiles are constructed by searching NR database with blastpgp, i.e. the legacy PSI-BLAST program [102]. Checkpoint files from PSI-BLAST search is then converted to MTX format sequence profiles. Conversion of DeepMSA alignments to MTX format is implemented by the "a3m2mtx.pl" script in the DeepMSA package. This script jump-starts a PSI-BLAST search using the MSA of DeepMSA against a dummy BLAST format database. The MTX file can then be recovered from the checkpoint file of the jump-start search. Similarly, for query proteins, we also construct both DeepMSA profiles and default profiles.

In Table 3, we list a comparison of template alignments obtained by HHsearch and MUSTER using different MSAs. The results are presented only for "Hard" targets in terms of the average TM-score [234], alignment coverage (number of aligned residues divided by query length), and RMSD of aligned regions, where all templates with a sequence identity >30% to the query have been excluded. It is shown that, for "Hard" threading targets, the TM-score of first template by MUSTER and HHsearch is increased by 10.9% and 7.5%, respectively, if the DeepMSA profiles instead of the default PSI-BLAST/HHblits profiles are used. Of note, the number of "Hard" targets with correctly identified templates (TM-score>0.5) is increased by 64.0% and 39.4% for MUSTER and HHsearch, respectively.

Table 3. Benchmark results for the first threading template on 211 "Hard" targets. Bold font indicates the highest value in each category.

Method	TM-score	*P*-value	RMSD (Å)	Coverage	#(TM-score>0.5)
HHsearch	0.308	5.70E-03	11.15	0.665	33
HHsearch$^{(D)}$	**0.331**	*	11.17	0.697	**46**
MUSTER	0.311	7.40E-04	13.62	0.872	25
MUSTER$^{(D)}$	**0.345**	*	12.87	0.851	**41**

(D) indicates threading with DeepMSA profile.

The observation that DeepMSA significantly boosts threading performance for "Hard" targets can be partially explained by improved quality of query-template alignments. To examine this point, we curate a subset of 143 "enriched" "Hard" targets, each of them having at least 30 templates of the correct fold (TM-score >0.5) detectable by TM-align with <30% sequence identity to the query. For each of these targets, we calculate average TM-score with all the templates aligned by HHsearch using DeepMSA sequence profile and compare it to that using the default HHblits profile used by HHsearch. Figure 6A lists the average TM-score difference on the top 30 templates for each of 143 targets. The data show that DeepMSA generated positive impact on the query-template alignments for 68.5% (=98/143) of the cases. Among the 98 cases, 69 (70.4%) have the TM-score difference with p-value <0.05 in the paired t-test (dark bars in Figure 6A), showing that the difference is statistically significant although only about 30 data points are involved in the paired t-test calculation for each target.

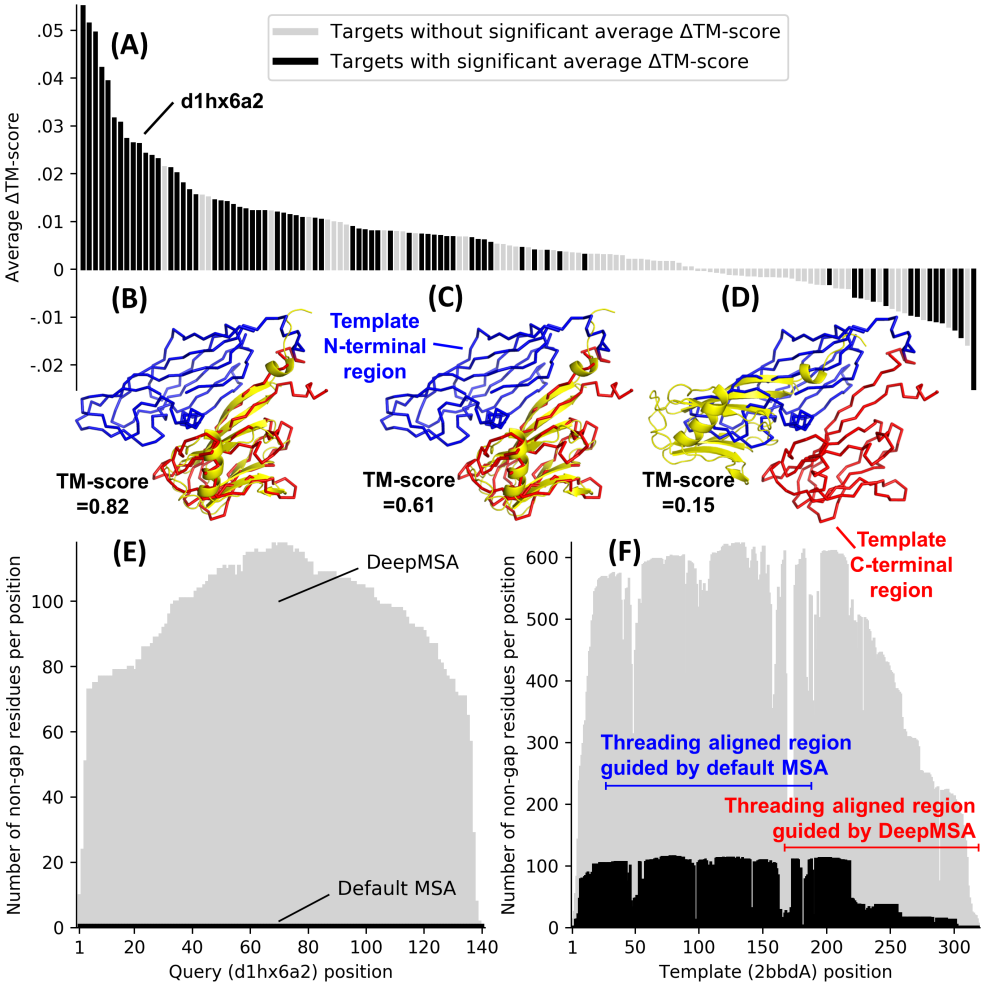


Figure 6. Contribution of DeepMSA to query-template alignment in HHsearch threading. (A) The TM-score of HHsearch guided by DeepMSA profile minus that by the default profile (ΔTM-score) is calculated for each template of a "Hard" target. The *y*-axis is the average ΔTM-score for each target, ranked in descending order (*x*-axis) of average ΔTM-score. The statistical significance of ΔTM-score for each target is calculated by a paired t-test between TM-score pairs (i.e. TM-score by DeepMSA versus TM-score by default profile) for all templates of the target. Targets with significant ΔTM-score are colored in black. (B, C, D) Alignment of query d1hx6a2 (cartoon) to template 2bbdA (upper left and lower right ribbons for N- and C- terminal regions, respectively) using TM-align (B), HHsearch alignment guided by DeepMSA profile (C) and that guided by the default profile (D). TM-score by DeepMSA profile guided HHsearch is lower than that by TM-align due to alignment shift. (E, F) Number of non-gap residues (y-axis) at each position (x-axis) in the DeepMSA profile (grey) and in the default HHblits profile (black) for query (d1hx6a2) (E) and template (2bbdA) (F).

To further illustrate the importance of DeepMSA profile in threading, we show a case study on query d1hx6a2 and its template 2bbdA. HHsearch threading based on DeepMSA profile correctly aligns query to C-terminal (residue 167 to 319) of template and achieves a TM-score =0.61 (Figure 6C); the alignment region is similar to that by the structure alignment from TM-align, although TM-align has an even higher TM-score (=0.82, Figure 6B). On the other hand,

HHsearch threading with the default HHblits profile only gets a TM-score=0.15 due to complete mis-alignment of query to the N-terminal (residue 27 to 188) of template (Figure 6D). Such differences can be explained by depths of MSAs for both query and template: the default HHblits run only detects 133 homologs for the template and no homolog for the query. On the other hand, DeepMSA profile is much deeper, with 624 and 118 homologs for the query (Figure 6E) and the template (Figure 6F), respectively. The lack of template homologs in the default run is particularly severe at the C terminal of the template, driving HHsearch to align the query to the template N terminal instead.

In addition to the creation of correct alignments, another reason for the performance improvement by DeepMSA on threading is that better MSA profiles can help improve the ranking of the template alignments. In Figure 7, we show an example from the query protein (d1yvua1) which is aligned on the template 3f73A2 using HHsearch. Although both default and DeepMSA profiles resulted in reasonable query-template alignments with a TM-score >0.5, their alignment scores are very different. While the HMM probability on the DeepMSA profile is 77.5% which puts the template as ranked the first, the probability score is 0.2% using the default profile which is ranked at 19,825th position among all templates. Thus, although the default profile can generate correct alignment on this query-template pair, the correct template cannot be selected by the threading program due to the poor alignment scores. In this case, an unrelated protein (3iz6D3, TM-score=0.08) was selected as the first template when using the default HMM profile alignments.

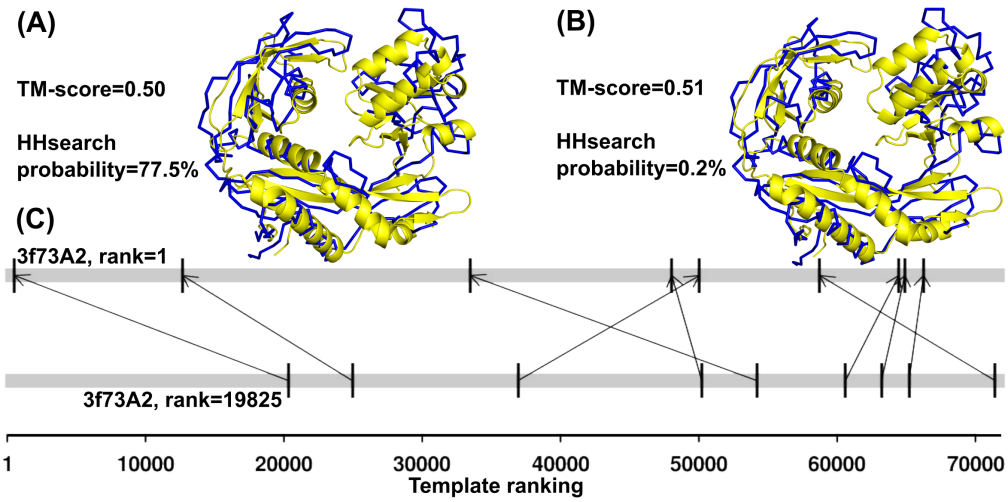


Figure 7. Contribution of DeepMSA to HHsearch template ranking for query d1yvua1. (A-B) Threading alignment between query (cartoon) and template 3f73A2 (ribbon), guided by DeepMSA profile (A) and by default HHblits profile (B). (C) Ranking of nine correct templates (TM-score>0.5, black vertical lines) among all 70,977 templates (grey horizontal bands) after excluding template proteins with a sequence identity >30% to the query. Template rankings guided by DeepMSA profile and that by the default profile are shown in upper and lower bands, respectively. The same template in the two cases is connected by a thin arrow.

2.3.5 DeepMSA profiles improve secondary structure prediction over traditional PSI-BLAST profiles

We further test the performance of DeepMSA in secondary structure (SS) prediction by PSIPRED 4.0 [206] and PSSpred [235]. By default, PSIPRED and PSSpred construct MTX format sequence profiles by searching UniRef90 or NR database with PSI-BLAST program [102]. MTX format DeepMSA profile for these two programs can also be obtained by a3m2mtx.pl.

The accuracy of the SS predictions by PSSpred and PSIPRED is evaluated by Q3 accuracy and SOV segment overlap measure [236] (Table 4). Compared to the default profiles, sequence profiles from DeepMSA improve the Q3 accuracy by 1.2% and 1.0% for PSSpred and PSIPRED, respectively. Similarly, SOV scores by PSSpred and PSIPRED are improved by 1.8% and 1.5%, respectively, when MSAs from DeepMSA are used. The differences are statistically significant, since the p-values in Student's t-test are all below 0.002.

Here, it important to note that the original models of PSSpred and PSIPRED were trained based on 2011 and 2016 sequence databases, respectively. Although secondary structure predictions, as well as the contact and threading programs studied in previous sections, are usually sensitive to the sequence databases and MSAs that the models are originally trained on, we do not attempt to re-train the models using the new DeepMSA profiles. In this context, the performance improvement should be mainly attributed to the sensitive and comprehensive information that DeepMSA provides, compared to the MSAs generated by other default programs.

Table 4. Summary of SS prediction by PSSpred and PSIPRED for 211 "Hard" targets. Bold font indicates the higher value in each category.

Predictor	MSA	Q3	P-value	SOV	P-value
PSSpred	PSI-BLAST + UniRef90	80.518	1.38E-03	77.257	1.05E-03
	DeepMSA	**81.472**	*	**78.660**	*
PSIPRED	PSI-BLAST + UniRef90	82.796	1.61E-03	79.401	2.00E-03
	DeepMSA	**83.616**	*	**80.601**	*

* Each *p*-value is calculated by one-tailed paired t-test to test whether DeepMSA has significant better SS prediction accuracy than the respective profile.

2.4 Discussion and Conclusion

We developed an open-source pipeline, DeepMSA, aiming to collect deep and sensitive multiple sequence alignments from whole-genome and metagenome sequence databases. Large-scale benchmark experiments show that DeepMSA consistently improves protein contact prediction, fold-recognition, and secondary structure prediction, compared to the widely-used HHblits, Jackhmmer and PSI-BLAST sequence searching programs. For example, the use of MSAs from DeepMSA improves top L long-range contact prediction precision of CCMpred by 24.4% compared to the default use of the HHblits MSAs by the program. Similarly, MUSTER threading identifies correct templates for 64.0% more "Hard" targets by switching the default PSI-BLAST profiles to the DeepMSA profiles. Notably, all improvements in contact prediction,

secondary structure prediction and threading have been achieved without retraining predictor model and parameters in neural networks or dynamic programming alignment.

The high quality of MSA by DeepMSA is partly due to the greater coverage and alignment depth resulted from the combination of diverse source of sequence databases. However, benchmark study shows that deeper MSA with more sequence homologs does not always lead to better contact prediction, since the final effect of MSAs is often a tradeoff of sequence coverage and alignment accuracy. Further analysis reveals that appropriate incorporation of multiple sequence search and alignment algorithms is the key to generate high quality MSAs by DeepMSA. In particular, HMMER alignment reconstruction by custom HHblits database generation is found to be especially helpful: a baseline method ("No custom db" in Table 1 and Table 2) without the custom HHblits database generation step results in 1.0% to 4.2% worse top L long-range contact prediction accuracies than DeepMSA, even when both methods use identical sequence databases.

The on-line server and the standalone program of DeepMSA are freely available at https://zhanglab.ccmb.med.umich.edu/DeepMSA/. An updated version of LOMETS [181] meta-server for threading-based protein structure prediction using sequence profiles generated by DeepMSA is available at https://zhanglab.ccmb.med.umich.edu/LOMETS/. The continuous developments of robust MSA and profile construction methods should help enhance the usefulness and impacts of the whole-genome and metagenomics initiatives on the structure and function prediction studies of the community. For example, the current DeepMSA program runs only with monomer proteins, while an extension of the program for protein-protein complex MSA construction is important and under progress[237].

Chapter 3 D-QUARK: *Ab Initio* Protein Folding Assisted by Deep Learning Predicted Distance and Orientations

3.1 Introduction

While reassembly and refinement of threading templates remains the most reliable protocol for protein structure prediction, the effectiveness of such template-based modeling (TBM) approaches are contingent upon the availability of good templates. The need to avoid template dependency for distant- and non-homology targets has led to the development of *ab initio* approach, or template-free modeling (FM) approach for protein structure prediction. Much recent progresses in *ab initio* protein folding are fueled by the usage of deep learning predicted contact maps[190,238]. Although a contact-map constraints which pairs of residues should be close to each other, it provides limited insight on the exact distance between interacting residue pairs due to its binary (contact versus non-contact) nature.

To address the limitation of contact map, three groups (AlphaFold[173,194], RaptorX-Contact[171], and DMPfold[172]) have almost simultaneously proposed the extension of contact map to distance map to guide *ab initio* protein folding. Instead of representing inter-residue interaction of a protein as the $L \times L$ binary contact map, where L is the target sequence length, a distance map has the shape of $L \times L \times K$. Each pixel in the distance map provides a probability distribution of the distance between a residue pair over a series of K distance bins. The use of distance bins enables the AlphaFold, a predictor based on relatively simple gradient descent-based conformation sampling approach to generate *ab initio* structure models with accuracy rivals to that of state-of-the-art TBM and FM methods.

Despite the success of distance-based protein folding, two issues remains. Firstly, the distant-based protein folding methods previous methods are mainly based on distant-geometry[171,172] and/or gradient descent based on Limited-memory Broyden-Fletcher-Goldfarb-Shanno (L-BFGS) optimization[173,194,196], which can be trapped in local minima. For example, AlphaFold needs to run 5000 separate L-BFGS runs for a single target protein to ensure the conformation space is sufficiently covered[194], making an otherwise light weighted simulation computationally intensive. Moreover, distance geometry or L-BFGS usually requires the energy function to be expressed as upper-lower bounds or as smooth and differentiable functions. This limits the possible functional forms the energy function can take, and prevents the incorporation of statistical energy functions which are usually not differentiable. Another inherent limitation of distance maps is its inability to differentiate different types of interactions. For example, even though close interaction is needed in both alpha helix packing and beta strand pairing, an inter-helix residue pair usually adopts antiparallel or perpendicular orientation, while beta pairing residues are almost always parallel. This is part of the reason orientation-dependent statistically energy functions almost always outperform orientation-independent energy[174,175,239]. The importance of orientation necessitates the prediction and incorporation of orientation maps[196] into folding simulations.

In this work, we present D-QUARK, a protein folding algorithm that combines deep learning predicted distance and orientation maps with inherent QUARK statistical energy[47] for comprehensive replica-change Monte Carlo (REMC) simulation. It differs from mainstream contact- and distance-based *ab initio* protein folding protocol in its unique flat-well distance potential, and a careful balance between statistical energy and deep learning derived restraints.

3.2 Methods

The pure *ab initio* protein folding pipeline of D-QUARK consists of three main steps (Figure 8): deep learning-based distance and orientation prediction, distance- and orientation guided REMC simulation, and clustering of simulation decoys for final atomic refinement.

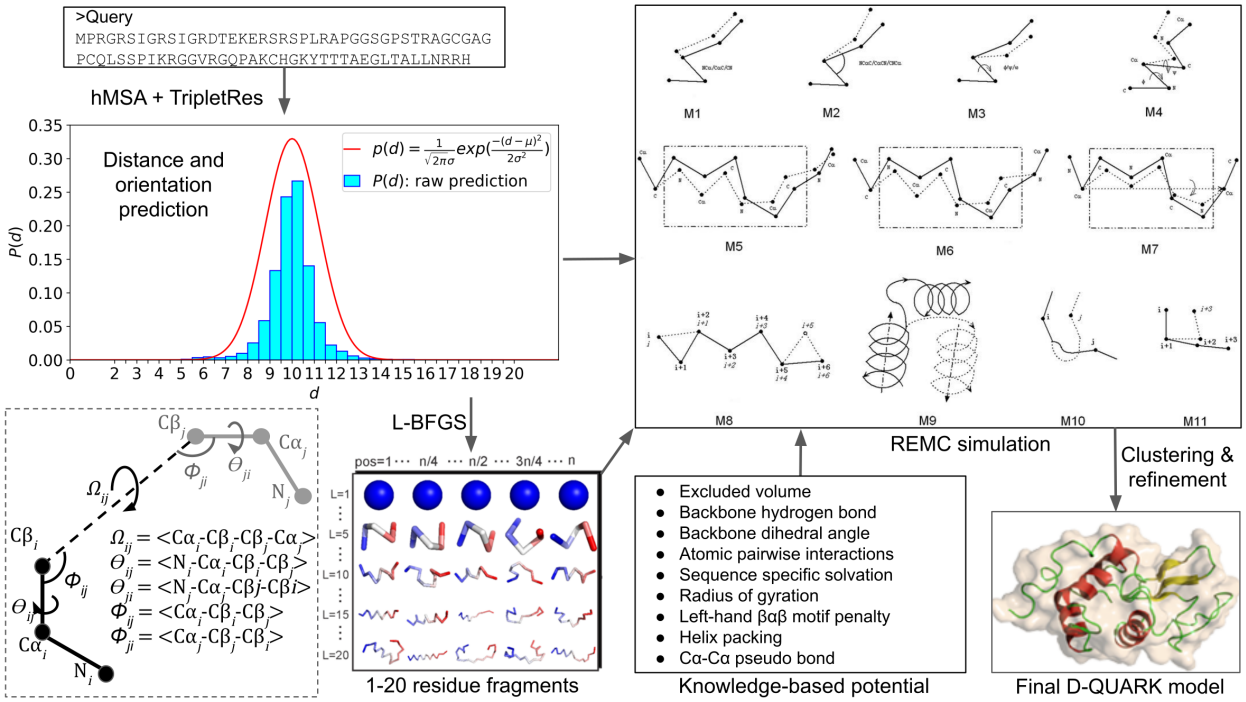


Figure 8. The D-QUARK pipeline for distance-based protein folding. D-QUARK consists of the following steps: residue-residue distance and orientation prediction by TripletRes deep learning model using hMSA sequence alignment; distance- and orientation-guided fragment generation; assembly of fragment by REMC simulation guided by a composite force field combining distance and orientation prediction and knowledge-based potential; clustering and refinement for final model generation. The lower left inset depicts the geometric definition of dihedral angles Ω and Θ as well as the angle Φ for orientation prediction between residue i (black) and residue j (grey), where Ω is symmetric ($\Omega_{ij}=\Omega_{ji}$) while Θ and Φ are asymmetric ($\Theta_{ij}\neq\Theta_{ji}$ and $\Phi_{ij}\neq\Phi_{ji}$).

3.2.1 Distance and orientation prediction

The input for our distance- and orientation map predictor is the multiple sequence alignments (MSA) constructed for the target sequence. This is performed by two complementary approaches: DeepMSA (Figure 9A) and qMSA (Figure 9B), using three metagenome sequence databases (Metaclust, BFD and Mgnify) and two whole-genome sequence databases (Uniclust30

and UniRef90). Here, DeepMSA[240] is our previous MSA construction program developed in CASP13. In the three stages of DeepMSA, HHblits2, Jackhmmer and HMMsearch were used to search the query against Uniclust30 (version 2017_04), UniRef90 and Metaclust, respectively. In Stage 2 and 3, homologs identified by Jackhmmer and HMMsearch, respectively, are constructed into a custom HHblits format database, which will be searched through by HHblits2 using the MSA input from the previous stage to generate new MSAs. As an extension of DeepMSA, qMSA (standing for "quadruple MSA") has four stages to perform HHblits2, Jackhmmer, HHblits3, and HMMsearch searches against Uniclust30 (version 2020_01), UniRef90, BFD, and Mgnify, respectively. Similar to DeepMSA Stage 2 and 3, the sequence hits from Jackhmmer, HHblits3 and HMMsearch in Stage 2, 3 and 4 of qMSA are converted into HHblits format database, against which the HHblits2 search based on MSA input from the previous stage is performed. These steps result in 7 MSAs in total (i.e., 3 from DeepMSA and 4 from qMSA). These MSAs are scored by a deep learning contact predictor, TripletRes[232], where a single MSA (referred to as hybrid MSA, or hMSA) with the highest probabilities for top 10L (L is the sequence length) all range contacts (Cβ-Cβ distances<8Å) will be selected.

The selected MSA is used by the full-version TripletRes program (Figure 9C) to calculate raw coevolutionary input features from covariance (COV) statistics and pseudo-likelihood maximization (PLM), which is shown to result in significantly more accurate neural network learning[232] than previous predictors[162,166,241] that use post-processed coevolution features. The output of this full-version TripletRes are a set of spatial restraints, including the Cα-Cα distance, Cβ-Cβ distance and inter-residue orientations. The distances are predicted in the form of 38 distance bins (1 bin for <2Å, 36 bins for 2 to d_{cut}=20Å with bin width 0.5Å, and 1 bin for ≥20Å),

while torsional angles are predicted with bin width of 15˚ plus an additional bin for no interaction (i.e. Cβ-Cβ distance ≥20Å).

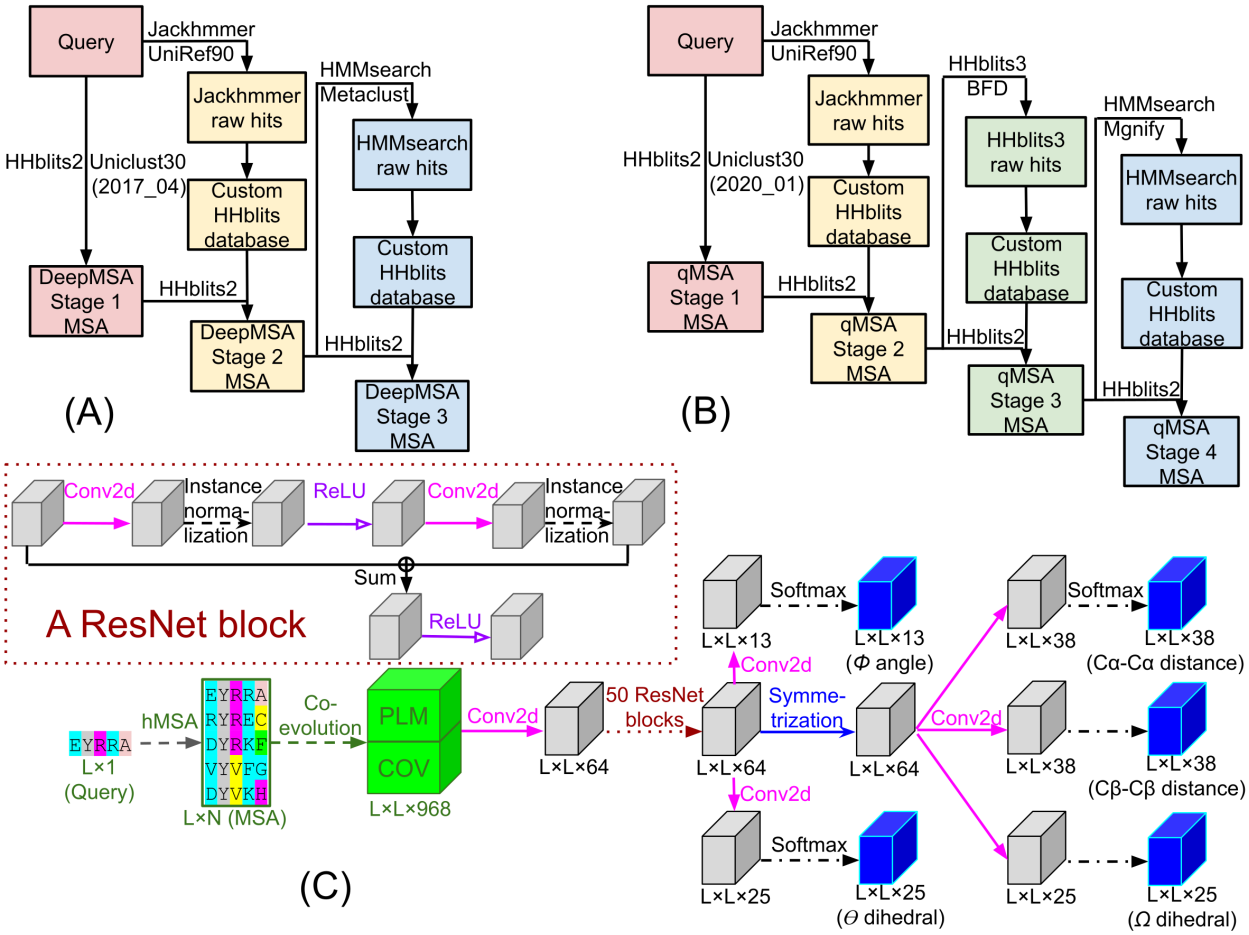


Figure 9. MSA generation and distance prediction in D-QUARK. (A-B) MSA generation by DeepMSA (A) and qMSA (B). (C) The neural network architecture of extended TripletRes for prediction of distances (Cα and Cβ) and orientations (Ω, Θ and Φ). TripletRes used 50 ResNet blocks, whose architecture is shown in the inset. The two sets of input features of TripletRes are calculated from the MSA by covariance (COV) and pseudo-likelihood maximization (PLM), with each set of feature in the form of an $L\times L\times 484$ matrix. Here, L is the length of query while 484=22×22 is the number of amino acid type combinations for a pair of residues (22=20 standard amino acid types plus 2 types for a gap and a non-standard amino acid).

3.2.2 Implementation of distance and orientation potential in protein folding simulation

These distance and torsion angle restraints are used to generate continuous fragments ranging from 1 to 20 residues by short L-BFGS simulations. The simulation is guided by the negative log probability potential for both Cα-Cα and Cβ-Cβ distances (d), and orientations (o):

$$E_{log}(d) = \begin{cases} -\log\left(\frac{P(d)+\epsilon}{P(d_{cut})+\epsilon}\right) + \alpha \cdot \log\left(\frac{d}{d_{cut}}\right), d < d_{cut} \\ \alpha \cdot \log\left(\frac{d}{d_{cut}}\right), d \geq d_{cut} \end{cases} \quad (3.1)$$

$$E_o(o) = -\log\left(\frac{P(o)+\varepsilon}{\varepsilon}\right) \quad (3.2)$$

ε=1e-4 is a pseudo-count to avoid division by or logarithm of zero. α=1.57 is a constant parameter for the distance-scaled finite ideal-gas reference state[242] of a distance potential. L-BFGS requires every individual energy term to be continuous and differentiable so that gradients can be calculated, while the raw probability distribution is binned (i.e. discontinuous). For L-BFGS purpose, the above energy terms are converted to smooth forms by cubic spline fitting. Using different cutoffs ranging from to 0.55 to 0.95 for the probability of no interaction, 30 L-BFGS runs were performed to generate 30 conformations. From these 30 conformations, continuous fragments ranging from 1 to 20 residues are extracted.

The fragments are assembled by a replica-exchange Monte Carlo (REMC) simulation extended from QUARK[47]. This REMC has 40 temperature replicas and 12 different types of movements, and is guided by a composite force field comprising of knowledge-based energy terms inherited from QUARK[47], the same orientation potential shown in Eq (3.2), and a different flat-well distance potential:

$$E_{well}(d) = \begin{cases} 1{,}3\sigma \leq |d-\mu| \\ 1 - \frac{1}{2}\left(\left|\frac{d-\mu}{\sigma}\right| - 1\right)^2, 2\sigma \leq |d-\mu| < 3\sigma \\ \frac{1}{2}\left(\left|\frac{d-\mu}{\sigma}\right| - 1\right)^2, \sigma < |d-\mu| < 2\sigma \\ 0, |d-\mu| \leq \sigma \end{cases} \quad (3.3)$$

μ and σ are the mean and standard deviation, respectively, of the distance prediction. These two parameters are from Gaussian fitting of the raw probability distribution by minimizing the following objective function using simplex optimization[243]:

$$f(\mu,\sigma) = -\sum_{k=1}^{38} P(k) \cdot log\left[\int_{lb_k}^{ub_k} \frac{1}{\sqrt{2\pi}\sigma} exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) dx\right] \quad (3.4)$$

Here, lb_k and ub_k are the lower and upper bounds of the kth distance bin, while $P(k)$ is the TripletRes predicted probability of the kth bin. In addition to the flat-well and negative log probability potential proposed above, we also tested REMC with three other forms of distance potentials, i.e., negative probability, flat bottom, and Hooke potentials in Eq (3.5), (3.6) and (3.7), respectively.

$$E_{prob}(d) = \begin{cases} -P(d), d < d_{cut} \\ 0, d \geq d_{cut} \end{cases} \quad (3.5)$$

$$E_{bottom}(d) = \begin{cases} \frac{(|d-\mu|-\sigma)^2}{2\sigma^2}, |d-\mu| > \sigma \\ 0, |d-\mu| \leq \sigma \end{cases} \quad (3.6)$$

$$E_{hooke}(d) = \begin{cases} \frac{(d-\mu)^2}{2\sigma^2}, |d-\mu| > \sigma \\ 0, |d-\mu| \leq \sigma \end{cases} \quad (3.7)$$

Unlike L-BFGS, the negative log probability and negative probability potentials in REMC can directly use the raw binned probability without spline fitting, as Monte Carlo simulation does not require energy terms to be differentiable. Additionally, α=0 is used in Eq (3.1) for REMC as it generates slightly more accurate results than those from α=1.57 in REMC.

3.2.3 Clustering and refinement

Approximately 25,000 decoy conformations are collected from the 10 replicas with the lowest temperatures in REMC. These decoys are clustered by pairwise RMSD using

SPICKER[244]. The cluster centroids from the five largest clusters are consecutively refined by ModRefiner[245] and FG-MD[188] to get the five final models, which are ranked in descending order of the cluster size. In this article, we mainly discuss the result for the first model from the largest cluster.

3.3 Results

3.3.1 Dataset

D-QUARK was benchmark on 301 non-redundant (pairwise sequence identity <30%) PDB chains ranging from 51 to 286 residues. Since D-QUARK was developed specifically for FM folding, the benchmark dataset only include target proteins determined as "hard" and "very hard" by LOMETS2[181] and with first threading template TM-score[246,247] less than 0.5.

3.3.2 Cα-Cα distances can be predicted more accurately than Cβ-Cβ distances

Most state-of-the-art distance-based protein folding algorithms [172,194,196] chose to use Cβ-Cβ distances, and only resort to use Cα atom when Cβ atom is unavailable (i.e. glycine residues). However, Cα-Cα distances should theoretically also be predictable by deep learning as Cα-Cα distances. Indeed, as shown in Table 5, TripletRes Cα-Cα distance predictions is actually even more accurate than its Cβ-Cβ distance predictions, which are in turn more accurate than those from third-party predictors (trRosetta and DMPfold).While we mainly measure the accuracy of distance prediction by Root Mean Square Error (RMSE) of distance prediction for the top *L* long range residues pairs (i.e. pairs of residues separated by ≥24 residues in sequence), we also indirectly measure the prediction accuracy in terms of precisions for top *L* long contact, where a distance map is converted to a contact map by summing up the predicted probabilities of all bins

for <8Å. Interesting, despite Cβ-Cβ has less accurate distance predictions than Cα-Cα in TripletRes, the former has more accurate contact predictions in terms of precision for top *L* long range. These data suggests that Cα-Cα and Cβ-Cβ provides complementary information to protein folding, and justify our incorporation of TripletRes distances for both atom types to D-QUARK.

Table 5. Top *L* long range contact precision and distance RMSE by different distance predictors.

Predictor	Atoms	Contact precision (*p*-value)	Distance RMSE (*p*-value)
TripletRes	Cα-Cα	0.466 (3.75E-21)	**1.771** (*)
	Cβ-Cβ	**0.516** (*)	1.896 (6.99E-05)
trRosetta	Cβ-Cβ	0.438 (1.43E-44)	2.350 (4.23E-18)
DMPfold	Cβ-Cβ	0.376 (1.20E-67)	3.153 (9.62E-27)

* All *p*-value is calculated by one tailed t-test against TripletRes predicted Cβ-Cβ contacts and Cα-Cα distances. The most accurate result (highest precision and lowest RMSE) is highlighted in bold.

3.3.3 Contact-based MSA selection improves the quality of MSA

Since the hybrid MSA in D-QUARK is selected by the highest sum of contact scores rather than largest number of effective sequence (*Nf*) as in previous studies[238,240], we tested whether this strategy indeed improves quality of selected MSA. The MSA quality is quantified by two metrics in the TripletRes prediction for top L long range pairs: Cβ-Cβ contact precision and Cα-Cα distance RMSE. As shown in Table 6, contact-selected MSA consistently outperforms *Nf*-selected MSA for both qMSA and DeepMSA in both contact and distance prediction. Additionally, although qMSA has a roughly similar performance (slightly better contact precision but slight worse distance RMSE) as DeepMSA, their combination by contact-based MSA selection leads to a consistently better quality for the final hMSA than DeepMSA or qMSA alone.

Table 6. MSA quality for different MSA generation and selection approaches, measured by TripletRes Cβ-Cβ contact precision and Cα-Cα distance RMSE for top *L* long range residue pairs.

Method	Selection	Contact precision (*p*-value)	Distance RMSE (*p*-value)
DeepMSA	*Nf*	0.497 (2.25E-7)	1.820 (3.83E-1)
	Contact	0.501 (3.96E-6)	1.820 (7.39E-2)
qMSA	*Nf*	0.501 (1.38E-3)	1.853 (1.30E-1)
	Contact	0.509 (2.92E-2)	1.823 (1.98E-1)
hMSA	Contact	**0.516** (*)	**1.771** (*)

* All *p*-value is calculated by one tailed t-test against hMSA, which is the final MSA used by D-QUARK. The best MSA quality (highest precision and lowest RMSE) is in bold.

3.3.4 Functional form of distance potential has a profound impact on protein folding

We access the performance of D-QUARK using different distance potentials, in comparison with four other state-of-the-art protein folding programs incorporating predicted contacts (C-QUARK and C-I-TASSER[238]) or distances (DMPfold[172], and trRosetta[196]), as well as the original QUARK algorithm[47] without predicted contacts or distances. The performance is evaluated by first model TM-score, first model RMSD, and the success rate (i.e., number of targets with TM-score>0.5 divided by total number of targets). For this benchmark, C-QUARK and C-I-TASSER use their built-in DeepMSA alignments for contact prediction, while DMPfold and trRosetta, which do not depend on specific built-in MSA generator, uses the same MSA as D-QUARK for distance prediction. For QUARK, C-QUARK and C-I-TASSER, which use either local templates fragments or full length templates, all structure templates sharing ≥30% sequence identity to the target protein in order to emulate real-life scenario of *ab initio* modeling.

Table 7. Performance of D-QUARK in comparison with other third-party programs.

Program	D-QUARK energy†	TM-score (*p*-value)	RMSD (*p*-value)	Success rate
QUARK		0.296 (6.24E-108)	13.6 (1.79E-68)	3.0%
C-QUARK		0.431 (6.72E-72)	9.87 (1.43E-32)	35.6%
C-I-TASSER		0.448 (1.35E-66)	9.48 (6.46E-30)	40.9%
DMPfold		0.503 (7.06E-41)	9.06 (6.34E-28)	51.5%
trRosetta		0.555 (5.04E-32)	7.91 (1.14E-15)	61.5%
D-QUARK	E_{log}	0.456 (3.77E-58)	9.29 (1.29E-30)	39.2%
	E_{bottom}	0.456 (3.59E-56)	9.50 (9.91E-31)	39.9%
	E_{prob}	0.462 (3.74E-55)	9.54 (1.09E-31)	40.5%
	E_{Hooke}	0.463 (8.21E-52)	9.39 (7.42E-30)	41.2%
	E_{well}	0.470 (4.00E-54)	9.20 (3.74E-29)	43.2%
	$E_{well} + E_o$	0.618 (*)	6.56 (*)	73.4%

† E_{log}, E_{bottom}, E_{prob}, E_{hooke}, E_{well}, and E_o represents negative log probability, flat bottom, negative probability, Hooke, flat well distance potentials, and the orientation potential, respectively.
* All *p*-values are calculated by one-tailed t-test with regard to the final D-QUARK version (E_{well} +E_o).

Several observations can be made from Table 7 and Figure 10 that summarize the benchmark. Firstly, even when using the same deep learning derived distance prediction, different functional forms of distance potential implemented by D-QUARK can lead to different performance, with the best potential (E_{well}) resulting in 3% higher TM-score than the worse potential (E_{log}) after weight tuning. Secondly, all three distance-based protein folding programs (DMPfold, trRosetta and D-QUARK) outperform protein folding programs that only use contact information (C-QUARK and C-I-TASSER), which in turn outperform protein folding algorithm without contact or distance (QUARK). These data show the advantage of distance-based protein folding over protein folding with contact or without any contact/distance restraints. Thirdly, D-QUARK with both deep learning predicted distance and orientations (E_{well}+E_o) significantly outperforms in-house and third-party distance-based protein folding protocols without orientations (D-QUARK E_{log}, E_{bottom}, E_{prob}, E_{Hooke}, E_{well} and DMPfold), suggesting that both distance and orientation potentials in protein folding. Fourthly, even though both D-QUARK and trRosetta combine deep learning derived distances and orientations, D-QUARK outperforms trRosetta due to both more accurate deep learning prediction (Table 6) and better protein folding simulation. Finally, while we are unable to compare D-QUARK and AlphaFold on the full benchmark dataset in Table 7 due to the lack of feature generation and protein folding programs from AlphaFold, we perform a comparison in Figure 10E for a smaller subset of CASP13 FM targets submitted by AlphaFold (group A7D in CASP13). D-QUARK outperforms AlphaFold in terms of average TM-score (0.626 versus 0.579, *p*-value=1.78E-2) and success rate (76.7% and 63.3%).

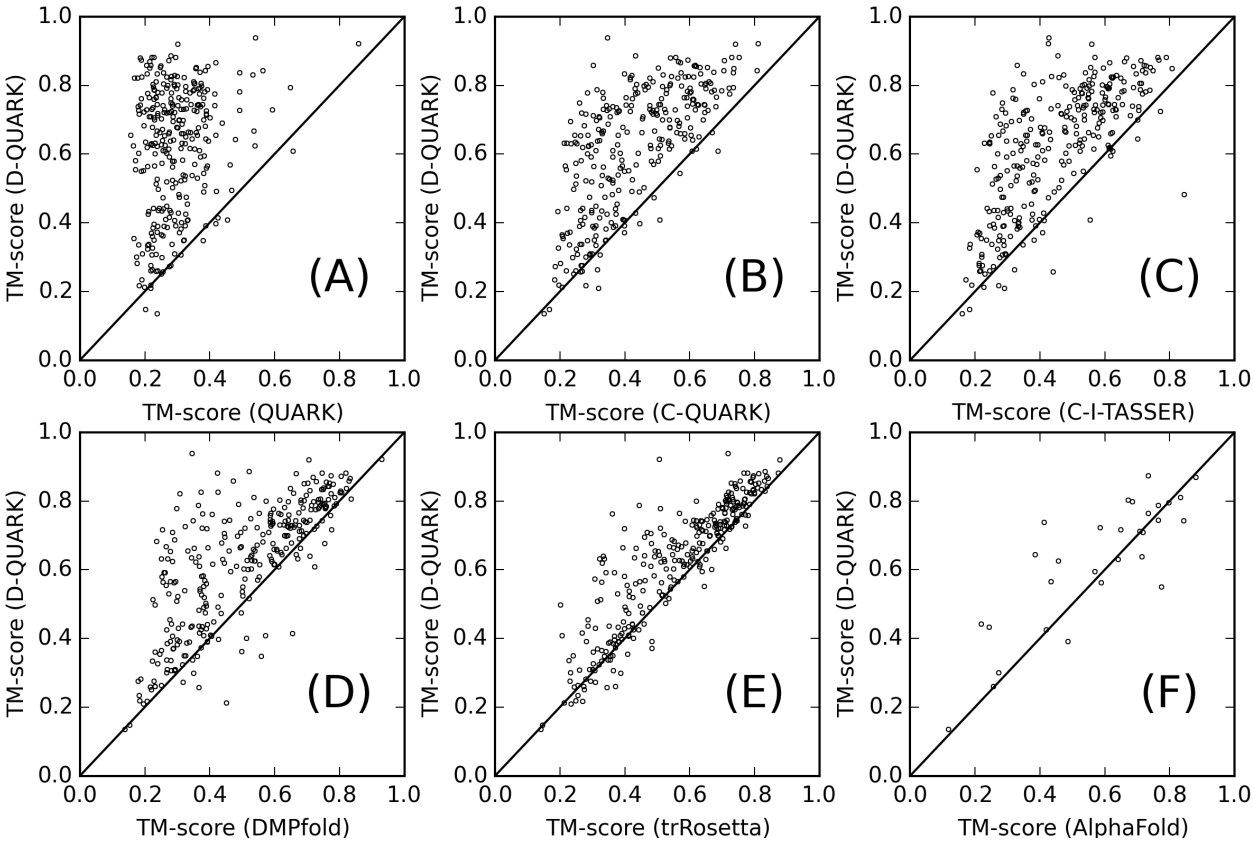


Figure 10. Head-to-head comparison of first model TM-score between (A) QUARK, (B) C-QUARK, (C) C-I-TASSER, (D) DMPfold, (E) trRosetta, (F) AlphaFold and D-QUARK. Each circle represents one target protein in the benchmark dataset. The comparison between D-QUARK and AlphaFold is on a subset of 31 FM targets that AlphaFold submitted their prediction in CASP13.

3.3.5 A case study based on preliminary assessment of D-QUARK in CASP14

D-QUARK participated in the most recent CASP14 challenge as two automated servers: "Zhang_Ab_Initio" and "QUARK". As a case study, we discuss T1040, which corresponds to a single domain (residue 1372-1501) from the RNA polymerase in Cellulophaga phage (PDB ID: 6vr4 Chain A). This is a particularly challenging target with no sequence homologs from DeepMSA and only 11 homologs from qMSA. All servers failed to generate a correct first model. Nonetheless, "Zhang_Ab_Initio" first model is closest to the native structure among all servers with TM-score=0.498, indicating a roughly correct topology, while model 2 (the best Zhang_Ab_Initio model) has TM-score=0.521 (Figure 11AB). Part of the reason for the

advantage of Zhang_Ab_Initio over other serves on this target is that all its structure templates are incorrect (TM-score=0.174 and 0.298 for first and best LOMETS templates, respectively, Figure 11CD). This shows the power of D-QUARK for non-homolog FM targets, especially when TripletRes has a reasonable prediction accuracy (distance RMSE=1.385Å, and top *L* contact precision 0.392). The main reason for the modest TM-score is the incorrect orientation of C-terminal helical segment (residue 97-130, black in Figure 11A-D for model and red for native structure), which is incorrectly packed against the residue 51-64 (green helix in Figure 11). This is caused by false positive prediction of interaction between residue 62 to 63 and residue 109 to 113 in the TripletRes prediction (red box in Figure 11E, grey arrow in Figure 11A), which are actually quite far apart in the native structure (red arrow in Figure 11A). This suggests that a small number of false positive predictions can mislead the protein folding process.

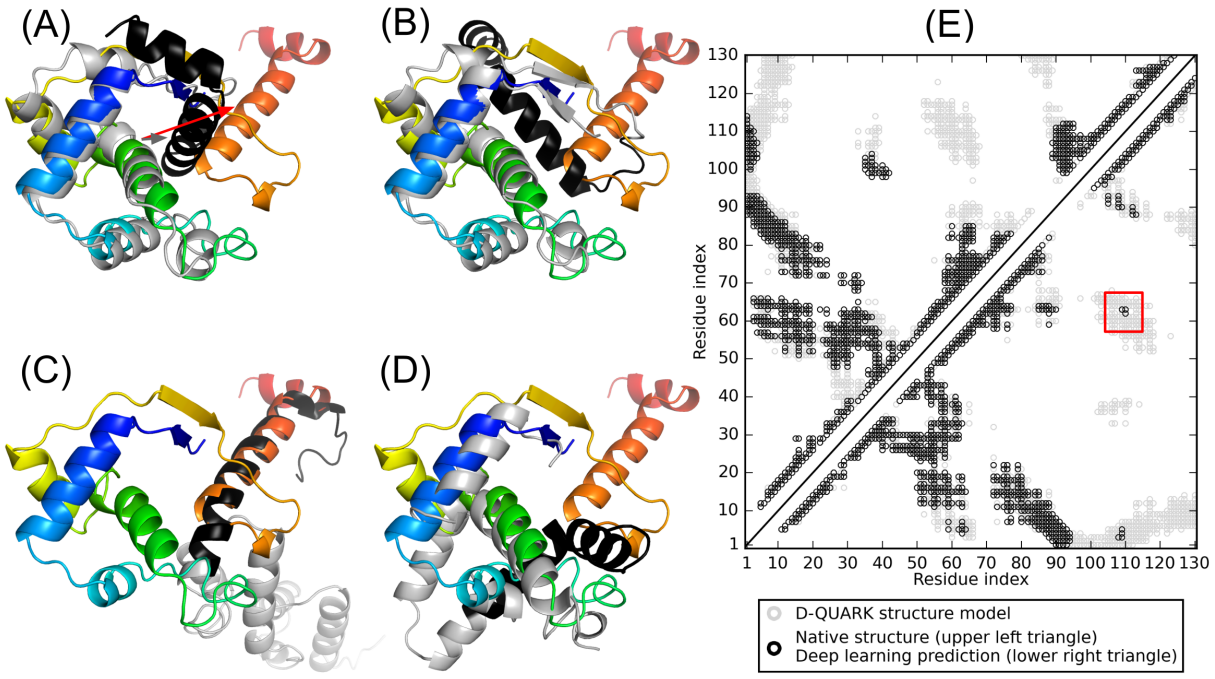


Figure 11. Modeling of T1040. (A-D) Structure models (residue 1 to 96 in grey cartoon; residue 97-130 in black) of D-QUARK Model 1 (A), Model 2 (B), LOMETS Template 1 (C; PDB ID: 3d5l Chain A) and Template 129 (D; PDB ID: 6iv9 Chain A), superposed to the native structure (PDB ID: 6vr4 Chain A; spectrum cartoon colored from N- to C- terminal in blue to red). (B) Distance map for residue pairs with Cα-Cα distance<13Å in native structure (upper left black circles), TripletRes prediction (lower right black circles) and D-QUARK Model 1 (light grey circles).

3.4 Discussion and Conclusion

We developed D-QUARK, an *ab initio* protein folding algorithm incorporating deep learning-predicted inter-residue distance and orientations. On a benchmark dataset of 301 hard targets D-QUARK consistently outperforms state-of-the-art template-based and *ab initio* protein structure predictions programs guided by deep learning predicted contacts, distances, and orientations. Detailed analysis showed that the advantage of D-QUARK can be attributed to better quality MSA to generate input features for deep learning, a more accurate deep learning model for distance/orientation prediction, and the REMC simulation with carefully designed energy terms to incorporated predicted distance and orientations restraints.

All deep learning derived inter-residue restraints as probabilities of different distance and orientation bins rather than real values. The binned prediction may inherently imposes a resolution limit in the restraint prediction, even though we try to circumstance this issue by deriving the real value distance through Gaussian fitting of the probability bins. In the future, a real value distance/orientation deep learning model may be able to better model the inter-residue restraints.

Chapter 4 COFACTOR: Structure and Interaction-based Protein Function Prediction[1]

4.1 Introduction

Due to recent advances in high-throughput sequencing technology, the gap between the numbers of protein sequences and number of those with experimentally characterized functions is quickly growing. As of 2017, for example, there are more than 60 million protein sequences deposited in the UniProt database [248], but fewer than 0.8% of these sequences have the functions manually annotated in SwissProt [249]. Automated and yet accurate *in silico* protein function prediction thus becomes crucial for making use of the recent explosion of genomic sequencing data. Most of the current function prediction approaches are based on sequence homologous transfer [250], which may not be able to achieve the remarkable task since more than 80% of un-annotated protein sequences lack close functional homologs (i.e., sharing greater than 60% sequence identity), and 25% of un-annotated proteins lack any homologs sharing a sequence identity above 30% in the current databases. Given that the function of a protein is ultimately defined by its structure, COFACTOR [70,251] has been previously proposed to transfer functional insights to the unknown proteins from structural homologies, which provides an alternative approach to annotating non-homologous targets that sequence-homology based methods cannot model effectively [250].

---

[1] This chapter was adapted from two previously-published works. The first work was Zhang C, Freddolino PL, Zhang Y (2017) "COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. " *Nucleic Acids Research*, 45(W1), W291-W299. The second work was C Zhang, W Zheng, PL Freddolino, and Y Zhang (2019) "MetaGO: Predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein-protein network mapping." *Journal of Molecular Biology*, 430(15), 2256-2265. In our earlier works, the full function pipeline was referred to as "COFACTOR" while its subroutine for Gene Ontology prediction is called "MetaGO". To avoid readers' confusion, this thesis will consistently use the name "COFACTOR".

Function annotation using structural homology alone, however, suffers several deficiencies. First, global structural similarity does not always lead to functional similarity. For example, the TIM barrel fold [197] is adopted by many proteins covering 60 distinct EC classification [252] as well as many non-enzyme proteins. Even for proteins with similar functions, global fold based comparisons may fail because the proteins often share only the local binding or active sites with complete different folds [253]. Second, the current structure-function database is far from complete. For around 88% of proteins with known functions from the UniProt-GOA [254], for example, there are no experimentally solved structures in the PDB database [255], seriously limiting the power of structure-based detection of functional homologies. Finally, although structure is essential to protein function, the structure of proteins in cells is far from static and many functions are associated with cellular environment of the molecules and the molecular motion of disordered regions that do not have a structure on their own [256]. Therefore, composite approaches combining multiple and complementary information from different resources of sequence homologs and interaction networks should help increase the accuracy and coverage of the structure-based function annotations.

In this chapter, we report our recent enhancement of the COFACTOR webserver [251] to make use of hybrid models combining information from structure and sequence homologies, as well as protein-protein interaction networks, for optimal protein function predictions. In addition, considerable effort has been made to improve user's experience and facility in analysing and visualizing the modelling results, which include the introduction of new animation tools to display structural templates and ligand-protein interactions, and directed acyclic graphs (DAG) to visualize the gene ontology annotation hierarchy. The new COFACTOR server and the functional libraries are freely available at http://zhanglab.ccmb.med.umich.edu/COFACTOR/.

4.2 Methods

4.2.1 Gene Ontology Prediction

The approach of GO prediction in COFACTOR consists of structure, sequence, and protein-protein interaction based pipelines (Figure 12).

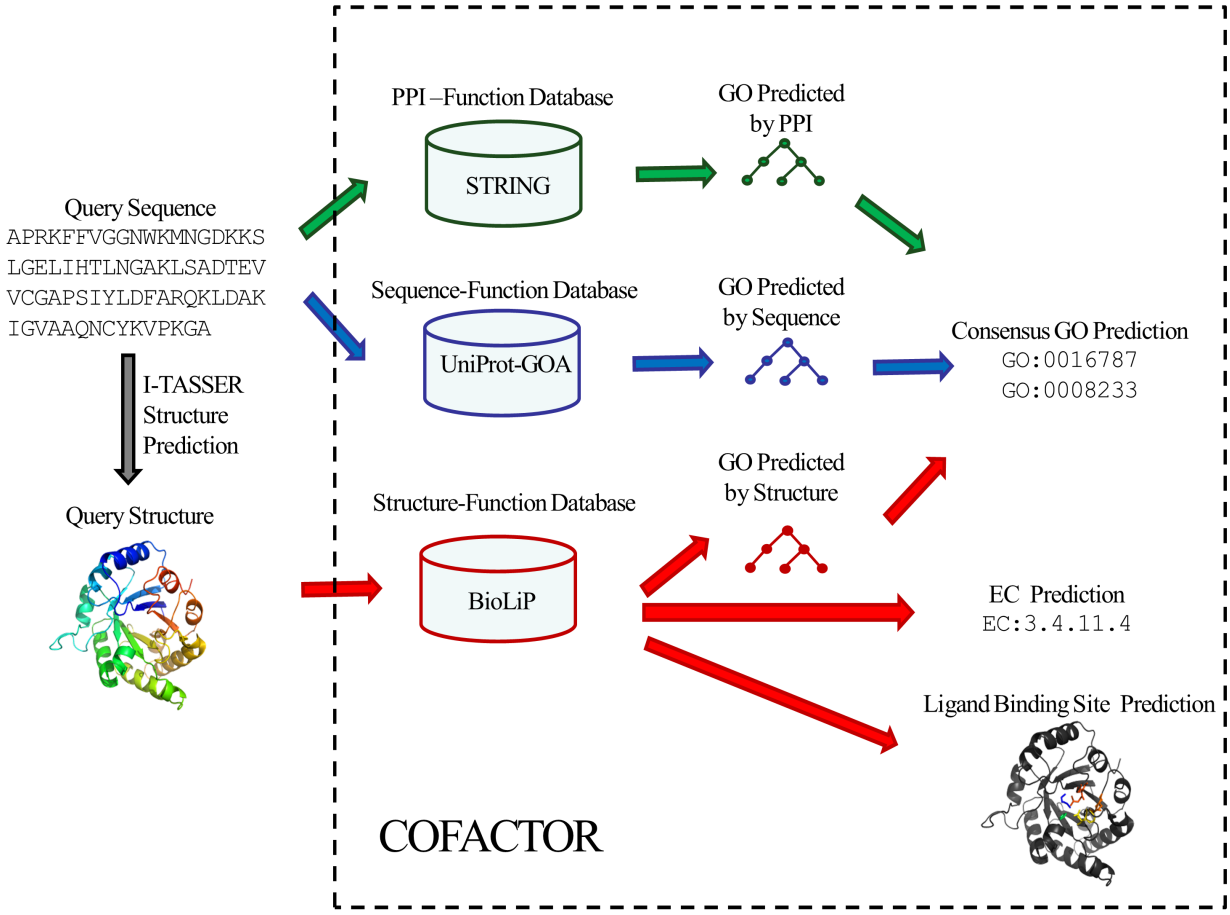


Figure 12. The workflow of COFACTOR for template-based function predictions. The method consists of three pipelines of functional template identifications. The GO models are derived from a consensus of the structure-, sequence- and PPI-based pipelines, while the EC and ligand-binding predictions are obtained from the structure-based template transfers.

Structure-based pipeline

The structure-homology based GO prediction method by COFACTOR was described previously [251]. Briefly, the query structure is compared to a non-redundant set of known proteins

in the BioLiP library [257], through two sets of local and global structural alignments based on the TM-align algorithm [258], for functional homology detections. Here, BioLiP is a semi-manually curated structure-function database containing known associations of experimentally solved structures and biological functions of proteins in terms of GO terms, EC number, and ligand binding sites. The local structure similarity between query and template is defined by

$$L_{sim} = \frac{1}{N_t}\sum_{i=1}^{N_{ali}}\left(\frac{1}{1+(d_i/d_0)^2} + M_i\right) \tag{4.1}$$

where N_t is the number of residues in the active/binding sites, N_{ali} is the number of aligned residue pairs, d_i is the $C\alpha$ distance between i^{th} aligned residue pair, d_0=3Å is the distance cut-off, and M_i is the BLOSUM62 substitution matrix score [259] between i^{th} pair of residues that has been normalized to the interval [0, 1]. The confidence score of a template hit is defined by

$$FCscore = \frac{2}{1+\exp\left(-(0.25\times L_{sim}\times SS_{bs} + TM + 2.5\times ID)\right)} - 1 \tag{4.2}$$

where TM is the global structure similarity in terms of TM-score [234] between query and template, ID is the sequence identity between query and template in the aligned region, and SS_{bs} is the sequence identity at the binding site. The overall confidence score for a particular GO term q is then calculated by

$$Cscore_{structure}(q) = 1 - \prod_{i=1}^{N(q)}\left(1 - FCscore_i(q)\right) \tag{4.3}$$

where $N(q)$ is the number of templates associated with the GO term q, and $FCscore_i(q)$ is the confidence score of the i^{th} hit associated with λ as defined in Equation (4.2). The predicted GO terms are reconciled using the PIPA algorithm [260].

Sequence-based pipeline

In the second pipeline, the query sequence is searched against the UniProt database through both sequence and sequence profile alignments by BLAST [261] and PSI-BLAST [102],

respectively. Only manually reviewed GO terms of sequence templates are considered, with GO terms annotated with IEA or ND evidence codes excluded. For BLAST, the query is directly searched against sequence template library with an e-value cut-off 0.01. The confidence score for a particular GO term q resulting from a BLAST search is defined by

$$GOfreq_{blast}(q) = \frac{\sum_{k=1}^{N(q)} s_k(q)}{\sum_{k=1}^{N} s_k} \quad (4.4)$$

where N is the number of templates identified, s_k is the sequence identity between the query and the k^{th} template, and $N(q)$ and $s_k(q)$ are those associated with a specific GO term q. For PSI-BLAST, a sequence profile is obtained by searching with the query sequence through the Uniref90 sequence library [223] by three iterations under an e-value cut-off 0.01. The sequence profile is used to jump start a PSI-BLAST profile-sequence search against the UniProt-GOA sequences. The confidence score for GO term λ is defined in the same way as in BLAST (Equation 4.4). The final weighted average confidence score of the sequence-based pipeline is calculated as

$$Cscore_{sequence}(q) = w \times GOfreq_{blast}(q) + (1 - w) \times GOfreq_{psiblast}(q) \quad (4.5)$$

where w equals to the maximum sequence identity of the query to all the template hits. In this way, BLAST hits have a stronger weight if close homologs are found, while the weight of the PSI-BLAST hits is increased for the non-homologous cases for which PSI-BLAST profile alignments are usually more efficient than the sequence based alignments.

PPI-based pipeline

In this pipeline, the query is first mapped to the STRING [262] protein-protein interactions (PPI) database by BLAST, with a sequence identity cut-off >0.9. GO terms of the interaction partners, as annotated in the STRING database, are then collected and assigned to the query protein. The underlying assumption is that the interacting partners tend to participate in the same

biological pathway at the same sub-cellular location and therefore may have similar GO terms. Finally, the confidence score for GO term q mapped by PPI is calculated by

$$Cscore_{PPI}(q) = \frac{\sum_{k=1}^{N(q)} str_k(q) \times S_k(q)}{\sum_{k=1}^{N} str_k \times S_k} \tag{4.6}$$

where N is the number of interacting partners, str_k is the confidence score of interaction between query and the k^{th} interaction partner as assigned by the STRING database, and S_k is the sequence identity in the first step of BLAST alignment for the k^{th} interaction partner. $N(q)$, $str_k(q)$, and $S_k(q)$ are those associated to the specific GO term q.

Consensus GO prediction

The final GO prediction is obtained by combining the GO terms from the structure, sequence, and PPI-based pipelines, with the confidence score calculated by

$$Cscore^{GO}(q) = 1 - \prod_{m} \left(1 - Cscore_m(q)\right)^{w_m} \tag{4.7}$$

where $m \in \{structure, sequence, PPI\}$. w_m is the relative weight for each of the three methods, with $w_{sequence} = w_{PPI} = 1$ and $w_{structure} = 1 - w$, where w equals to the maximum sequence identity among identified function templates. Hence, the weight of the structure-based model becomes stronger for the cases that have no homologous templates.

4.2.2 Enzyme Commission Number Prediction

The pipeline of EC number prediction is similar to the structure-homology based method used in GO prediction. Enzymatic homologs are identified by aligning the target structure, using TM-align [258], to a library of 8,392 enzyme structures from the BioLiP library [263], with the active site residues mapped from the Catalytic Site Atlas database [264]. The confidence score for each

predicted EC number is estimated based on the global and local similarity between the target and top template hit:

$$Cscore^{EC} = \frac{2}{1 + \exp\left(-(0.25 \times L_{sim} \times SS_{as} + TM + 2.5 \times ID)\right)} - 1 \quad (4.8)$$

where TM is the TM-score between query and template, ID is the sequence identity, SS_{as} is the sequence identity at the active sites, and L_{sim} is local structure similarity as defined in Equation (4.1).

4.2.3 Ligand Binding Site Prediction

Ligand binding prediction in COFACTOR consists of three steps. First, functional homologies are identified by matching the query structure through the BioLiP library [263], which contains 58,416 structure templates harbouring in total 76,679 ligand-binding sites for interaction between receptor proteins and small molecule compounds, short peptides, and nucleic acids. The initial binding sites are then mapped to the query from the individual templates based on the structural alignments.

Next, the ligands from each individual template are superposed to the predicted binding sites on the query structure using the superposition matrices from the local alignment of query and template binding sites. To resolve atomic clashes, the ligand poses are refined by a short Metropolis Monte Carlo simulation under rigid-body rotation and translation, guided by an empirical energy function of

$$E_{pose} = RMSD + N_{clash} - \sum_{i=1}^{N_{lig}} \frac{1}{1 + \left|d_i^t - d_i^q\right|} \quad (4.9)$$

where $RMSD$ is the RMSD of current ligand pose and the origin ligand pose, N_{clash} is the number of atomic clashes between ligand and protein, N_{lig} is the number of ligand atoms, d_i^t is

the distance between i^{th} ligand atom and the Cα atom of the template residue in contact with the ligand atom, and d_i^q is the distance between the same ligand atom and the closest query Cα atom.

Finally, the consensus binding sites are obtained by clustering all the ligands that are superposed to the query structure, based on distance of the centroids of mass of the ligands using a cut-off of 8 Å. Different ligands within the same binding pocket will be further grouped by the average linkage clustering with chemical similarity, using the Tanimoto coefficient [265] with a cut-off of 0.7. The model with the highest ligand-binding confidence score ($Cscore^{LBS}$) among all the clusters is selected, which is defined by

$$Cscore^{LBS} = \frac{2}{1 + \exp\left(-\frac{N}{N_{tot}}\left(0.25 \times L_{sim} + TM + 0.25 \times ID + \frac{2}{1 + \langle D \rangle}\right)\right)} - 1 \qquad (4.10)$$

where N is the number of ligands in the ligand cluster, N_{tot} is the total number of ligands collected from all the homologous template, L_{sim} is the local similarity at the binding site defined in Equation (4.1), TM is TM-score between query and template, ID is the sequence identity between query and template in the structurally aligned region, and $\langle D \rangle$ is the average distance between ligands within the cluster.

4.3 Results

4.3.1 Benchmark results on GO predictions

The COFACTOR GO pipelines have been benchmarked on a non-redundant set of 1,224 *E. coli* proteins from UniProt database, with lengths ranging from 38 to 968 residues and pairwise sequence identity <40%. The input structures for COFACTOR were predicted by I-TASSER [16] with all homologous structural templates with a sequence identity >30% to the query excluded, thus simulating predictions for a target without any close homologs. Similar to the

Critical Assessment of Function Annotation (CAFA) experiments [250,266], the GO performance is mainly assessed by the F-measure, which is defined as the harmonic average between precision and recall:

$$F_{max} = \max_{t}\left\{\frac{2 \times pr(t) \times rc(t)}{pr(t) + rc(t)}\right\} \quad (4.11)$$

where t is the confidence score threshold (ranging between 0 and 1), and $pr(t)$ and $rc(t)$ are the precision and recall at a threshold t.

Figure 13 shows the performance of the COFACTOR server on three aspects of GOs: molecular function (MF), biological process (BP) and cellular component (CC); results are shown in control with those of the GoFDR program [267], one of the top performing methods in CAFA2 [250], and three baselines methods: Naïve Baseline, BLAST and PSI-BLAST, as implemented in CAFA [250,266]. To examine the effect of the combination of complementary pipelines, we also show the results from individual COFACTOR components from structure, sequence and PPI pipelines. To test the dependence of the pipelines on the homologies, four levels of sequence identity cut-offs at 20%, 30%, 50%, and 90% were used separately to filter out homologous templates. Several interesting observations arise from this figure. First, whereas the performance of sequence-based methods (GoFDR, BLAST/PSIBLAST, and the sequence module of COFACTOR) declines rapidly below 50% sequence identity, the structure module of COFACTOR shows almost no loss of performance even down to 20% sequence identity, and at that point it outperforms all sequence-based methods. For example, $F_{\max}$ for MF is 0.538 at the 20% sequence identity cut-off, very close to 0.541 obtained at 50% cut-off. Second, the new sequence component of COFACTOR is a strong performer on its own, with performance exceeding all other sequence-based methods including GoFDR (except for the cases at very low homology cut-off), and thus provides a useful complement to the structure-based module in the

high sequence homology region. Finally, the hybrid COFACTOR model outperforms all other methods used in our comparison (including, interestingly enough, the Naïve method for CC term predictions, which was not beaten by any prediction set in the CAFA2 competition [250]), at all levels of sequence identity cut-offs.

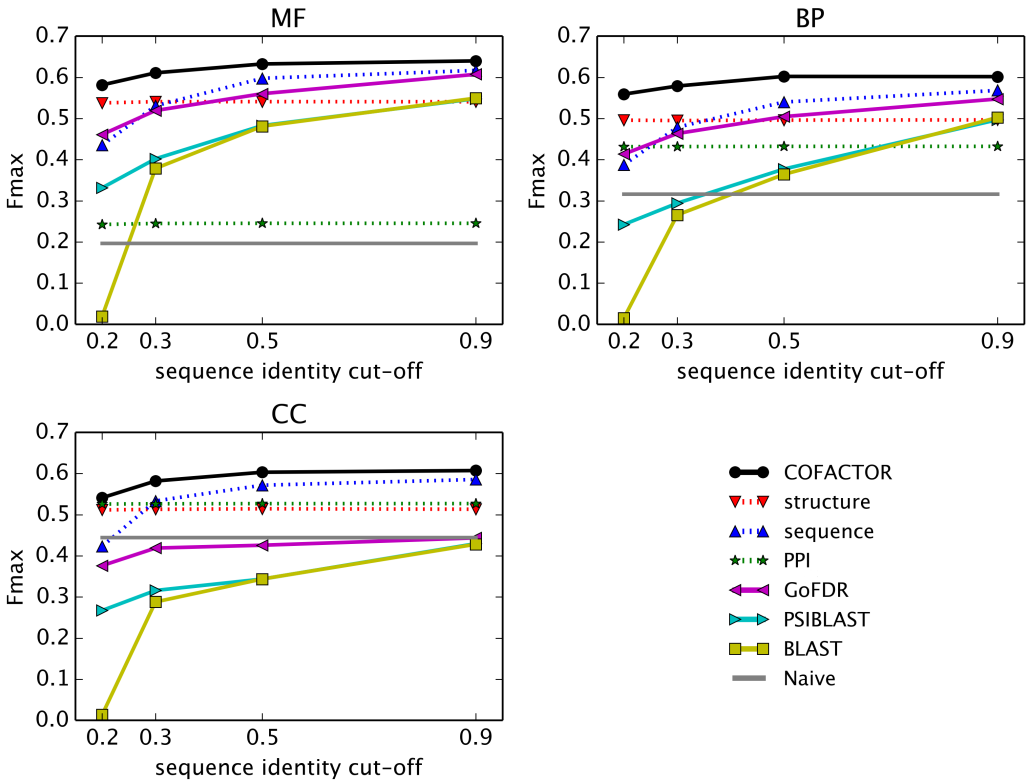


Figure 13. Accuracy of GO annotations by COFACTOR and control methods at different sequence identity cut-offs on a test set of 1,224 non-redundant proteins. Accuracy is evaluated by maximum F-measure. No sequence identity cut-off is imposed on Naïve, as it is not relevant. "structure", "sequence", and "PPI" are the individual structure-, sequence- and PPI-based pipelines in COFACTOR. "COFACTOR" is the consensus prediction combining the three pipelines. Only GO terms annotated by UniProt-GOA with experimental evidence codes (EXP, IDA, IMP, IGI, IEP, TAS, or IC) are considered as "gold standards". All parent GO terms of annotated GO terms are also considered annotations of each target. For the predicted GO terms, all their parent terms are also recursively propagated toward the root such that each parent term receives the highest confidence score among its children terms. The root term of the three GO aspects (MF, BP, and CC) and the extremely common "protein binding" term are excluded.

To examine the specificity of the COFACTOR prediction, we present in Figure 14A a histogram of precision of the GO prediction versus the confidence score by COFACTOR, where a strong correlation is found for all aspects of GO terms, with the Pearson correlation coefficient (PCC) being 0.96, 0.94, and 0.86 for MF, BP, and CC terms, respectively. Consistent with Fig.

S1, at the same $Cscore^{GO}$ cut-off the precision of MF and BP is generally higher than that of CC. For example, the precision for both MF and BP will be >0.3 when $cscore^{GO} > 0.6$, while the precision of CC is only marginally close to 0.3 when $cscore^{GO} > 0.8$.

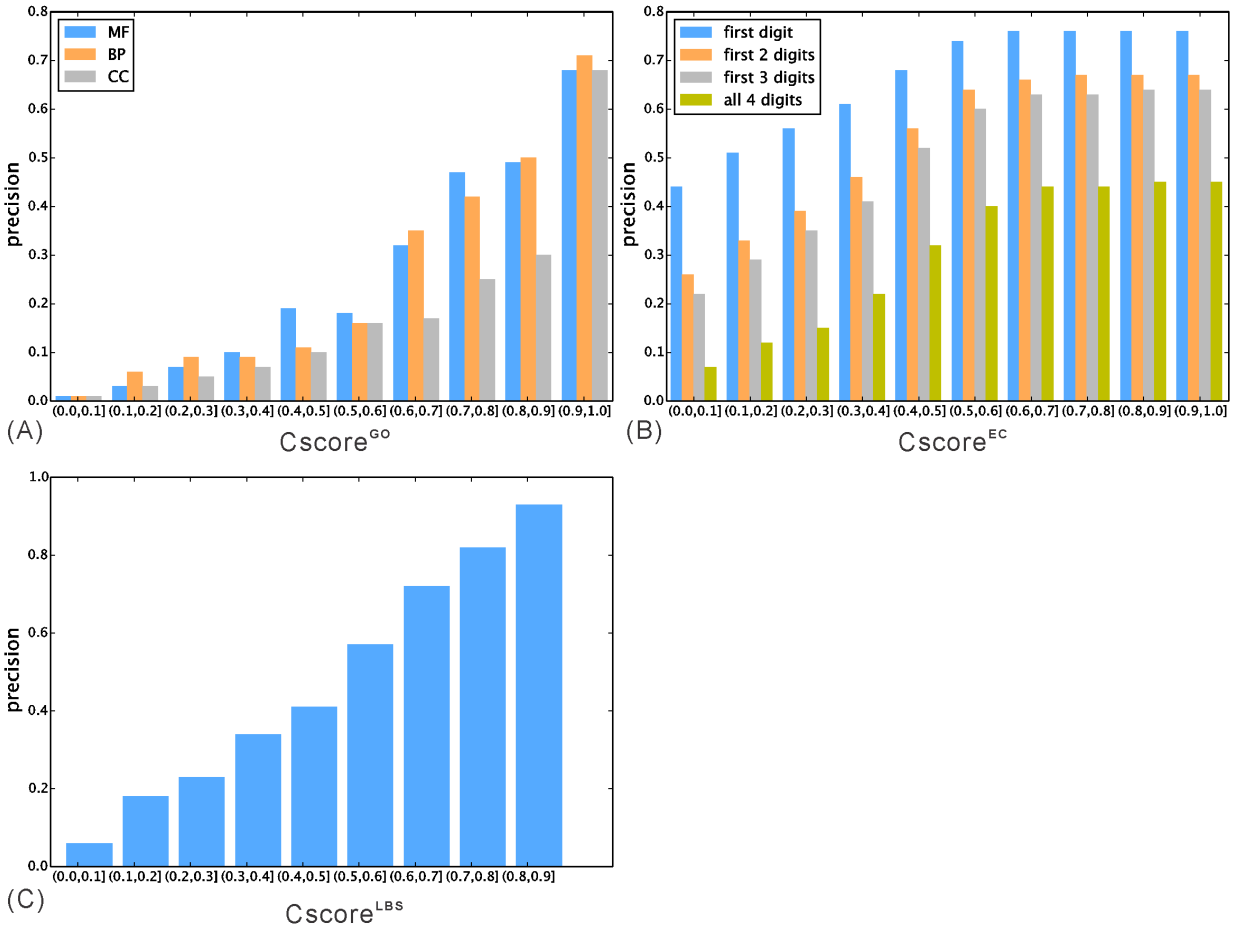


Figure 14. Precision of COFACTOR models versus the confidence score in each category of function annotation. (A) GO, (B) EC, and (C) Ligand binding sites.

As implemented in the CAFA experiments [268], we have used *localID* (i.e. the sequence identity normalized by number of aligned residues) as the confidence score for the "BLAST" and "PSI-BLAST" in the control studies. However, such a score might not be the best choice for these baseline methods. In Figure 15, we compare the performance of BLAST and PSI-BLAST using different confidence scores, including *localID*, *globalID* (sequence identity normalized by

the query length), *evalue* (lowest E-value), and *frequency* (the number of homologs annotated with a GO term of interest among all identified homologs). It shows that *frequency* consistently has the highest Fmax at all cutoffs through all GO aspects, indicating that a consensus of multiple template hits is probably a more robust indicator than the score of the best individual template. We therefore recommend the use of the frequency of (PSI-)BLAST hits as a more reliable and challenging baseline method in future assessment experiments. We also find that sequence identity is more indicative of GO annotation similarity than E-value, where both *globalID* and *localID* have consistently a higher Fmax than evalue. Based on these observations, in Equations (4.3) to (4.5), we have combined the homologous templates from both BLAST and PSI-BLAST, with the sequence identity as the weight of the combinations. The result shows that the combination outperforms the simple counting of the frequency in each individual program. Thus, even the poorly performing BLAST and PSI-BLAST methods may be substantially improved by careful consideration of the applied scoring schemes.

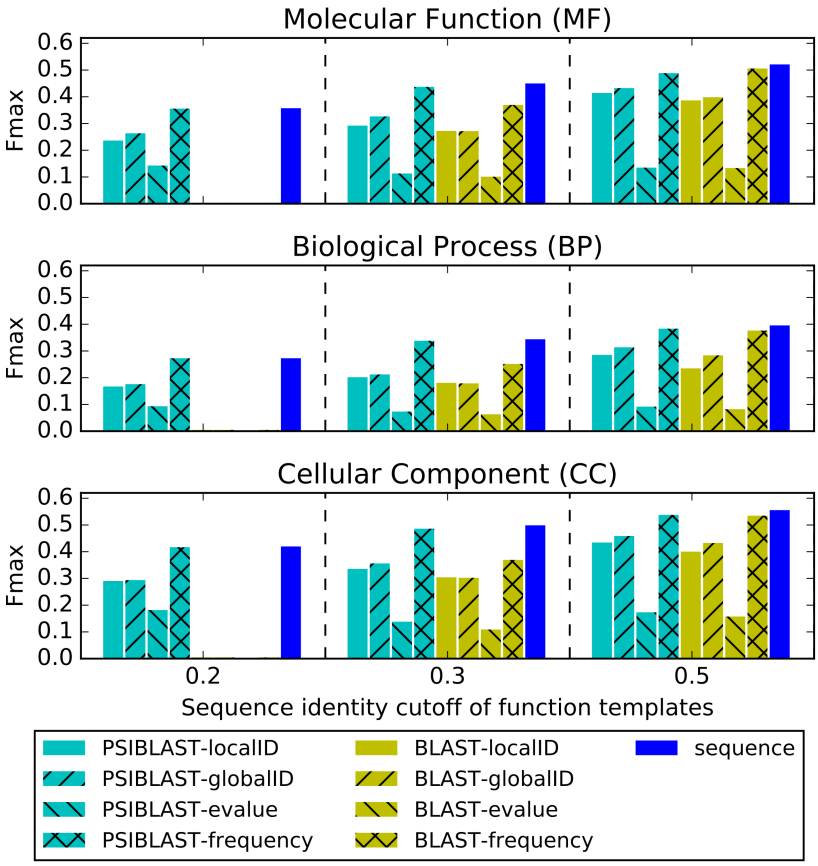


Figure 15. The Fmax score of the GO prediction by PSI-BLAST and BLAST using four different scoring functions (*localID*, *globalID*, *evalue*, and *frequency*) for selecting the functional templates. "*sequence*" indicates the sequence-based pipeline in COFACTOR, which combines the prediction results from PSI-BLAST and BLAST hits. This figure was generated using a separate dataset of 1000 CAFA3 targets non-redundant to the COFACTOR benchmark dataset.

4.3.2 Structure-based approach for EC number prediction

COFACTOR's ability to predict EC numbers was tested on a set of 318 non-homologous enzymes, with the benchmark EC numbers extracted from the PDB entries. The structural models are again predicted by I-TASSER, which are used for the EC template detection as in Equation (4.8). As with the GO term predictions above, to simulate a challenging case with no close sequence homologues available, both structural and function templates homologous to the query (with a sequence identity >30%) were excluded from the I-TASSER and COFACTOR template libraries. Figure 16 presents the benchmark results of COFACTOR on EC number

prediction compared with the BLAST and PSIBLAST baseline predictors at the same homology cut-off. The data shows a significant advantage of COFACTOR's use of structural homology transfers over the sequence-homology approach of BLAST and PSIBLAST. For example, the F-measure for the first three digits of EC number for the first template of COFACTOR is 0.702, while those for the BLAST and PSIBLAST baseline predictors are just 0.243 and 0.450, respectively.

Figure 14B shows a correlation of the precision of the EC models versus the confidence score ($Cscore^{EC}$), while a strong correlation with a PCC=0.95 is obtained between $Cscore^{EC}$ and the precision for the first enzyme homolog identified for each target. Generally, the precision of the prediction goes above 0.5 for any models with a $Cscore^{EC} > 0.4$ (Figure 14B).

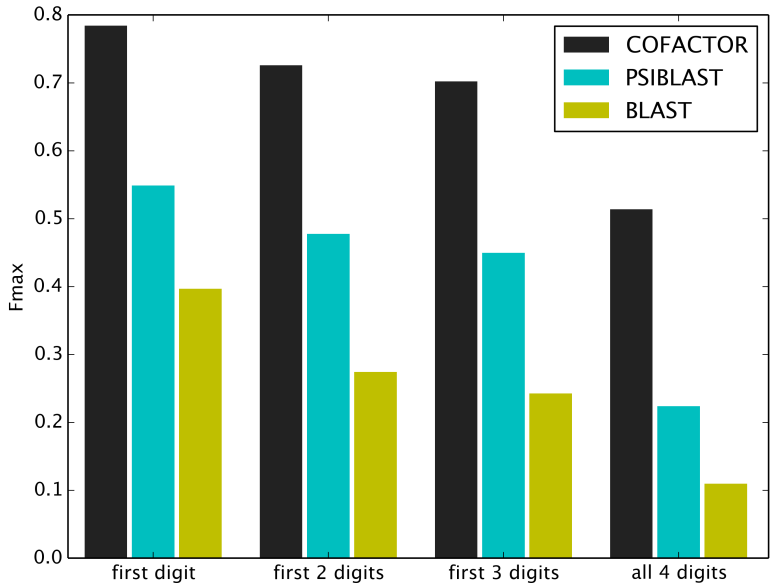


Figure 16. Accuracy of EC number prediction by COFACTOR and control methods at 30% sequence identity cut-off. Accuracy is evaluated by maximum F-measure. The BLAST and PSIBLAST baseline method is implemented as in Figure 13, but use the same EC library as COFACTOR.

4.3.3 Ligand binding site prediction

The performance of COFACTOR in ligand binding site prediction was benchmarked on 814 ligand-binding sites from 500 non-homologous proteins from the PDB. As in the tests above, both structural and functional templates with a sequence identity >30% have been excluded from the I-TASSER structure prediction and COFACTOR binding site template recognitions, to avoid homologous contamination. Although no homologous templates were used, COFACTOR identifies at least one binding residue correctly in 88% of the test proteins. The overall Matthews correlation coefficient (MCC, as defined in Figure 17) between actual and predicted binding sites is 0.465. This compares favourably to other state of the art binding site predictors including Concavity [269] and Findsite [270] which have overall MCCs of 0.378 and 0.454, respectively, for the same set of proteins

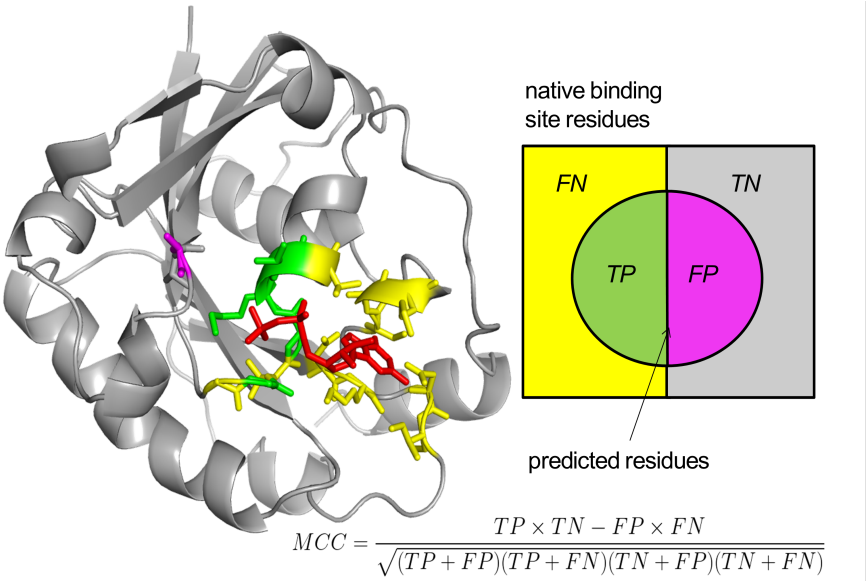


Figure 17. An illustrative example of ligand binding site prediction on the C-chain of the GDPRan-NTF2 complex (PDB ID: 1a2k). Red: native ligand GDP. Green: residues correctly predicted (TP). Magenta: residues incorrectly predicted (FP). Yellow: native binding site residues that are not predicted (FN). The prediction is evaluated by Mathews Correlation Coefficient (MCC).

In Figure 17, we show an illustrative example from the C-chain of the GDPRan-NTF2 complex (PDB ID: 1a2k), where 5 residues were predicted by COFACTOR as ligand-binding

sites and 4 of them were correct, resulting in an MCC=0.723 for this case. Figure 14C displays the precision values of COFACTOR binding predictions versus the confidence score ($Cscore^{LBS}$), which shows a strong correlation with PCC=0.99. According to the data, 62.6% of the binding sites are predicted correctly for the models with a $Cscore^{LBS}$>0.5.

4.3.4 Webserver

Server input. The mandatory input for the webserver is a single-chain protein structure file for the query protein in PDB format. If the input structure contains multiple chains or multiple models, only the first chain of the first model will be parsed. In the absence of an experimentally solved structure, the user can use models generated by the on-line structure prediction tools, such as I-TASSER [15,271], QUARK [272], Rosetta [33], HHpred [273] or Phyre2 [274]. If the user does not additionally specify the amino acid sequence, the sequence of query will be extracted from "SEQRES" records of PDB file, or "ATOM" records if "SEQRES" is absent. If the input PDB structure is not completely consistent with its corresponding biological sequence, the user is always recommended to provide the full sequence as additional input, so that the sequence-based components of COFACTOR can generate correct results.

Server output. Upon job completion, the user will be notified by email with a link to the result page in the COFACTOR server website. The result page consists of four major panels: structural analogies, GO terms, EC numbers, and ligand binding sites; an example is shown in Figure 18. The first panel displays the top ten analogous structures from the PDB library that are structurally closest to the query protein. The structural superimpositions are displayed in an interactive JSmol applet that allows users to rotate and annotate the pictures [275]. The analogous template is shown together with the TM-score, RMSD of aligned region, sequence identity, and

query coverage, and two links are given to the addresses for downloading the PDB template structure and the superposed query/template models from TM-align, respectively. By clicking on each of the radio buttons, user can explore the JSmol applet of all different templates (Figure 18A). The second panel shows the consensus GO prediction results, with models for the MF, BP, and CC aspects listed separately. The predicted GO terms are listed alongside the $Cscore^{GO}$ and their common name. For each of the three GO aspects, the predicted GO terms are plotted together with their parent terms as a directed acyclic graph, in which the predicted GO terms are highlighted by a $Cscore^{GO}$-specific color code, with blue to red representing the terms with $Cscore^{GO}$ from [0.4-0.5] to [0.9-1.0]. Since there are usually multiple terms predicted for each target, only the confident predictions with $Cscore^{GO}$ ≥0.5 are displayed, although the full set of predictions are available for download. If none of the GO terms has a $Cscore^{GO}$ ≥0.5, the GO terms with the highest $Cscore^{GO}$ will be displayed (Figure 18B). The third panel shows the top five EC number predictions, each associated with the template structure and marked with predicted active sites that can be visualized in the accompanying JSmol applet. In addition, the predicted EC number, the confidence score, TM-score between query and template, RMSD of aligned region, sequence identity, query coverage, and predicted active sites are also listed for each model (Figure 18C). The last panel shows the ligand binding site prediction results. For each set of binding sites, the structure templates are presented in order of descending confidence score, together with their TM-score, RMSD of aligned region, sequence identity, coverage, and binding site residues. The positions of the ligand binding site residues are highlighted in the target structure, and can be viewed and interpreted using the JSmol applet (Figure 18D). For every target protein, all the prediction results are packed in a tarball file named “result.tar.bz2” that can be conveniently downloaded from the output page.

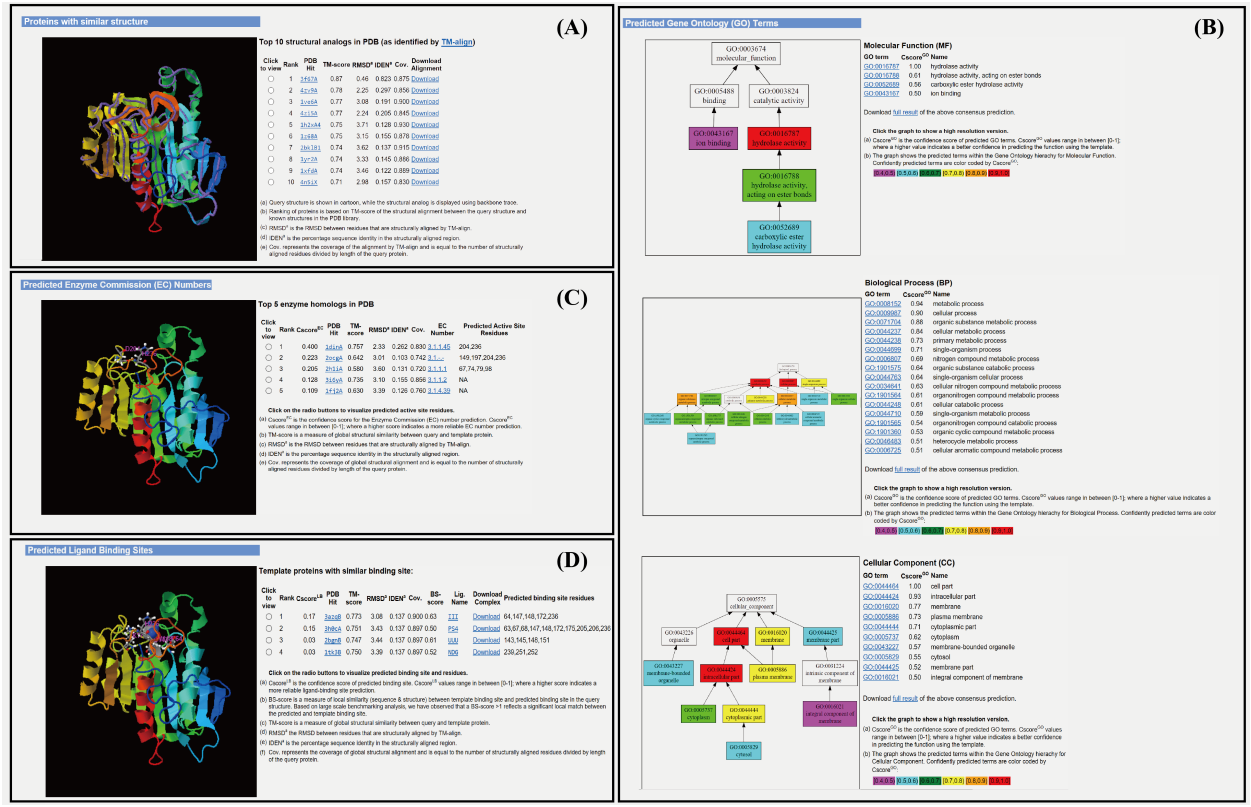


Figure 18. An illustration of the COFACTOR webserver output consisting of four annotation panels. The example is from the E. coli protein ysgA (UniProt accession: P56262) with a structural model generated by I-TASSER [16]. (A) Top ten analogous structures that are structurally closest to the query structure, displaying the structural similarity between YsgA and known hydrolases. (B) GO prediction results in three aspects of molecular function (MF), biological process (BP), and cellular component (CC), which are consistent with UniProt annotation of YsgA as a putative carboxymethylene butenolidase and EcoCyc [276] annotation as a predicted hydrolase. (C) EC prediction results from top-five enzyme homologous templates, suggesting carboxymethylene butenolide hydrolase activity (EC 3.1.1.45) and directly predicting the enzyme's active site. (D) Ligand-binding site prediction results from the top ten homologous templates, including residues surrounding putative active sites that are in proximity to the ligand.

4.4 Discussion and Conclusion

We report recent advancements made to the on-line COFACTOR server for hybrid protein function annotations. In general, the biological function of a protein can be intricate and often contains multiple levels of categorizations. The COFACTOR server focuses on the three most widely-used and computationally amenable categories of function: gene ontology, enzyme commission, and ligand-binding sites. Compared with the previous version of COFACTOR, which generated function annotations mainly based on structural homology transfer, the updated

server introduced several new pipelines built on sequence profile and protein-protein interaction network information to enhance the accuracy and coverage of the structure-based function predictions. Accordingly, several new sources of function templates, including sequence and PPI function terms, have been incorporated into the default function library (BioLiP) of the COFACTOR server. Our large-scale benchmark tests have shown that the new composite pipelines can generate function predictions with accuracy outperforming many state of the art methods in the literature.

To facilitate the use and interpretation of the prediction results, a confidence scoring system has been introduced, which can help user to quantitatively estimate the accuracy of the predictions. Meanwhile, new directed acyclic graphs combined with animation software are introduced to facilitate the viewing, analysis and manipulation of the prediction models. These developments and updates will significantly enhance the accuracy and usability of an already widely applied structure function service system, and make it continue to be a powerful tool both for rapid annotation of uncharacterized proteins and for providing a starting point to understand and further characterize targets that may be identified in high-throughput experimental studies.

Chapter 5 Structure-based Annotation of uPE1 Proteins in Human Proteome[1]

5.1 Introduction

As the direct carriers of biological functions in the human body, proteins participate in nearly all biological events, including catalysis of endogenous metabolites, regulation of most biological pathways, and formation of many subcellular structures. Understanding the function of human proteins has become an important prerequisite to uncover the secrets of human diseases and diverse phenotypes in modern biomedical studies. As a protein usually must be folded into specific tertiary structure in order to be functionally active, determining protein structure is an important avenue in protein function annotation.

Despite many years of community efforts in protein characterization, there is still a substantial number of proteins whose structure and biological functions are incomplete or unknown. Among all the 17470 confidently identified (PE1) human proteins in the neXtProt[277] release 2018-01-17, there are 1260 uPE1 entries which do not have specific functional annotation. In the same neXtProt release, there are 6188 out of 17470 PE1 entries with experimental 3D structures but only 32 among the 1260 uPE1 proteins. The lack of structure and function annotations for many proteins in the human proteome limits our capability to understand their functional roles even in tissues with high expression. For example, of the 26 uPE1 proteins on chromosome 17 with immunohistochemistry data in Human Protein Atlas[278]

[1] This chapter was adapted from two previously-published works. The first work was C Zhang, X Wei, GS Omenn, and Y Zhang (2018) "Structure and protein interaction-based Gene Ontology annotations reveal likely functions of uncharacterized proteins on human chromosome 17" *Journal of Proteome Research*, 17(12), 4186-4196. The second work was C Zhang, L Lane, GS Omenn, and Y Zhang (2019) "Blinded testing of function annotation for uPE1 proteins by the I-TASSER/COFACTOR pipeline using the 2018-2019 additions to neXtProt and the CAFA3 challenge." *Journal of Proteome Research*, 18(12), 4154-4166.

(retrieved on 2018-05-09), 24 have "high" expression in at least one tissue as detected by antibody studies. Similarly, 52 of the 66 uPE1 proteins on chromosome 17 (as of neXtProt 2017-08-01) have median RNA expression levels higher than 10 Transcripts per Million (TPM) in at least one tissue, as reported in GTEx[279] version 7.

To alleviate the issue in protein structure and function annotations, we developed a hybrid pipeline which creates 3D structure prediction using I-TASSER[280], with the functional insights deduced by COFACTOR[281]. Both I-TASSER and COFACTOR pipelines have been tested in community-wide blinded experiments, which demonstrate considerable reliability of structure modeling and functional annotations. For example, in CASP12, for 53 targets with template structures identified in PDB, I-TASSER generated correct folds with a TM-score >0.5 for 47 cases, where in 41 cases structures were driven closer to the native than the templates. For 39 free-modeling (FM) targets which do not have any similar fold in the PDB database, 11 were correctly folded by I-TASSER.[190] In CASP9, the COFACTOR algorithm[282] achieved a functional residue prediction precision of 72% and Matthews correlation coefficient 0.69 for the 31 function prediction targets, which were higher than those by all other methods in the experiment.[282]

The original version of COFACTOR[251] was built on the transfer of function from structural templates detected by homologous and analogous structure alignments. That version of COFACTOR was used to suggest structure and function for dubious proteins in the human proteome (PE5).[283] Recently, COFACTOR was extended with additional sequence and Protein-Protein Interaction (PPI) pipelines, which was tested in the most recent CAFA3 function annotation experiment.[281,284] According to the CAFA3 evaluation (https://www.synapse.org/#!Synapse:syn12299467) for GO term prediction in MF, BP, and CC

aspects, COFACTOR achieved F1-scores (defined in Equation 1 below) 0.57, 0.60, and 0.61, respectively, which are 43%, 81%, and 17% higher in accuracy than the best baseline methods used by assessors. Additionally, we have used the I-TASSER/COFACTOR pipeline for proteome-wide structure and function modeling of *E. coli* proteins, and the predicted functions of three proteins have been validated by enzymatic assay and mutation experiments.[285]

In light of recent progress, we applied this pipeline to better annotate the human proteome as part of the HUPO Chromosome-centric Human Proteome Project (C-HPP).[286] As a proof-of-principle study, we applied the I-TASSER/COFACTOR pipeline to all 66 uPE1 proteins from human chromosome 17 in neXtProt 2017-08-01 release to decipher the structure and function of these poorly annotated human proteins. Additionally, to rigorously benchmark the performance of our pipeline, we performed two rigorous time-elapsed blind tests of protein function prediction. In the first blind test, we evaluated the performance of COFACTOR in the CAFA3 GO term prediction challenge. On the 267 and 912 CAFA3 human targets used for MF and BP evaluations, respectively, we found that a clear advantage of COFACTOR compared to simple sequence homology search or background probability modeling, though its performance is still dependent on the availability of high scoring templates. In the second blind test, an independent assessor (co-author L.L.) identified a set of 44 neXtProt[287] proteins undergoing function curation based on manually gathered publications during 2018[288]. Meanwhile, predictors (co-authors C.Z., G.S.O., and Y.Z.) performed protein structure and function predictions using the same automated pipeline as in our 2018 study. Based on the automatically predicted GO terms, these three predictors assign a free-text function interpretation for each query protein. Both the automated GO predictions and the manual free-text interpretations were performed blind to the pending curation of the proteins and were submitted to the assessor before

the neXtProt 2019-01 release. For both predicted GO terms and the respective free-text interpretation, consistency of the predicted functions with neXtProt curation was assessed upon the publication of neXtProt 2019-01 release. These analyses should serve as an incremental step towards completion of structure and function modeling of all remaining uPE1 and even PE2,3,4 proteins in the human proteome[289]. The full blinded testing dataset is available at https://zhanglab.ccmb.med.umich.edu/COFACTOR2/nx2019addition/GOterm.html, while our predicted functions for the chromosome 17 uPE1 proteins are available at https://zhanglab.ccmb.med.umich.edu/COFACTOR/chr17/. Additionally, the structure and function modeling results for all targets modeled in this study are provided as a link on neXtProt (https://www.nextprot.org/entry/NX_P0C870/gh/zhanglabs/COFACTOR, where "NX_P0C870" can be replaced by neXtProt ID for each other target of interest).

5.2 Methods

5.2.1 Protein structure and function prediction pipelines

Our computational workflow for structure-based function annotation of a given protein consists of two main components: structure modeling by I-TASSER and function annotation by COFACTOR (Figure 19). The pipeline is fully automated with the query sequence as the sole input.

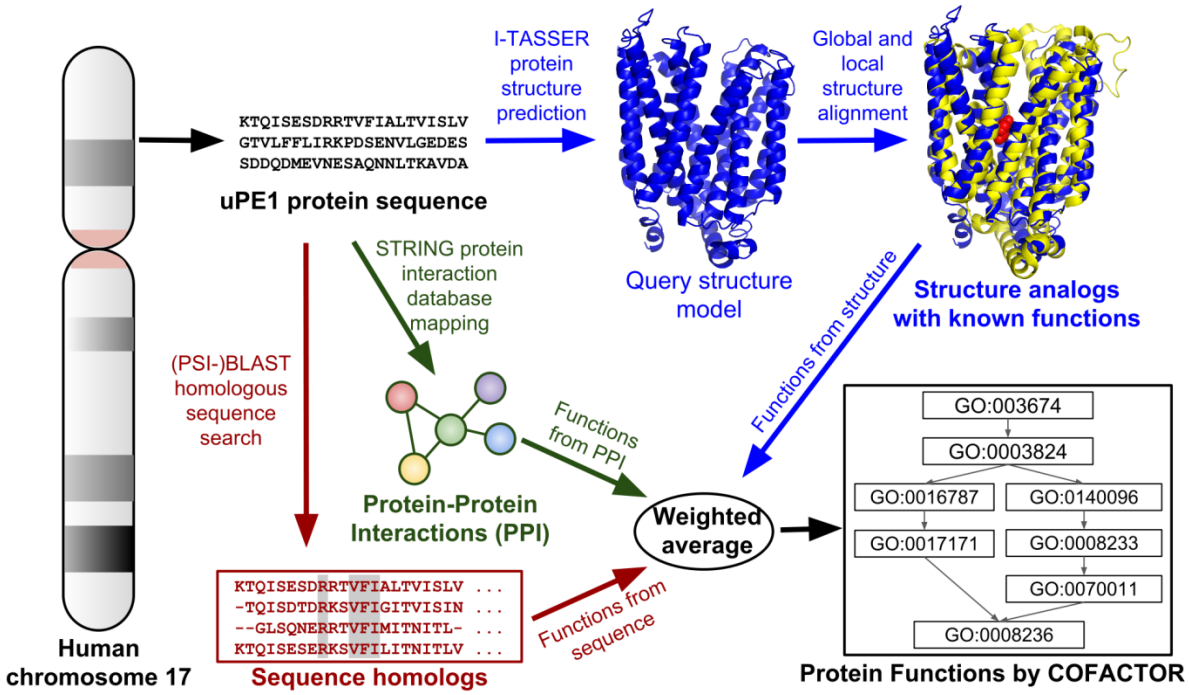


Figure 19. Flowchart of the automated I-TASSER/COFACTOR pipeline for protein structure and function prediction, applied to uPE1 proteins from human chromosome 17.

I-TASSER protein structure prediction

In the I-TASSER structure prediction stage, the query protein sequence is first threaded through a non-redundant PDB library (https://zhanglab.ccmb.med.umich.edu/library/) by LOMETS[290], which is a locally-installed meta threading algorithm combining 10 different state-of-the-art threading programs[99,100,291-297], to identify structure templates. Continuous fragments are excised from these template structures, which are subsequently assembled into full length structure by replica-exchange Monte Carlo (REMC) simulation implemented by I-TASSER. Tens of thousands of decoy conformations from the REMC simulation trajectory are then clustered by SPICKER[187] by structure similarity. The centroid of the largest cluster, which corresponds to the conformation with lowest free energy, is selected to undergo structure refinement by FG-MD[188] to obtain the final structure model. While I-TASSER typically reports up to five structure models, ranked in descending order of the size of cluster from which a model

came, we use only the first I-TASSER model for subsequent function modeling. That is because the first model has the highest confidence score and on average is closer to native structure than the lower-ranked models.[298]

COFACTOR automatic structure-based function annotation

To obtain function annotation for the query structure model, the COFACTOR structure-based function prediction approach uses a modified TM-align[227] structure alignment program to search the query structure against entries templates from the BioLiP[257] structure-function database to identify structure templates with function annotations. The functions of structure templates are then transferred to query according to global structure similarity, active site local similarity and matching of sequence profiles between query and template, as measured by a combination of global and local structure alignments[284]. The combination of global and local structure similarity is critical to structure-based function annotation, as shown previously.[284] If only global similarity is considered, the annotation result can be misled by fold promiscuity, where proteins sharing highly similar global topology can have very different functions.[197] On the other hand, relying only on active site local structure similarity can also lead to false positive hits: ligand binding pockets with similar conformation can be associated with unrelated biochemical functions due to the very limited number of possible pocket structures.[299] To further disentangle the structure promiscuity issue, the above structure-based function annotation is supplemented by the sequence-based approach, which extracts function annotations from BLAST and PSI-BLAST[102] hits in the UniProt[300] database search. Meanwhile, the protein-protein interaction (PPI) based approach infers function from UniProt sequences homologous to the query's PPI partners, as defined by the STRING[262] database. In the earlier version of COFACTOR[201], the functions are inferred from GO terms annotated to the PPI partners. When

we later developed MetaGO[284], an extension for the GO term prediction component of COFACTOR, we found that functions inferred from UniProt[300] sequences homologous to the PPI partners are more accurate than functions directly inferred from PPI partners. Therefore, the current COFACTOR program uses this improved PPI based method originally developed for MetaGO, where functions are predicted from UniProt sequences homologous to PPI partners of query. For a given GO term q, the confidence of final consensus prediction ranges between 0 and 1, and is a weighted average of the three approaches (structure, PPI, and sequence):

$$Cscore(q) = 1 - \prod_{m\epsilon\{structure,PPI,sequence\}} [1 - Cscore_m(q)]^{w_m} \quad (5.1)$$

Here, w_m is the weighting score for method m. $Cscore_m(q)$ is the confidence score of the mth method for GO term q and takes the following form:

$$Cscore_m(q) = \frac{\sum_{i=1}^{N^m(q)} S_i^m(q)}{\sum_{i=1}^{N^m} S_i^m} \quad (5.2)$$

N^m is the total number of templates detected by method m. S_i^m is the weighting score of the ith template detected by method m. The template weighting score could be (PSI-)BLAST sequence identity for sequence based method, and interaction score assigned by STRING database for PPI based method. $N^m(q)$ and $S_i^m(q)$ are the template number and weighting score of ith template, respectively, in method m for the subset of templates associated with GO term q. Instead of using the most confident template for each GO term, Equation 5.2 represents a weighted k-nearest-neighbor approach where all N templates are considered in the consensus voting for each predicted GO term. Therefore, if all templates are associated with q, the nominator and denominator in Equation 5.2 are the same, and $Cscore(q)$ is one, i.e. 100% confident, even when none of the templates share high sequence similarity to the query.

5.2.2 Manual free-text function interpretation

We follow three steps to assign free-text annotation for automated GO term prediction:

(a) Examine MF and BP GO terms from I-TASSER/COFACTOR, excluding general terms, either those defined in neXtProt SPARQL NXQ_00022 or terms like "cellular process".

(b) Select the most specific GO term in MF or BP with C-score>0.5. If there is no GO term with C-score>0.5, consider terms with C-score>0.4.

(c) If the aspect (MF/BP) with the term selected in step (b) also has other high confidence unrelated GO terms, proceed to the complementary aspect (BP/MF) and repeat step (b).

For example, even though the BP prediction for C1QTNF8 (P60827-1) includes multiple terms with C-score ≥0.5 (GO:0009987 “cellular process”, C-score=0.93; GO:0048518 “positive regulation of biological process”, C-score=0.67; GO:0032502 “developmental process”, C-score=0.65; GO:0044238 “primary metabolic process”, C-score=0.54; GO:0048584 “positive regulation of response to stimulus”, C-score=0.53), these GO terms are not informative for the purpose of free-text function interpretation because they only vaguely suggest the protein’s involvement in biological regulation of unspecified pathways. Meanwhile, this protein does not have any MF GO term predicted with C-score ≥0.5 (after excluding the GO terms considered by neXtProt as too general). Therefore, for this protein, we alternatively use MF GO terms predicted with C-score ≥0.4 (GO:0005102 “signaling receptor binding”, C-score=0.41) and assign the free-text interpretation “signal receptor binding”.

In the event that predicted GO terms are too diverse in one of the three GO aspects to conclusively interpret the function, other aspects of GO are used for interpretation. For example, BP prediction of RFPL1 (O75677-1) is too diverse (GO:0016567 “protein ubiquitination”, C-score=0.55; GO:0010468 “regulation of gene expression”, C-score=0.56; and GO:0002376

"immune system process", C-score=0.63); we instead used its high confidence MF GO term predictions, which are exclusively related to ubiquitin-protein transferase activity (GO:0004842, C-score=0.78).

All free-text interpretations strictly use phrases in the definitions of predicted GO terms selected by the above criteria.

This exercise of our free-text annotation was performed to emulate how biologists would interpret a list of computationally predicted GO terms for a protein. It is only performed for the small neXtProt dataset of 44 proteins, because manual inspection of the full CAFA3 dataset with 20 197 human proteins was impractical. As exemplified by REPL1 (O75677-1) above, to simplify our interpretation, the free-text annotation derived from predicted GO terms only attempted to cover the most likely function of a protein. Therefore, such a free-text annotation may not be as comprehensive as the respective UniProt/neXtProt free-text annotation, which aims to cover as many different functions of a protein as possible so long as there is conclusive literature evidence. This difference in how our free-text annotations and those of UniProt/neXtProt are derived also affects how we evaluate the performance of our free-text annotations, as discussed later.

5.2.3 Assessment metrics for function prediction

Biologically meaningful metrics for assessing protein function prediction should not focus only on the precision of predicted GO terms. For example, a protein function predictor that only predicts shallow and generic GO terms such as "protein binding" or "cellular process" could have a very good precision but is rarely useful in practice. In fact, neXtProt does not consider 11 MF and 2 BP GO terms for being too general and does not use Cellular Component (CC) at all

when retrieving uPE1 proteins (https://www.nextprot.org/proteins/search?mode=advanced&queryId=NXQ_00022). The 11 general MF terms are GO:0005524 "ATP binding", GO:0000287 "magnesium ion binding", GO:0005515 "protein binding", GO:0042802 "identical protein binding", GO:0008270 "zinc ion binding", GO:0051260 "protein homooligomerization", GO:0005509 "calcium ion binding", GO:0003676 "nucleic acid binding", GO:0003824 "catalytic activity", GO:0046914 "transition metal ion binding", and GO:0046872 "metal ion binding"; the 2 general BP terms are GO:0007165 "signal transduction", and GO:0035556 "intracellular signal transduction". We have accepted those exclusions in this analysis of neXtProt data. To simultaneously assess the precision and recall of our prediction, we follow the standard practice of CAFA and evaluate the accuracy of automatic GO term prediction by maximum F1-score, i.e., the Fmax.

$$Fmax = \max_{t\in(0,1]} \left\{ \frac{2 \cdot pr(t) \cdot re(t)}{pr(t) + re(t)} \right\} \qquad (5.3)$$

$$pr(t) = \frac{tp(t)}{tp(t) + fp(t)}, re(t) = \frac{tp(t)}{tp(t) + fn(t)} \qquad (5.4)$$

In the above equations, $pr(t)$, or "precision", is the number of correctly predicted GO terms, true positive $tp(t)$, over the number of all GO terms predicted with confidence score $\geq t$, i.e., $tp(t) + fp(t)$. $re(t)$, or "recall", is $tp(t)$ divided by all true positive plus false negative GO terms annotated to query by UniProt/neXtProt ground truth, i.e., $tp(t) + fn(t)$. The concept of Fmax is illustrated in Figure 20.

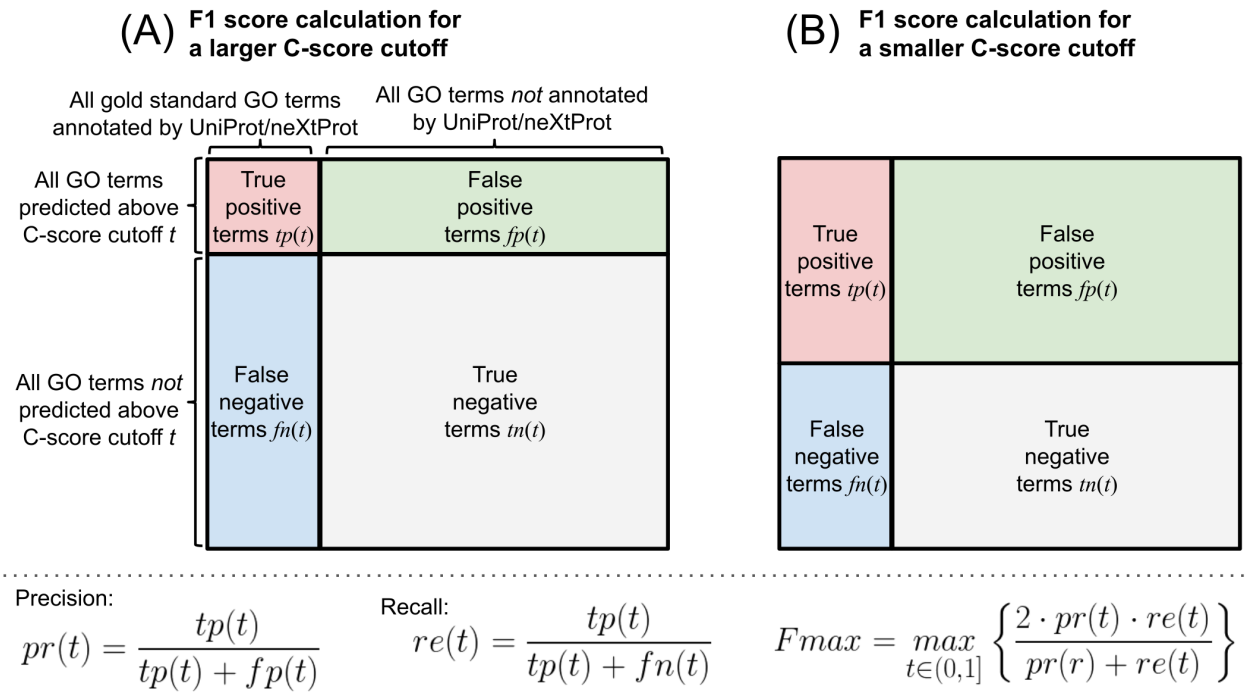


Figure 20. (A) Graphic explanation of Fmax, the standard metric for evaluating the overlap between of the set of predicted GO terms (the two red and green rectangles on the top) and the set of ground-truth GO terms (usually being experimental annotations in UniProt/neXtProt, the two red and cyan rectangles on the left). The big square represents all possible GO terms. Precision is the portion of predicted GO terms that are correct (the set of ground-truth GO terms), and recall is the portion of ground-truth standard terms that are predicted. (B) For the same protein, the set of "predicted" GO terms depends on the C-score cutoff t ranging between 0 and 1, and less stringent cutoff (smaller t value) results in larger set of predicted terms (bigger area for the two rectangles on the top), which makes both precision and recall dependent on the C-score cutoff t as well. The harmonic average of precision and recall is called F1 score, whose maximum over the entire range of t is Fmax.

Two further clarifications should be made for the Fmax, as a measure of consistency between our prediction and the UniProt/neXtProt GO annotation. First, although Fmax=0.5 means half of the predicted GO terms exactly match GO terms annotated by UniProt/neXtProt, and half of the UniProt/neXtProt GO terms are among the predicted GO terms, a predictor achieving Fmax=0.5 should not be interpreted as being no better than guessing the two faces of a flipped coin. Unlike a flipped coin whose probability for the landing of the two faces are half-half, the average probability for a GO term to get annotated (or not annotated) to a protein is far from half-half in the database: for 92.1%, 98.5% and 99.8% of the 47 340 GO terms defined by

the Gene Ontology Consortium, each of them is annotated to less than 0.1%, 1%, and 10% of any UniProt proteins, respectively. Therefore, predicting GO terms with 50% precision is indeed a challenging task and a significant success.

Second, Fmax should not be confused with C-score. The C-score is for each predicted GO term of a query protein, while Fmax is an overall statistic for a protein or a set of proteins. C-score is estimated by COFACTOR without knowing the ground-truth, while Fmax can only be calculated if we know both the predicted GO terms and the ground-truth GO terms.

Compared to GO term evaluation, assessment of free-text annotation is more challenging as there is no agreed-upon metric to quantify the similarity between two free-text biological function descriptions. Moreover, free-text function annotations for a protein, especially one that performs multiple functions or is involved in complicated pathways, are affected by subjective judgment by the function curators for both our predictions and by UniProt/neXtProt curators. In this blinded analysis, we compared both GO terms and free-text interpretation from I-TASSER/COFACTOR prediction and from the UniProt/neXtProt literature-based free-text curation. Another complication for head-to-head comparison between the two kinds of free-text annotation is that, as mentioned above, our free-text interpretation from I-TASSER/COFACTOR only attempts to cover the most likely function of the target protein, while UniProt/neXtProt free-text annotation attempts to more comprehensively cover different functions of a protein. Therefore, if free-text interpretation from I-TASSER/COFACTOR matches at least part of the neXtProt free-text annotation for a target protein by manual inspection, we consider the pair of free-text annotations is consistent (see Table 1 at https://zhanglab.ccmb.med.umich.edu/COFACTOR2/nx2019addition/GOterm.html).

While both free-text and MF/BP GO terms are considered "function annotations", free-text annotations curated by UniProt/neXtProt may not be fully reflected by the GO terms annotated for the same protein, partly due to the complexity of data source and curation process. neXtProt function annotations have the following major sources. First, all manual annotations performed by Swiss-Prot curators from experimental papers are generally captured as free-text, MF/BP GO terms (using the closest possible terms), keywords and, in the case of enzymes, Enzyme Commission (EC) numbers. Sometimes there is no existing GO term to describe a particular function, resulting in only a free text description without GO terms, which happens to 17 and 3 for MF and BP, respectively, for the 25 blindly-tested neXtProt targets. In most cases, GO terms assigned in this way are more generic than the respective free-text. Secondly, MF and BP GO terms are also manually annotated by other members of the Gene Ontology Consortium, such as HGNC and MGI. Finally, MF and BP annotations computationally assigned by UniProt or the Gene Ontology Consortium are considered. Apart from free-text and GO terms, neXtProt includes other function annotations such as pathway annotations from KEGG and Reactome, and transporter classification from TCDB. For this paper, we mainly focus on GO terms and free-text.

5.3 Results

5.3.1 Datasets

The 66 uPE1 proteins from chromosome 17 were compiled from neXtProt release 2017-08-01. The detailed protocol for generating this list is specified in supplementary Text S2. While most of these uPE1 proteins do not have any GO term annotations for MF and BP, some of them have GO terms that are considered too generic by neXtProt to be qualified as "annotated"

proteins, including protein binding, calcium binding, zinc binding, identical protein binding, and protein homooligomerization. As neXtProt does not consider GO CC terms when defining uPE1 proteins in the SPARQL query, some of these uPE1 proteins do have GO CC term annotations. For example, SYNGR2 (neXtProt ID: NX_O43760-1) is annotated as being located at "neuromuscular junction" (GO:0031594) and at "synaptic vesicle membrane" (GO:0030672) for CC based on its known role in modulating the localization of synaptophysin into synaptic-like microvesicles.[301,302] Due to this known bias in how neXtProt treats GO CC terms for uPE1 proteins, we later discuss instances where our CC term prediction is different from existing neXtProt annotations. The numbers of uPE1 proteins are "moving targets" due to new experimental evidence as well as evolving criteria reflected in excluded MF and BP terms. Thus neXtProt release 2017-08-01, which this study was based on, had 1218 uPE1 proteins proteome-wide and 66 uPE1 chromosome 17 proteins; neXtProt release 2018-01-17 has 1260 and 70, respectively.

This study additionally used three datasets: one benchmark neXtProt dataset for recalibrating the C-score of COFACTOR, and two time-elapsed blindly-tested datasets from CAFA3 human targets and newly annotated PE1 entries from neXtProt 2019-01. The "recalibration" set is used to establish the relation between C-score and function prediction precision in COFACTOR, while the performance of I-TASSER/COFACTOR is evaluated on the two blindly-tested datasets.

The recalibration set consists of 1 995 well-annotated human PE1 proteins with up to 750 residues in neXtProt release 2019-01. Similar to the benchmark set of 100 Chromosome 17 PE1 proteins in our 2018 report[303], each protein in this recalibration set has ≥ 3 Gold MF terms, ≥3 Gold BP terms, and ≥3 Gold CC terms.

The blindly-tested CAFA3 human dataset included 20 197 human protein targets, among which 267, 912, and 347 targets acquired new GO terms in UniProt between 2017-02-02 and 2017-11-15 for MF, BP, and CC, respectively, which is listed as part of the supplementary data for the CAFA3 report[304] (See supplementary_data/cafa3/benchmark20171175.tar at https://figshare.com/articles/Supplementary_data/8135393). These targets are further divided into two types: 147, 240, and 214 "No Knowledge" targets do not have any experimental GO annotation before CAFA3 for MF, BP, and CC, respectively; 120, 672, and 133 "Limited Knowledge" targets have at least one experimental GO annotation before CAFA3 for MF, BP, and CC. Statistical analysis of function predictions on this dataset, released by the CAFA Consortium in May 2019[304], is evaluated based on the GO term predictions our group submitted during CAFA3 challenge before 2017-02-02.

As of the neXtProt release 2019-01, 25 of the 44 proteins submitted for curation in 2018 acquired new function annotations. While all 25 targets receive free-text function annotation in neXtProt, only 8 and 22 acquired GO terms for MF and BP, respectively, excluding GO annotations deemed by neXtProt as being too general in the neXtProt SPARQL query NXQ_00022 as explained above in Methods. We make available our predictions for all 44 (https://zhanglab.ccmb.med.umich.edu/COFACTOR/nx2019addition/GOterm.html#3), so that comparison with future neXtProt releases will be facilitated. Among these 25 recently curated neXtProt targets, the function annotation for one target (P0C870-1, https://www.uniprot.org/uniprot/P0C870?version=85&version=87&diff=true) was updated by UniProt on 2019-02-13, and was not in time to be included in neXtProt release 2019-01. For this particular target, we use the more recent UniProt annotation on 2019-02-13 instead of that from neXtProt 2019-01. We do not separately evaluate our result on Gold GO terms and on Gold plus

Silver GO terms as in previous study[303], because all newly annotated MF and BP GO terms for this set of 8 and 22 targets have Gold status.

5.3.2 Recalibration of COFACTOR C-score for human proteins

While COFACTOR assigns a C-score for each predicted GO term for a target protein, the C-score is strongly correlated with, but does not strictly equal, the probability of the GO term being associated with the target. When we originally reported the GO term prediction method of COFACTOR, we calibrated this C-score to the corresponding probability of GO term association, i.e. the precision of GO term prediction, given the C-score, on a prokaryotic set of 1244 *E. coli* proteins[201]. Due to the later improvement of our function prediction method[284] and the change of species of interest (*E. coli* to human), it became necessary to recalibrate the current COFACTOR algorithm for the recalibration set of 1995 human proteins. To calculate GO term prediction precision given C-score, all GO term predictions for each of the three aspects (MF, BP, and CC) were grouped into 10 bins by C-score with bin width =0.1. To examine whether the calibration depends on the availability of close homology templates, we performed two separate calibration runs by excluding function templates sharing ≥0.3 and ≥0.9 sequence identity (ID) to the query. The calibration curve and the precision-recall curve are shown in Figure 21. For the PPI-based pipeline in COFACTOR, there are two rounds of sequence homolog search: the first round maps query sequence to its STRING entry, which is used to identify PPI partners interacting with the query; and the second round of sequence search maps PPI partners to UniProt proteins to obtain the GO annotations. Because the function annotations in the PPI-based pipeline are eventually derived only from the PPI partner homologs in UniProt, the sequence

identity cutoffs in Figure 21 are applied only between query and the PPI partner homologs in UniProt.

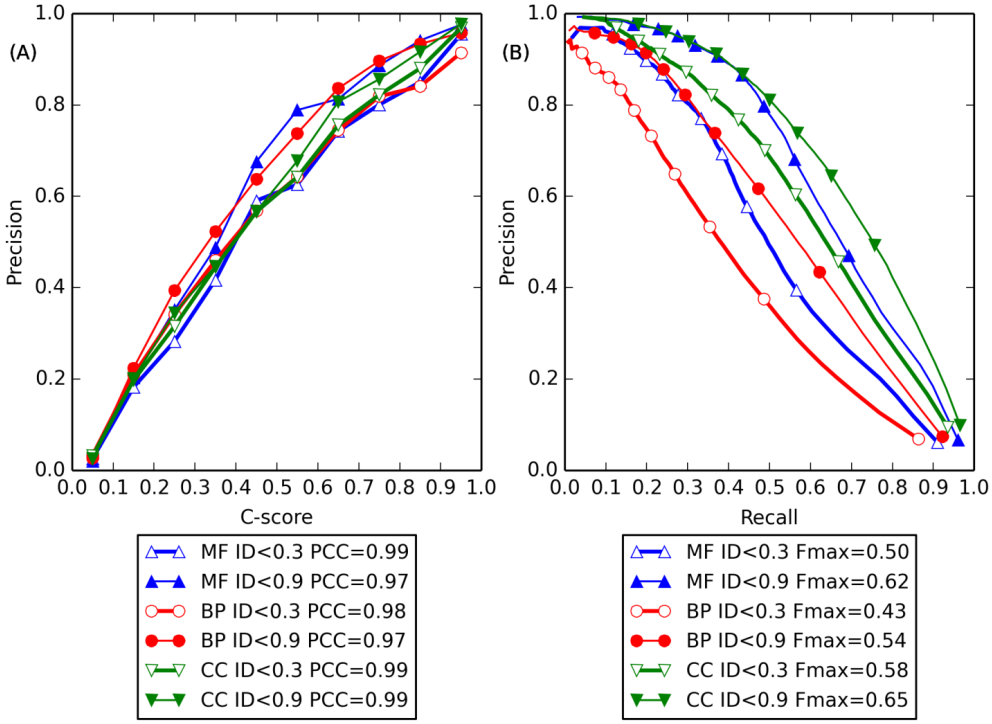


Figure 21. (A) Calibration curve for GO term prediction precision versus C-score for MF (blue up-pointing triangles), BP (red circles), and CC (green down-pointing triangles). Curve for template sequence identity (ID) <0.3 and <0.9 to the query are shown in hollow and solid markers, respectively. The lower legend shows the Pearson's correlation coefficient (PCC) between precision and C-score. (B) Precision-recall curve for GO term prediction. The lower legend shows the Fmax for each curve.

As shown in Figure 21B, GO term prediction accuracy of COFACTOR, as measured by Fmax, is higher by 24%, 28% and 11% when high sequence identity (ID<0.9) templates are available, compared to low sequence identity cutoff (ID<0.3) cases. Nevertheless, the values are still quite high for the lower cutoff. On the other hand, the correlation between precision and C-score does not have as strong dependency on sequence identity cutoff, even though the low sequence identity cases still have slightly lower precision given the same C-score. For example, for 0.4< C-score ≤0.5, the precision is 0.69, 0.64, and 0.57 for MF, BP, and CC for ID<0.9, compared with 0.60, 0.57, and 0.57 for ID<0.3 (Figure 21A). Considering the fact that most

poorly characterized or uncharacterized proteins have sequence identity around 0.3 to the closest functionally characterized homolog (see our earlier methodology paper[284]), we recommend the use of the recalibration curve obtained at ID<0.3 for interpretation of COFACTOR function prediction for human targets.

To determine reasonable GO term prediction confidence (C-score) cutoffs in the I-TASSER/COFACTOR pipeline, we show in Figure 22 the relation between C-score and prediction accuracy (F-measure). The highest F-measures for MF, BP, and CC are achieved when we choose C-score cutoffs > 0.59, > 0.55, and > 0.56, respectively.

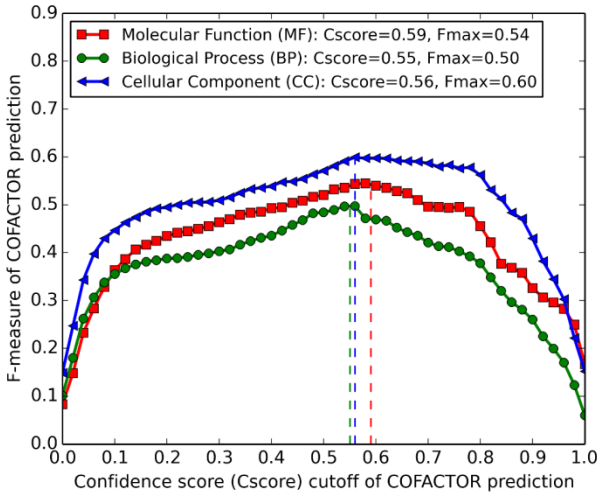


Figure 22. F-measures of COFACTOR prediction versus confidence score cutoffs for the three aspects of GO terms. From left to right, the three vertical dashed lines indicate C-scores 0.55 (green), 0.56 (blue), and 0.59 (red) which are C-score cutoffs corresponding to the highest F-measure for BP, CC, and MF, respectively.

5.3.3 Performance of GO term prediction by COFACTOR in CAFA3

A preliminary version of COFACTOR[201] was used in CAFA3, the latest CAFA community-wide challenge for protein function prediction, by team "Zhang-Freddolino lab". In the recently released official CAFA3 result[304], our team was ranked as one of the top performing groups (ranked second, third, fourth, and fifth for prediction of motility, biofilm formation, CC, and BP, respectively, but not within top ten for MF) among 68 teams (See Figure 3 and Figure 4

of CAFA3 report[304]). Such performance was obtained by a partial implementation of COFACTOR with just the sequence- and PPI-based pipeline for 82 903 (63.4%) of the 130 827 prediction targets from 23 species, as the structure-based pipeline of COFACTOR was not ready for high-throughput prediction when we participated in CAFA3 in 2017[201]. To further save time, among the reduced CAFA3 set of 47 924 structure-based function prediction targets, the full length structure models of query proteins were generated by LOMETS threading followed by MODELLER[305] homology modeling for 43 953 targets (91.7%) while the full I-TASSER pipeline was only used for the remaining 3971 targets (8.3%). The lack or lower quality of structure information is part of the reasons for our limited CAFA3 performance in MF, because the specificity of molecular function such as biomolecule binding and catalytic activity is determined by structure.

Our performance on the human subset of CAFA3 is shown in Figure 23. Since prediction models from other CAFA3 predictor teams are not publicly available, we compare our predictions obtained during CAFA3 challenge in 2017-02-02 with two baseline methods implemented by CAFA3 assessors: (1) the "BLAST" method searching a query against UniProt, where the prediction C-score of GO term *q* equals the sequence identity at the BLAST-aligned region between query and the top BLAST hit annotated with *q*; and (2) the "Naïve" method, equivalent to the background distribution of GO terms: for any target, "Naïve" predicts every GO term, where the C-score of GO term *q* equals the number of UniProt proteins experimentally annotated with *q* divided by the total number of experimentally annotated UniProt proteins.

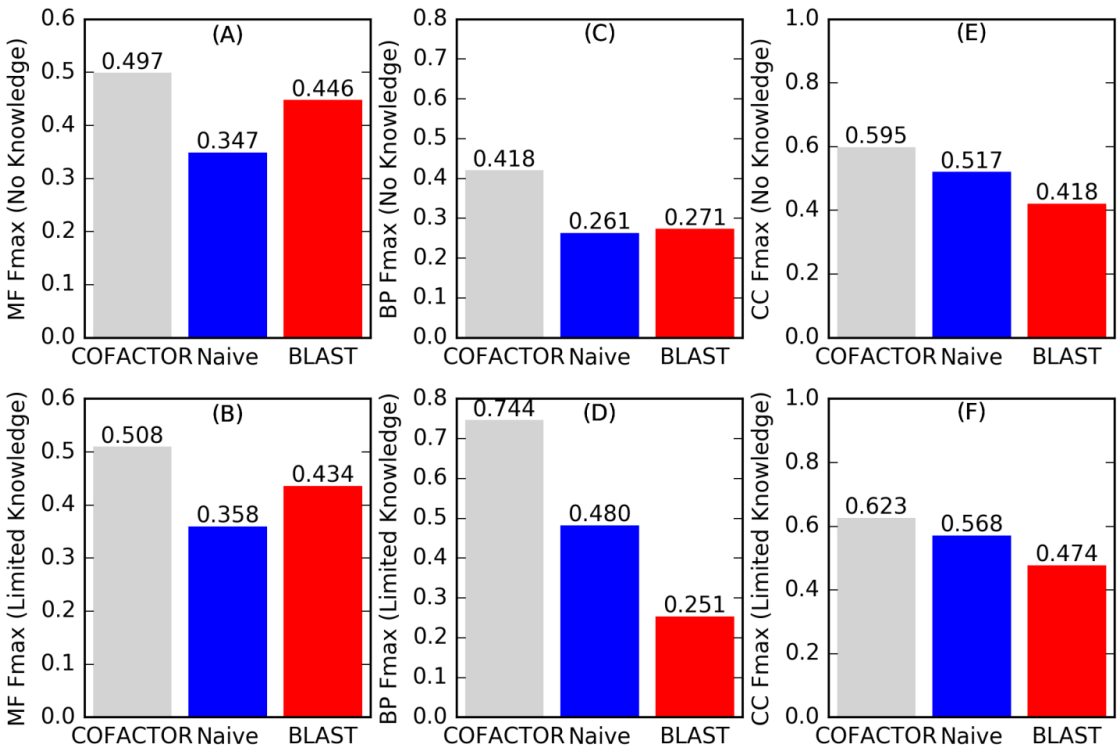


Figure 23.Fmax for MF (A, B), BP (C, D) and CC (E, F) GO term prediction by COFACTOR (Zhang-Freddolino lab) and two baseline methods, "Naïve" and "BLAST" for "No Knowledge" (A, C, E) and "Limited Knowledge" (B, D, F) targets. Fmax calculations exclude GO terms annotated before 2017-02-02.

As shown in Figure 23, COFACTOR prediction consistently outperformed the baseline methods in CAFA3 for all assessment categories. The advantage is particularly evident for BP, where our Fmax was 54% and 55% higher than the best performing baseline methods for "No Knowledge" and "Limited Knowledge" types. Moreover, COFACTOR outperforms "Naïve" on CC by 15% and 10% for "No Knowledge" and "Limited Knowledge" targets. No computational method outperformed "Naïve" in CAFA2.[306] Even though our predictions ultimately derive function annotation from UniProt annotated GO terms similar to the "Naïve" and "BLAST" baseline methods, COFACTOR more effectively identifies functional templates and combines their GO annotations, instead of relying on simple sequence similarity search ("BLAST") or accepting background distribution of GO terms ("Naïve").

To understand better why the functions of some targets are easier to predict than other targets, we computed the Pearson's correlation coefficient (PCC) between various features of the query protein in the target set and the Fmax of GO term prediction accuracy for MF and BP (Figure 24) based on the supplementary data accompanying the CAFA3[304] report (https://figshare.com/articles/Supplementary_data/8135393). For a meta-server (such as COFACTOR) that combines multiple features (identities of multiple sequence homologs, STRING scores of PPI partners, and similarities of structure templates) to derive a consensus prediction, it is often impossible for the consensus prediction to be dependent only on one feature. Nevertheless, it is still possible to identify whether the quality of a feature affects the accuracy of final prediction in a statistically significant manner. For example, while we did not observe significant dependence of Fmax on query sequence length (Figure 24A), Fmax of the sequence-, PPI-, and structure-based component methods of COFACTOR is significantly dependent upon the availability of templates or interaction partners, as quantified by their sequence identity (Figure 24B), STRING score (Figure 24C), and TM-score (Figure 24D), especially when the template score is modest (sequence identity<0.5, STRING score<0.7, or TM-score<0.6). However, for all three methods, the correlation coefficient between Fmax and the score of first template is not high ($|PCC| \leq 0.3$), partly because each of the three component methods is a consensus approach to simultaneously consider all template hits (Equation 2), so that the prediction result for a GO term will not be completely biased by a single high scoring template.

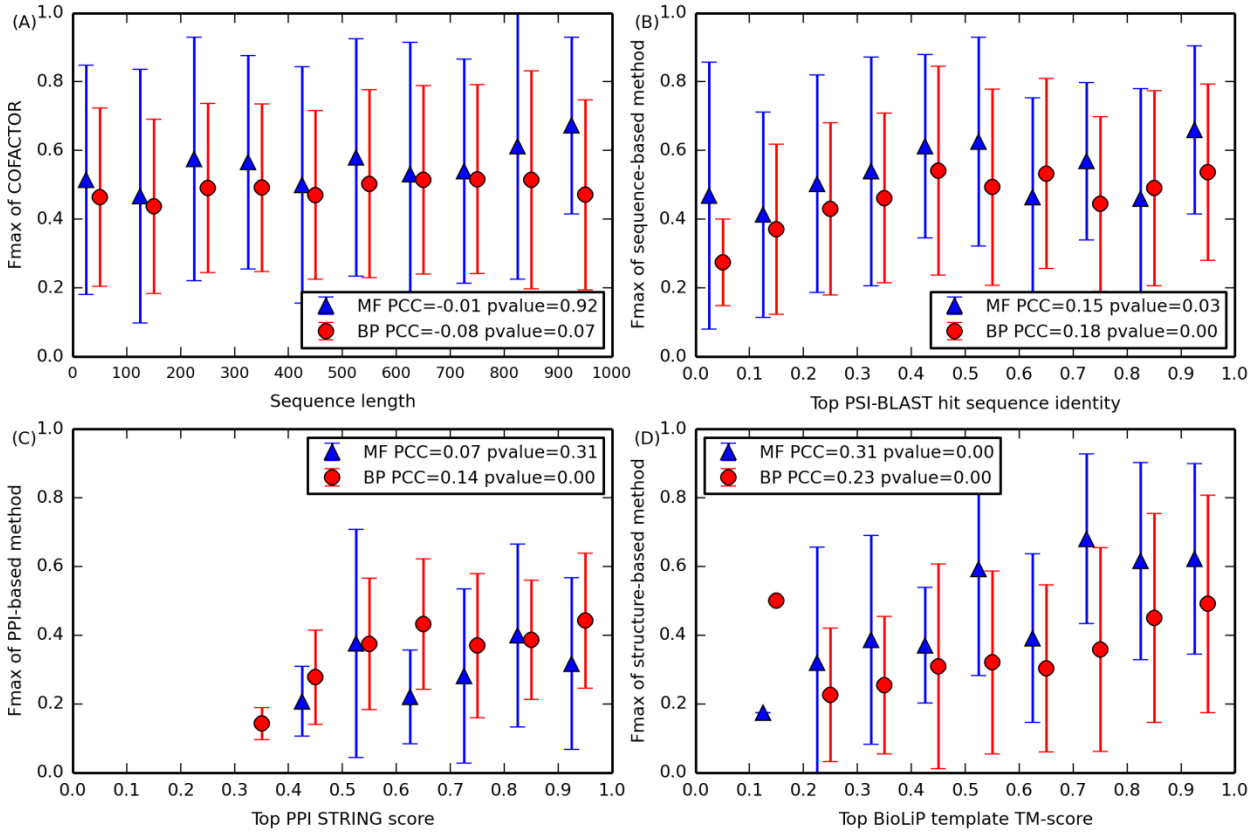


Figure 24. Fmax of MF (blue triangle) and BP (red circle) GO term prediction versus (A) sequence length, (B) global sequence identity of closest PSI-BLAST hit, (C) highest PPI interaction score (STRING score), and (D) TM-score between query structure and the closest BioLiP template. A pair of error bars marks the standard deviation of Fmax at each bin. Inside each figure legend, the two numbers are the PCC and its p-value, respectively[307]. Among the set of 267 and 912 CAFA3 human targets for MF and BP, all were subjected to function prediction based on sequence and PPI by COFACTOR; only 88 and 227 targets, respectively, were predicted by the structure-based pipeline of COFACTOR (D).

In short, our analysis indicates that, while COFACTOR is a good function predictor that goes far beyond simple sequence homology searching, it still has the intrinsic limitation of a template-based predictor, where target proteins with less reliable templates are more difficult to model.

5.3.4 Evaluation of free-text and GO term prediction using newly-annotated uPE1 proteins in neXtProt 2019-01-11

While CAFA3 provides a large blindly-tested set for relatively robust statistical analysis of our GO term prediction performance, we could not evaluate the performance of free-text function as it was neither required nor feasible given the very large set of targets in the CAFA3 challenge. To complement the CAFA3 evaluation and contribute to the C-HPP uPE1 CP50 Challenge[308], we assessed the I-TASSER/COFACTOR pipeline on a narrowly focused blindly-tested set of 25 previously unannotated proteins with new function annotation in the neXtProt 2019-01-11 release. The detail findings are presented in our online supplementary data at https://zhanglab.ccmb.med.umich.edu/COFACTOR2/nx2019addition/GOterm.html#2, while a simplified table is presented in Table 8.

Among the 25 targets in this time-elapsed blindly-tested set, 3 have I-TASSER models that are predicted to have correct structure topology (estimated TM-score >0.5: #2, 17, 18 in Table 8), while another 10 are predicted to have approximately correct fold (estimated TM-score in the range [0.4,0.5]: #1, 4, 5, 8, 9, 10, 12, 14, 20, 24 in Table 8).

Among the 25 targets, we did not assign free-text function annotation for 3 (O75363-1, Q8NDM7-1, and Q9BZH6-1; #8, 16, and 11, respectively, in Table 8), because the GO terms we predicted for these targets are too general to infer the function. For the remaining 22 targets, our manual free-text function interpretations are consistent with neXtProt annotation for 9 of them, as marked by asterisks (*) in Table 8 (#1, 2, 3, 6, 7, 9, 15, 18, 19). Meanwhile, of the 8 and 22 targets with UniProt/neXtProt curated MF or BP GO terms, 3 (#1, 2, 3) for MF and 4 (#1, 2, 3, 4) for BP have Fmax≥0.5 for our GO term prediction. That makes a total of 4 different targets of the 25 with good matches for GO terms, only one (#4) of which is in addition to the 9 above with

good matches for free-text, making a total of 10 that have good matches by either free-text or GO terms or both.

Table 8. Comparison of I-TASSER/COFACTOR function annotation and UniProt/neXtProt curation for 25 uPE1 with newly provided function annotation in neXtProt release 2019-01-11. Full detail of this table is available at https://zhanglab.ccmb.med.umich.edu/COFACTOR2/nx2019addition/GOterm.html#2

(a) An asterisk (*) marks a target if our free-text annotation matches neXtProt free text annotation.
(b) A plus (+) marks a target whose Fmax for either MF or BP is >0.5 but the free-text annotation does not match. Fmax for MF/BP quantitatively measures the consistency between COFACTOR predicted GO terms and neXtProt curated GO terms. "NA", or not applicable, means neXtProt did not assign GO term for a target. The table is ranked in descending order of Fmax.
(c) In the last column, phrases at top are free-text annotations, followed by MF and BP GO terms. Red shades indicate free-text phrases consistent between I-TASSER/COFACTOR prediction and neXtProt annotation.

#	accession, gene	*	our annotation	Fmax MF/BP	excerpt of UniProt/neXtProt annotation
1	Q96M27-1, PRRC1	*	Protein kinase A regulation	1.00, 0.88	Activation of protein kinase A activity. Protein kinase A regulatory subunit binding.
2	P0C870-1, JMJD7	*	Histone demethylation	0.55, 0.90	Endopeptidase cleaving histones N-terminal tails at the carboxyl side of methylated arginine or lysine residues. Fe2+ and 2-oxoglutarate-dependent monoxygenase.
3	Q7Z5A7-1, FAM19A5	*	Regulation of microglial cell activation	0.67, 0.50	Stimulates chemotactic migration of macrophages. Blocks osteoclast formation from macrophages. Negatively regulating vascular smooth muscle cell (VSMC) proliferation and migration. Inhibits injury-induced cell proliferation and neointima formation in the femoral arteries
4	Q5T0D9-1, TPRG1L	+	Phosphatidylinositol-4-phosphate phosphatase	NA, 0.55	Regulates synaptic release probability by decreasing the calcium sensitivity of release.
5	Q96D15-1, RCN3		Catalytic activity, acting on a protein	NA, 0.47	Molecular chaperone assisting protein biosynthesis and transport in endoplasmic reticulum. Pulmonary surfactant homeostasis. Anti-fibrotic activity by negatively regulating the secretion of collagens.
6	Q8WTR8-1, NTN5	*	Anatomical structure morphogenesis	NA, 0.44	Neurogenesis. Prevents motor neuron cell body migration out of the neural tube.
7	Q9C0D6-1, FHDC1	*	Binding of cytoskeleton	0.44, 0.33	Microtubule-associated formin. Regulates actin and microtubule dynamics. Induces microtubule acetylation and stabilization and actin stress fiber formation. Regulates Golgi ribbon formation. Required for normal cilia assembly.
8	O75363-1, BCAS1		(for CC: neuron part)	NA, 0.40	Myelination.
9	P60827-1, C1QTNF8	*	Signaling receptor binding	NA, 0.40	Relaxin receptor RXFP1 binding.
10	Q8IUY3-1, GRAMD2A		Binding of GTPase from Ras superfamily	0.11, 0.38	Organization of endoplasmic reticulum-plasma membrane contact sites. STIM1 recruitment and calcium homeostasis.
11	Q9BZH6-1, WDR11			NA, 0.35	Involved in Hedgehog (Hh) signaling pathway. Essential for normal ciliogenesis.
12	Q6ZNE9-2, RUFY4		Regulation of protein folding	0.26, 0.32	Positively regulates macroautophagy in primary dendritic cells. Increases autophagic flux by stimulating autophagosome formation and facilitating tethering with lysosomes. Binds to phosphatidylinositol 3-phosphate (PtdIns3P).
13	Q9GZU8-1, FAM192A		Hydrolase of protein	NA, 0.32	Promotes the association of the proteasome activator complex subunit PSME3 with the 20S proteasome and regulates its activity. Inhibits PSME3-mediated degradation of proteasome substrates.
14	Q494U1-1, PLEKHN1		Transmembrane transport of nucleotide	0.05, 0.29	Controls the stability of the leptin mRNA harboring an AU-rich element (ARE) in its 3' UTR.
15	Q8IUW5-1, RELL1	*	Regulation of apoptosis through TNF	NA, 0.29	Induces activation of MAPK14/p38 cascade.
16	Q8NDM7-1, CFAP43			NA, 0.29	Flagellar protein involved in sperm flagellum axoneme organization and function.
17	Q8TDG2-1, ACTRT1		Regulation of chromosome organization either though histone acetylation or binding of cytoskelton used in chromsome segregation	0.03, 0.29	Negatively regulates the Hedgehog (SHH) signaling. Binds to the promoter of the SHH signaling mediator, GLI1, and inhibits its expression.
18	O75677-1, RFPL1	*	Ubiquitin-protein transferase activity	NA, 0.27	Negatively regulates the G2-M phase transition, by promoting cyclin B1/CCNB1 and CDK1 proteasomal ubiquitin-dependent degradation.
19	Q5VTQ0-1, TTC39B	*	Protein ubiquitination regulation	NA, 0.26	Regulates high density lipoprotein (HDL) cholesterol metabolism by promoting the ubiquitination and degradation of the oxysterols receptors.
20	Q96S16-1, JMJD8		Histone demethylation	NA, 0.21	Positive regulator of TNF-induced NF-κB signaling. Regulates angiogenesis and cellular metabolism.
21	Q9H9L7-1, AKIRIN1		By binding to RNA polymerase, regulate expression of genes such as cytokines	NA, 0.18	Signal transducer for MSTN during skeletal muscle regeneration and myogenesis. Regulates chemotaxis of macrophages and myoblasts by reorganising actin cytoskeleton, leading to more efficient lamellipodia formation via a PI3 kinase dependent pathway.
22	Q96KV7-1, WDR90		Regulation of transcription by nucleic acid binding	NA, 0.17	Required for efficient primary cilium formation.
23	Q6AI39-1, BICRAL		Sodium:potassium ion transporter	NA, NA	Enzyme component of SWI/SNF chromatin remodeling subcomplex GBAF, changing chromatin structure by altering DNA-histone contacts in an ATP-dependent manner.
4	Q96J88-1, EPSTI1		Cytoskeleton binding	NA, NA	M1 macrophage polarization. Regulation of gene expression during macrophage differentiation. RELA/p65 and STAT1 phosphorylation and nuclear localization upon activation of macrophages.
25	Q9BZD6-1, PRRG4		Serine-type endopeptidase	NA, NA	Axon guidance across the CNS. Prevents the delivery of ROBO1 at the cell surface and downregulates its expression.

Of course, these newly-annotated proteins represent the 7% of PE1 proteins that have resisted functional annotation. Thus, the overall low Fmax of agreement between GO term predictions and literature curation (0.19 and 0.23 for MF and BP, respectively, for the 8 and 22 proteins) is partly attributable to incompleteness in GO term annotation. In fact, our BP prediction accuracy is >15% higher than three state-of-the-art GO term prediction programs, GoFDR[267], GOtcha[309], and DeepGOplus[310] (Table S2). In many scenarios, both our method and UniProt/neXtProt curation may only capture some of the many functions a protein performs. For example, target RFPL1 (O75677-1, #18 in Table 8) regulates the cell cycle by promoting ubiquitin-dependent protein degradation according to UniProt/neXtProt. While our MF prediction inferred the ubiquitin-dependent protein degradation function, our BP term prediction did not correctly predict the cell cycle regulation function, resulting in a low Fmax of 0.27 for BP GO terms despite partially consistent function annotation.

Table 9. Comparison of GO terms prediction accuracy (Fmax) between I-TASSER/COFACTOR our function annotation by I-TASSER/COFACTOR and state-of-the-art methods for 8 and 22 neXtProt proteins with newly annotated MF and BP GO terms. Bold font indicates the most accurate algorithm in each aspect for this dataset. While COFACTOR is on average more accurate than both GoFDR and GOtcha for all three aspects of GO terms as shown in large-scale benchmark studies[201,303], its MF prediction accuracy (Fmax) is lower than GoFDR and GOtcha for this set of targets in MF prediction, probably due to the very small dataset size of only 8 proteins.

Program	Fmax for MF of 8 proteins	Fmax for BP of 22 proteins
I-TASSER/COFACTOR	0.19	**0.23**
GoFDR	**0.28**	0.20
Gotcha	0.20	0.11
DeepGOplus	0.17	0.16

Such incompleteness of GO term annotation is not uncommon for UniProt/neXtProt literature curation. C1QTNF8 (P60827-1, #9 in Table 8) binds the G protein-coupled receptor RXFP1 (MF) to regulate cell motility (BP). Swiss-Prot curators annotated the protein with the free text "May play a role as ligand of RXFP1" to convey its molecular function without a GO term; the GO consortium annotated GO:2000147 “positive regulation of cell motility” for BP

based on the same experimental paper[311]. This causes the lack of an appropriate MF GO term for this protein such as GO:0001664 “G protein-coupled receptor binding” or GO:0005102 “signaling receptor binding”. Consequently, even though COFACTOR indeed predicts GO:0005102, we cannot calculate Fmax for MF and have a modest Fmax=0.40 for BP, despite our consistent free-text interpretation “signaling receptor binding”. While the incompleteness of function curation partly accounts for the low Fmax on this dataset, our earlier benchmark performed last year on 100 PE1 proteins resulted in much higher Fmax of 0.69 and 0.57 for MF and BP, respectively,[303] as that benchmark dataset only included deeply annotated targets with at least 3 Gold GO terms for each of the three GO terms aspects (MF, BP, and CC). Partly due to incompleteness of GO term annotation in the small dataset reported in this study, Fmax of COFACTOR GO term prediction does not have apparent correlation with features of targets such as template availability (Figure 25).

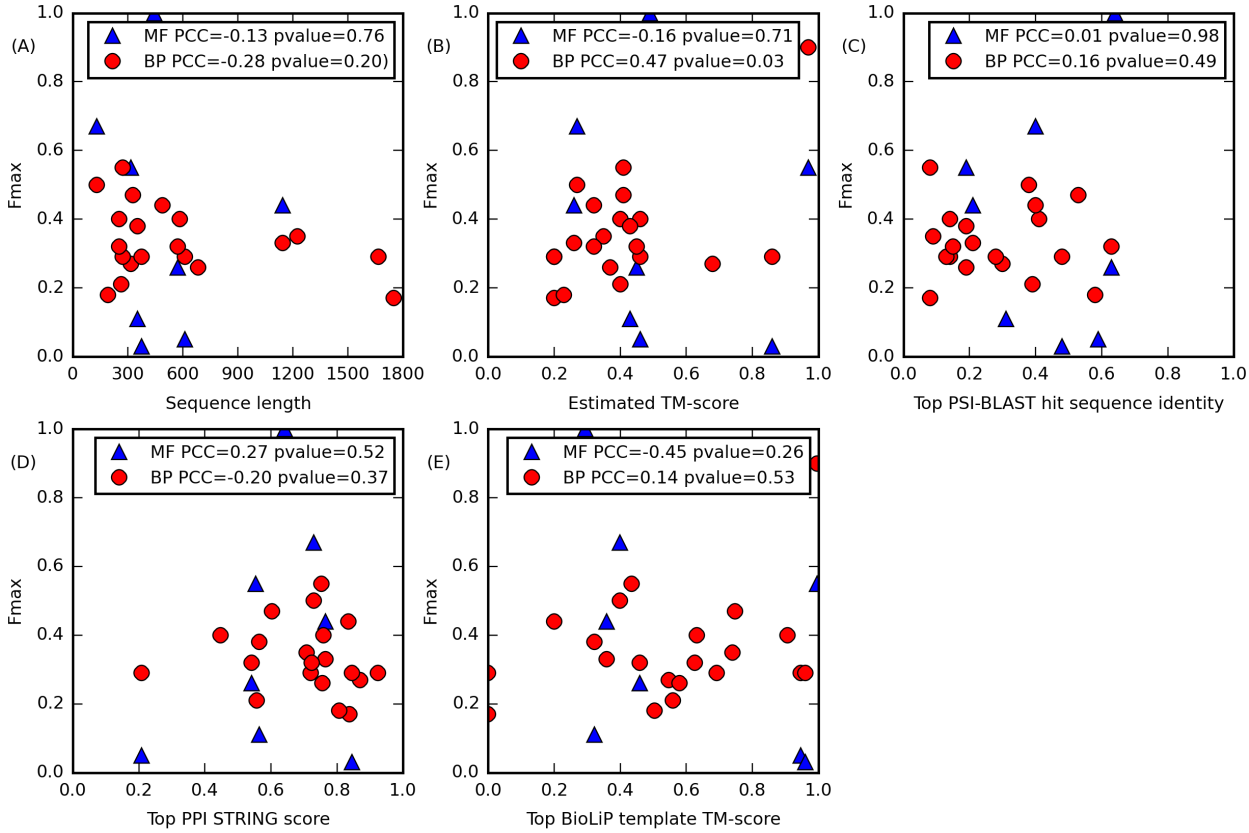


Figure 25. Fmax versus features of target protein in time-elapsed set of 8 and 22 proteins with MF and BP GO terms by UniProt/neXtProt. Inside each figure legend, the two numbers are the Pearson's correlation coefficient (PCC) between Fmax and target protein feature, followed by the p-value of PCC.

Among the 25 proteins in this time-elapsed neXtProt blindly-tested set, we highlight three representative function predictions. As the first example, we discuss P0C870-1 (JMJD7, #2 in Table 8), a recently characterized endopeptidase and monooxygenase[312,313], to illustrate the importance of structure template alignment and local sequence homolog hits in function prediction. As an endopeptidase, JMJD7 cuts histones at methylated arginine residues (GO:0035064 "methylated histone binding", GO:0004177 "aminopeptidase activity", GO:0004175 "endopeptidase activity" for MF); as a Fe^{2+} and 2-oxoglutarate-dependent monooxygenase, JMJD7 catalyzes hydroxylation of DRG1 and DRG2 translation factors (GO:0016706 "2-oxoglutarate-dependent dioxygenase activity", GO:0004497 "monooxygenase

activity" for MF, GO:0018126 "protein hydroxylation" for BP). The I-TASSER structure model of JMJD7 displays a typical "Jelly roll" fold and shares a high TM-score[247] of 0.98 for both of its recently solved structures (PDB IDs 5nfn Chain A and 5nfo Chain A, Figure 26A), even though neither of the two experimental structures was used in the I-TASSER modeling or function prediction. The structure of JMJD7 is similar to two human oxidoreductases: PDB IDs 3al5 Chain B (TM-score=0.69, Figure 26B), and PDB ID 4b7e Chain A (TM-score=0.70, Figure 26C), which are tRNA hydroxylase and hypoxia-inducible factor-asparagine dioxygenase, respectively. Despite the matching of these two structure analogs and the correct prediction of GO:0016706 "2-oxoglutarate-dependent dioxygenase activity" at C-score=0.32 by COFACTOR structure-based method, the I-TASSER model also shares high structure similarity to many other proteins that perform other unrelated functions such as GO:0070492 "oligosaccharide binding" and GO:0005215 "transporter activity", both at C-score=0.53 by structure-based method, partly because the Jelly roll fold is a common topology in a wide variety of proteins. In the sequence-based method, the closest oxidoreductase hit is Lysine-specific demethylase 8 (UniProt ID B2GUS6), with only 27% sequence identity at the aligned region. Despite the low sequence identity with the top hit, 56% of the BLAST and PSI-BLAST hits are annotated with oxidoreductase activity (GO:0016706), resulting in highly confident prediction of this term at C-score=0.62 for sequence-based method and C-score=0.76 for the final consensus COFACTOR prediction. Although our predicted GO terms for both MF and BP overlap very well with neXtProt annotation (Fmax=0.55 and 0.90, respectively), our blinded manual free-text interpretation process chose the term "histone demethylase activity" (GO:0032452, C-score=0.63) to derive the free-text function annotation "histone demethylation", which is not fully consistent with UniProt/neXtProt annotation, even though it correctly indicates the

oxidoreductase activity of JMJD7 on methylated histones. This reflects the difficulty of manual interpretation of the function given only partially correct GO term predictions. Nevertheless, in Table 8, we designated this protein (#2) as partially matching the free-text annotations from the curators.

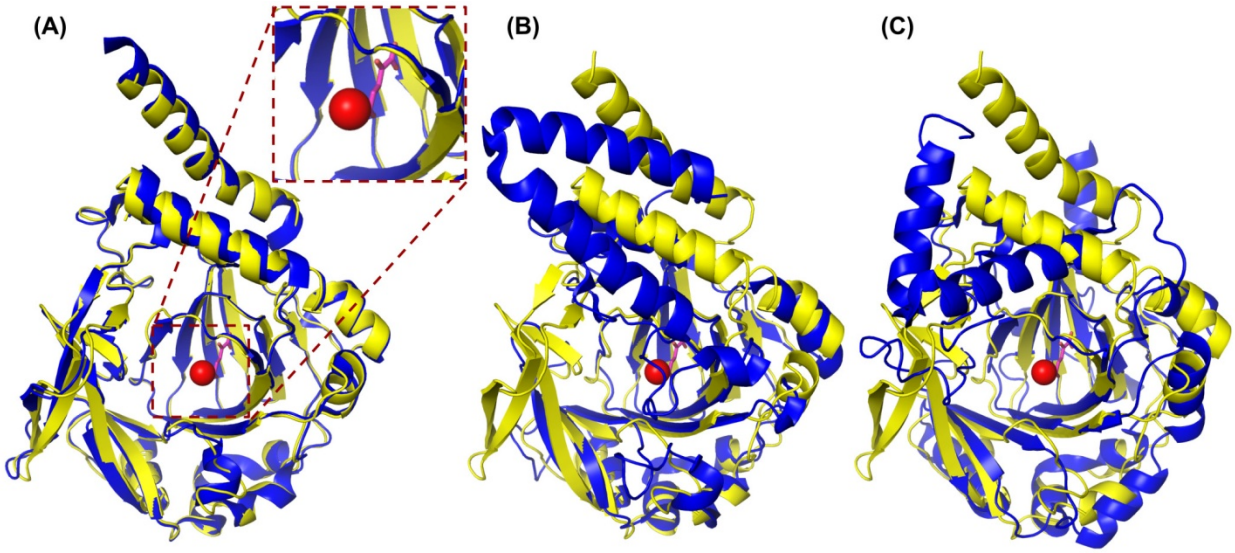


Figure 26. I-TASSER model of human JMJD7 (yellow cartoon) superposed to (A) its native structure (PDB ID 5nfn Chain A), (B) a human tRNA hydrolase (PDB ID 3al5 Chain B), and (C) a human hypoxia-inducible factor-asparagine dioxygenase (PDB ID 4b7e Chain A) in yellow blue cartoons. The JMJD7 ligand binding site (dashed inset) shows the COFACTOR predicted ligands, including Fe2+ ion (red sphere) and 2-oxoglutarate (magenta stick), both of which are known to participate in the catalytic activity of JMJD7.

Q5VTQ0-1 (TTC39B, #19 in Table 8) regulates high density lipoprotein (HDL) cholesterol metabolism by promoting the ubiquitination and degradation of the oxysterol receptors LXR (NR1H2 and NR1H3)[314]. I-TASSER/COFACTOR correctly predicts its protein ubiquitination regulation (but unfortunately not the cholesterol metabolism regulation function, resulting in low Fmax of 0.26 for MF). For this target, the prediction of protein ubiquitination regulation (GO:0006508 “proteolysis”, C-score=0.52; GO:0016567 “protein ubiquitination”, C-score=0.50), is mainly due to its structure similarity to anaphase-promoting complex subunits (Apc/C , Figure 27). This protein also has an asterisk for free-text annotation matching curators.

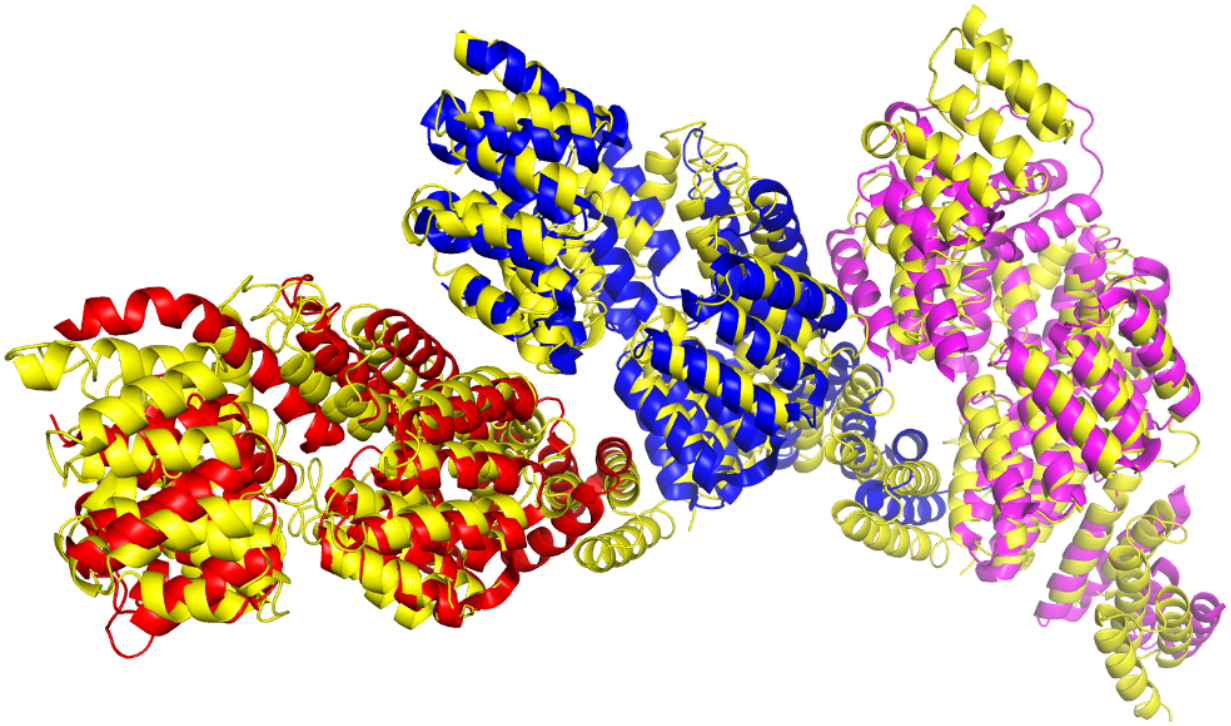


Figure 27. I-TASSER model of TTC39B (yellow cartoons) superposed to three subunits of Apc/C (5a31 Chain F, J, P in red, blue, and magenta cartoons, respectively) with TM-scores ranging from 0.66 to 0.76. Subunits of this complex are involved in regulation and catalysis of protein ubiquitination.

Q8IUW5-1 (RELL1, #15 in Table 8) is a receptor of Tumor Necrosis Factor (TNF) and induces activation of MAPK14/p38 cascade and apoptosis[315,316]. Our prediction correctly describes regulation of apoptosis through tumor necrosis factor (TNF), but did not include the MAPK14/p38 cascade regulation. On the other hand, neXtProt BP GO term annotation only includes "positive regulation of p38MAPK cascade" (GO:1900745) but did not include the TNF-mediated apoptosis, resulting in low Fmax=0.29 for BP prediction. The prediction of TNF-mediated apoptosis regulation (GO:0097190 "apoptotic signaling pathway", C-score=0.51, for BP and GO:0005031 "tumor necrosis factor-activated receptor activity", C-score=0.40, for MF) is not due to one single highly significant hit but due to multiple consensus (PSI-)BLAST hit with consistent GO term annotations. The closest sequence homolog is TNF receptor

superfamily member 19L (UniProt ID Q969Z4), which shares 30% sequence identity with the query. This protein ends up with a match for free-text but not for GO terms.

5.3.5 Summary of Predicted Structure and Functions of the 66 uPE1 Proteins

For the 66 chromosome 17 uPE1 proteins, when using the I-TASSER/COFACTOR pipeline, homologous templates are not excluded, because we want to obtain the best possible structure and function modeling results for these real prediction targets. Among the first ranked I-TASSER model of these uPE1 proteins, models of 12 proteins are predicted to have correct fold (estimated TM-score >0.5), while 13 are predicted to have roughly correct fold (estimated TM-score >0.4 and ≤0.5).

For prediction of GO terms for these uPE1 proteins, using C-scores >0.59, >0.55, and >0.56 established by Figure 22 as thresholds for reliable COFACTOR prediction for MF, BP, and CC, respectively, we obtained confident predictions for 13, 33, and 49 proteins for the respective GO term aspects (Figure 28). If these stringent C-score cutoffs are slightly relaxed such that we also consider predicted GO terms with C-score > 0.5, the number of uPE1 proteins with predicted GO terms will be increased to 30, 39, and 58 for MF, BP, and CC, respectively.

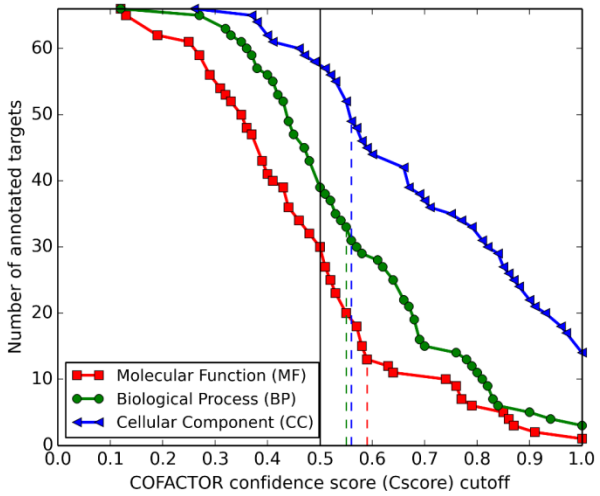


Figure 28. Number of uPE1 proteins with GO term prediction at different C-score thresholds. The solid black vertical line marks the C-score=0.5, while the red, green, and blue dashed vertical lines indicate C-score cutoffs 0.59, 0.55, and 0.56 for MF, BP, and CC, respectively. Here, GO terms associated with more than 20% of proteins in the UniProt database are excluded, because these GO terms, such as "protein binding", are too general to provide meaningful insight into their specific function.

An online supplementary table at https://zhanglab.ccmb.med.umich.edu/COFACTOR2/chr17/ann_small_table.html summarizes all predicted functions for all 66 uPE1 proteins. As a concise entry to the full table, we list the top 13 uPE1 proteins with highest C-scores for MF GO terms in Table 10.

Table 10. A concise table for 13 uPE1 proteins with high confidence predicted functions for MF. For each of the three aspects, MF, BP and CC, the GO term with the highest confidence and the GO term with C-score >0.5 that can provide specific biological insight are listed, with the C-score enclosed by parentheses. The four entries discussed as case studies in the following sections are indicated with asterisks. The entries are in descending order according to MF C-score.

	NeXtProt ID (Gene Name)	Molecular Function (MF)	Biological Process (BP)	Cellular Component (CC)
1*	NX_Q8TBR7-2 (FAM57A)	GO:0016740 (1.00) transferase activity GO:0050291 (0.99) sphingosine N-acyltransferase activity	GO:0032502 (0.69) developmental process GO:0007420 (0.54) brain development	GO:0005887 (1.00) integral component of plasma membrane GO:0005886 (1.00) plasma membrane
2	NX_Q12767-1 (TMEM94)	GO:0022892 (0.91) substrate-specific transporter activity GO:0046873 (0.57) metal ion transmembrane transporter activity	GO:0065008 (0.80) regulation of biological quality GO:0030001 (0.56) metal ion transport	GO:0005654 (1.00) nucleoplasm
3	NX_Q5BKU9-1 (OXLD1)	GO:0016491 (0.87) oxidoreductase activity GO:0004128 (0.73) cytochrome-b5 reductase activity, acting on NAD(P)H	GO:0015701 (0.90) bicarbonate transport GO:0008652 (0.53) cellular amino acid biosynthetic process	GO:0005739 (0.90) mitochondrion GO:0005737 (0.66) cytoplasm
4*	NX_A6NGC4-1 (TLCD2)	GO:0016740 (0.86) transferase activity GO:0050291 (0.76) sphingosine N-acyltransferase activity	GO:0006643 (0.76) membrane lipid metabolic process GO:0006672 (0.73) ceramide metabolic process	GO:0016021 (1.00) integral component of membrane GO:0005783 (0.75) endoplasmic reticulum
5*	NX_O43934-1 (MFSD11)	GO:0005215 (0.85) transporter activity GO:0005351 (0.66) sugar:proton symporter activity	GO:0006810 (0.82) transport GO:0008643 (0.68) carbohydrate transport	GO:0016021 (1.00) integral component of membrane GO:0005887 (0.77) integral component of plasma membrane
6	NX_Q9P298-1 (HIGD1B)	GO:0016740 (0.79) transferase activity GO:0061630 (0.71) ubiquitin protein ligase activity		GO:0043234 (0.88) protein complex GO:0005634 (0.71) nucleus

	NeXtProt ID (Gene Name)	Molecular Function (MF)	Biological Process (BP)	Cellular Component (CC)
7	NX_Q2TAL5-1 (SMTNL2)	GO:0008092 (0.77) cytoskeletal protein binding	GO:0016043 (0.70) cellular component organization GO:0048856 (0.59) anatomical structure development	GO:0005737 (0.66) cytoplasm GO:0044430 (0.50) cytoskeletal part
8	NX_Q9BQS6-1 (HSPB9)	GO:0042802 (0.76) identical protein binding GO:0051082 (0.52) unfolded protein binding	GO:0050896 (0.82) response to stimulus GO:0042981 (0.51) regulation of apoptotic process	GO:0005634 (0.97) nucleus GO:0005737 (0.96) cytoplasm
9	NX_Q96LD4-1 (TRIM47)	GO:0004842 (0.76) ubiquitin-protein transferase activity	GO:0031323 (0.54) regulation of cellular metabolic process GO:0019538 (0.54) protein metabolic process	GO:0005737 (0.57) cytoplasm
10	NX_Q8N7B9-1 (EFCAB3)	GO:0043169 (0.74) cation binding	GO:0019538 (0.58) protein metabolic process	GO:0016020 (0.82) membrane GO:0005737 (0.68) cytoplasm
11*	NX_Q6AI12-1 (ANKRD40)	GO:0008092 (0.62) cytoskeletal protein binding GO:0030507 (0.57) spectrin binding	GO:0060255 (0.62) regulation of macromolecule metabolic process GO:0016043 (0.60) cellular component organization	GO:0005737 (0.77) cytoplasm GO:0043234 (0.51) protein complex
12	NX_Q6UX52-1 (C17orf99)	GO:0004872 (0.63) receptor activity GO:0019199 (0.50) transmembrane receptor protein kinase activity	GO:0032502 (0.68) developmental process GO:0030030 (0.54) cell projection organization	GO:0031224 (1.00) intrinsic component of membrane GO:0005887 (0.63) integral component of plasma membrane
13	NX_Q3MHD2-1 (LSM12)	GO:0003723 (0.59) RNA binding	GO:0090304 (0.79) nucleic acid metabolic process GO:0016070 (0.73) RNA metabolic process	GO:0005576 (0.55) extracellular region

It can also be observed that the number of confidently annotated proteins is smaller for MF compared to BP and CC. This is partially due to the fact that, while most of these 66 uPE1 proteins lack close sequence homologs, the majority (56 of 66) have known or inferred PPI information, which COFACTOR can take advantage of in BP and CC prediction. For example, the uPE1 protein C17orf82 (neXtProt ID: NX_Q86X59-1) does not have any strong sequence or structure template hit, but interacts with proteins known to be involved in developmental processes or cellular component organization (https://string-db.org/network/9606.ENSP00000335229). Using the homologs of these PPI partners, COFACTOR deduces that the target protein is involved in “cellular component organization”

(GO:0016043, C-score=0.55) and “developmental process” (GO:0032502, C-score=0.52). While PPI is informative of BP and CC, it is not as useful for MF prediction, because proteins that physically interact with each other do not necessarily share the same molecular function (MF), even though they generally are involved in the same pathway (BP) at the same subcellular location (CC).

Among the uPE1 proteins with relatively confidently predicted functions (Figure 29), 7 are associated with cytoskeleton (GO:0008092 “cytoskeletal protein binding” for MF and GO:0044430 “cytoskeletal part” for CC), while another 7 are putative transmembrane transporters (GO:0022857 “transmembrane transporter activity” for MF). Other notable predicted biological functions shared by multiple uPE1 proteins include nucleic acid binding (GO:0003676 “nucleic acid binding” for MF and GO:0090304 “nucleic acid metabolic process” for BP), ubiquitin-dependent protein degradation (GO:0004842“ubiquitin-protein transferase activity” for MF and GO:0006511 “ubiquitin-dependent protein catabolic process” for BP), and N-acylsphingosine synthesis (GO:0050291 “sphingosine N-acyltransferase activity” for MF). Here we include both GO terms predicted with the stringent C-score cutoffs 0.59, 0.55, and 0.56 for MF, BP, and CC, respectively (Figure 29, gray), and the GO terms predicted with the relaxed C-score cutoffs 0.50 for all three aspects (Figure 29, white). There is no major difference in the source of prediction (structure, sequence, or PPI), the distribution of prevalent GO terms or the Fmax that resulted from the two sets of C-score cutoffs.

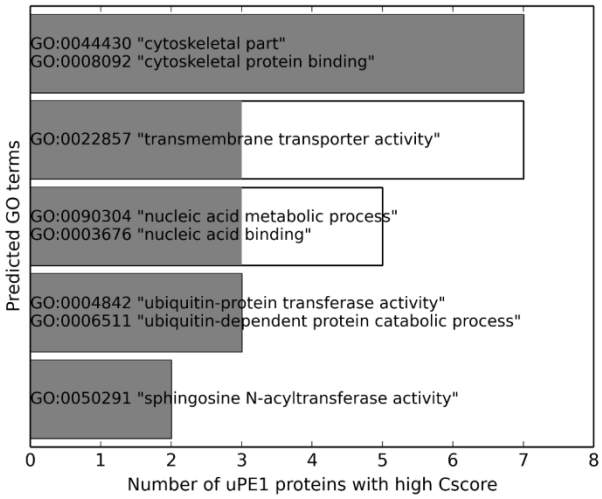


Figure 29. Notable GO terms predicted with high C-score for multiple uPE1 proteins. White bars show the number of proteins predicted with C-score > 0.5 for given GO terms, while the gray bars show the number of proteins predicted with C-score > 0.59, 0.55, and 0.56 for MF, BP, and CC, respectively.

Case Studies of Predicted Function of uPE1 Proteins

For this section, we selected four uPE1 proteins whose specific biological functions are predicted with a high MF C-score by COFACTOR plus one uPE1 protein predicted with a high CC C-score for manual interpretation of their likely structure and function, as well as the origin of the function assertion by our pipeline.

MFSD11 (neXtProt ID: NX_O43934-1) is a hard function prediction target with neither experimentally solved structure nor any functionally characterized sequence homolog sharing >30% sequence identity. The I-TASSER structure model of this target shows a multi-pass transmembrane helical protein topology with high confidence: the TM-score of the model, as estimated by statistical significance of threading template hits and convergence of folding simulation,[298] is as high as 0.86. The structure model superposes well to a proton:xylose symporter (PDB entry 4gby chain A, Figure 30), from which COFACTOR asserted that the MF for the target protein of interest is “sugar transmembrane transporter activity” (GO:0051119, C-score=0.74). This function prediction is consistent with a previous study[317], which suggested that

MFSD11 may be a membrane protein that transports soluble molecules and is involved in energy regulation.

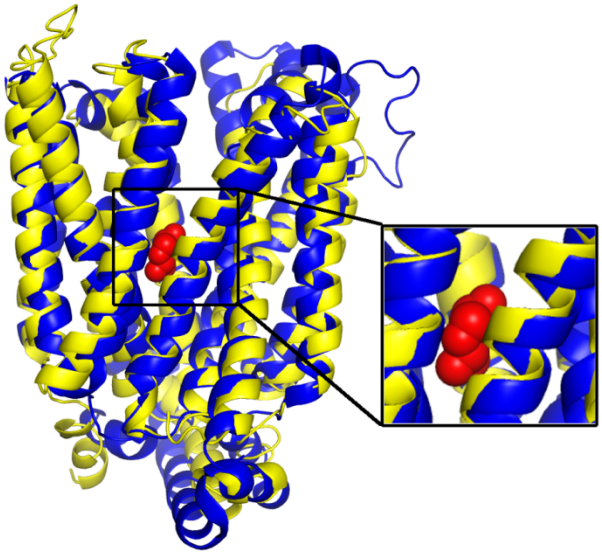


Figure 30. I-TASSER model of MFSD11 (yellow) superposed to the *E. coli* proton:xylose symporter (PDB entry 4gby chain A, blue) with TM-score=0.86. The xylose ligand from 4gbyA is shown in red spheres in the inset.

FAM57A and TLCD2 (neXtProt ID: NX_Q8TBR7-2 and NX_A6NGC4-1, respectively) are two protein coding genes located at p13.3 region on chromosome 17, separated from each other by 0.96 million base pairs. COFACTOR considers both proteins as sphingosine N-acyltransferases (GO:0050291, C-score=0.99 for FAM57A and C-score=0.76 for TLCD2) in terms of MF. These proteins have sequence identity of only 0.24; the lack of confident predictions for the binding sites makes it infeasible to assess the active site similarity for these proteins. Sphingosine is an important phospholipid constituent of the cell membrane, and is consistent with both proteins' I-TASSER structure models, which adopt a fold typical of membrane-associated proteins (Figure 31). Moreover, FAM57A is homologous to FAM57B (neXtProt ID: NX_Q71RH2-1) with sequence identity 0.46. FAM57B is already annotated as sphingosine N-acyltransferases, which further confirms the function assertion.

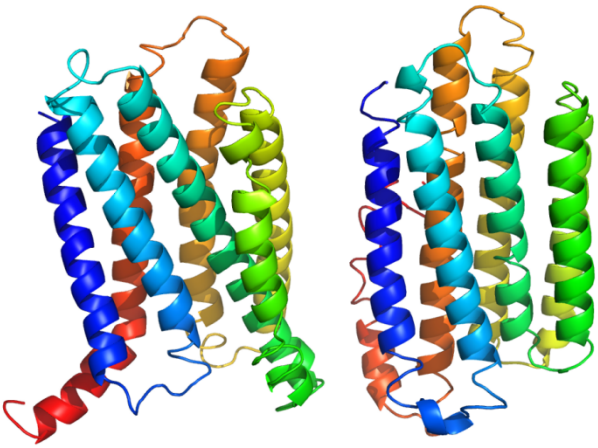


Figure 31. I-TASSER models of FAM57A (left) and TLCD2 (right). Both proteins are colored in spectrum with blue to red marking N- to C-termini.

ANKRD40 (neXtProt ID: NX_Q6AI12-1) is another hard function prediction target without functionally characterized close sequence homologs. I-TASSER predicts the target as an ankyrin repeat (Figure 32) with an estimated TM-score of 0.51. Based on the known role of ankyrin repeat-containing proteins in cytoskeleton anchoring, COFACTOR predicts the molecular function of ANKRD40 as "cytoskeletal protein binding" (GO:0008092, C-score=0.62), "spectrin binding" (GO:0030507, C-score=0.57), and "cytoskeletal adaptor activity" (GO:0008093, 0.57).

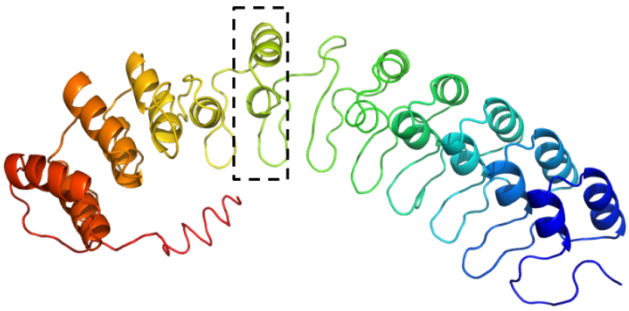


Figure 32. I-TASSER structure of ANKRD40 with nine consecutive ankyrin repeat units, each consisting of two helices linked by a loop. One ankyrin repeat unit is indicated in dashed rectangle.

Another interesting protein, based on CC prediction, is CCDC57 (neXtProt ID: NX_Q2TAC2-1), a large protein with 916 residues. While neither the sequence-based nor the PPI-based pipeline gives much hint to the function, the structure-based pipeline found that 17 of all 19 structure templates identified by the I-TASSER model belong to "phosphatidylinositol 3-kinase complex" (GO:0005942, C-score=0.89) for CC (Figure 33). This is consistent with COFACTOR's molecular function annotation "phosphatidylinositol 3-kinase activity" (GO:0035004, C-score=0.31) and biological process annotation "inositol lipid-mediated signaling" (GO:0048017, C-score=0.41), even though both function predictions have relatively low to moderate C-scores. Phosphatidylinositol triphosphate (PI3P) is a phospholipid found in membranes that helps to recruit a range of proteins, many of which are involved in protein trafficking; we conclude that CCDC57 has a related function.

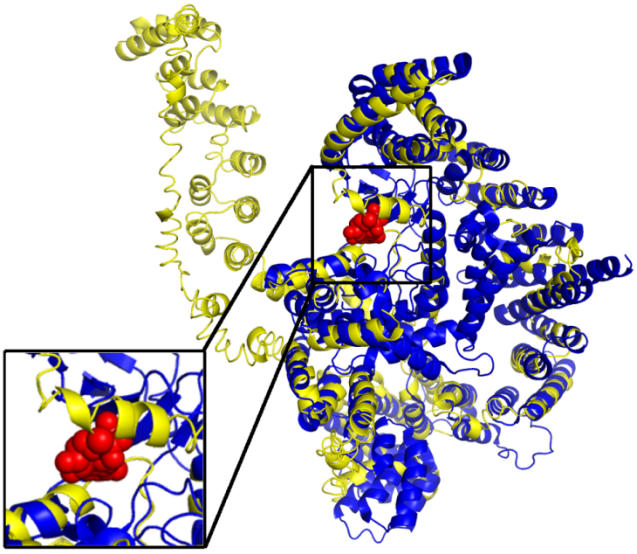


Figure 33. I-TASSER model of CCDC57 (yellow) superposed to PDB entry 4jsp chain A (blue), one of the many structure templates associated with phosphoinositide 3-kinase complex. The ligand bound to the 4jsp structure is phosphothiophosphoric acid-adenylate ester (red spheres), which is a small molecule analog of ATP, one of the substrates of phosphoinositide 3-kinases.

5.3.6 Comparing COFACTOR Prediction with Very Recent Function Annotations

The list of 66 uPE1 proteins was originally curated based on the lack of function annotations in neXtProt release 2017-08-01. Two previously unannotated proteins have new characterized functions. When we were drafting this manuscript, neXtProt release 2018-01-17 became available, with a finding that EVI2B (neXtProt ID: NX_P34910-1) regulates hematopoietic stem cell division and granulocyte differentiation.[318] COFACTOR failed to predict the highly specific BP function of this protein, only suggesting it is an "integral component of plasma membrane" (GO:0005887, C-score=1.00) for which UniProt gave the same CC term. In contrast, a recently published report characterized TRIM47 (neXtProt ID: NX_Q96LD4-1) as an E3 ubiquitin ligase;[319] the corresponding function annotation has not yet been updated in neXtProt 2018-01-17. I-TASSER/COFACTOR predicted the GO MF for TRIM47 as "ubiquitin-protein transferase activity" (GO:0004842, C-score=0.76).

5.3.7 Function Predictions that are Inconsistent with Database Annotations

For the uPE1 proteins investigated in this study, there are two cases where the I-TASSER/COFACTOR prediction is conflicting with existing annotations especially for subcellular localization (GO CC terms).

The first protein, TMEM94 (neXtProt ID: NX_Q12767-1), is annotated as "integral component of membrane" (GO:0016021) for CC in both neXtProt and UniProt with 10 predicted transmembrane helices based on automated annotation with IEA (Inferred from Electronic Annotation) evidence code by UniProt (https://www.uniprot.org/keywords/KW-0812) without experimental validation. Consistent with that database annotation, COFACTOR assigns "substrate-specific transporter activity" (GO:0022892, C-score=0.91) for MF and "metal ion

transport" (GO:0030001, C-score=0.56) for BP, both of which are associated with transmembrane transport.

We present TMEM94 as an example for inconsistency of CC prediction and neXtProt annotation. The CC result of COFACTOR for this protein is "nucleoplasm" (GO:0005654, C-score=1.00). This COFACTOR annotation, which has no counterpart in neXtProt, is generated by our sequence-based pipeline, whose function library contains the UniProt GO term of TMEM94 from year 2017 (line 382 of https://www.uniprot.org/uniprot/Q12767.txt?version=119). This UniProt annotation, labeled by UniProt with evidence "IDA:HPA" (inferred from direct assay, as reported by Human Protein Atlas database), originated from immunofluorescence experiments conducted in three human cell lines reported in the Human Protein Atlas (https://www.proteinatlas.org/ENSG00000177728-TMEM94/cell). Interestingly, while UniProt up to version 2017_02 contained the "nucleoplasm" annotation, this annotation is recently dropped by UniProt (https://www.uniprot.org/uniprot/Q12767?version=119&version=120&diff=true) even though the Human Protein Atlas experiments have not been invalidated. Since we do not exclude sequence homologs when predicting uPE1 functions, the COFACTOR sequence-based pipeline ends up hitting the TMEM94 protein itself as the "template" for its CC prediction. These differences in database annotations require further experimental efforts to determine the true or at least primary cellular component/localization of this protein.

Another example is C17orf99 (neXtProt ID: NX_Q6UX52-1), a putative human cytokine. The mouse ortholog of C17orf99 was recently established as a new 27 kDa cytokine called Interleukin 40 (IL-40), which is secreted by activated B cells.[320] Since the UniProt annotation was updated during the peer review process of this manuscript, neither the

COFACTOR function library nor the current neXtProt database (version 2018-01-17) includes this annotation. In our PSI-BLAST search for C17orf99 against human proteome (https://www.uniprot.org/proteomes/UP000005640, protein list last modified May 26, 2018), none of the top hits is cytokine, whereas the most significant hits within the human proteome are FCRL2 (neXtProt ID: NX_Q96LA5) and FCRL5 (neXtProt ID: NX_Q96RD9); both are transmembrane receptors involved in B cell development, which resulted in our pipeline's predicted CC term of C17orf99 is "intrinsic component of membrane" (GO:0031224, C-score=1.00). Nevertheless, the UniProt CC designation as "extracellular region" (GO:0005576) due to the predicted N-terminal signal peptide (https://www.nextprot.org/entry/NX_Q6UX52/sequence) and reported cytokine function may be preferable.

These contradictions in function annotations underscore the difficulty in CC prediction, which is a common challenge among many function prediction programs. In fact, it was observed in the CAFA2 experiment that almost none of the state-of-the-art programs could outperform the "Naïve" baseline in terms of CC prediction.[321] In the future, we will address the challenges in CC prediction by incorporation of amino acid composition and local sequence signatures such as predicted transmembrane regions and signal peptides into the COFACTOR function annotation algorithm.

5.4 Discussion and Conclusion

As a pilot study on prediction of functions for uncharacterized human proteins, we have carried out a comprehensive survey of PE1 proteins on chromosome 17 using the composite I-TASSER and COFACTOR structure and function annotation pipeline, which has been

extensively tested in the community-wide CASP and CAFA experiments.[190,282,284] The prediction accuracy of the pipeline was examined on 100 randomly-selected well-characterized proteins from this chromosome, and achieved a high F-measures of 0.69, 0.57, and 0.67 for MF, BP, and CC aspects of GO term predictions, respectively. The structure-based function prediction component of this pipeline is the main contributor of prediction accuracy for the non-homologous protein targets. Applying the pipeline on all of the 66 poorly- or non-characterized uPE1 proteins coded by genes on chromosome 17, we are able to infer the specific biological function with high confidence for 13, 33, and 49 uPE1 proteins for MF, BP, and CC aspects, respectively. The majority of these function inferences could not be achieved using traditional sequence-based function annotation approaches. We give extensive details for the 13 highest-rated predictions for Molecular Functions, plus structural findings for 5 case studies.

As a proof-of-concept, we started with the set of 66 uPE1 proteins on human chromosome 17 only. The pipeline can be readily extended to all 1260 uPE1 proteins from the entire human proteome, as well as 677 additional unannotated human proteins in neXtProt categories PE2, PE3, and PE4 (https://www.nextprot.org/proteins/search?mode=advanced&queryId=NXQ_00022). The work along this line is in progress.

We hope our modeling results will stimulate the interest of molecular and cell biologists and assist them to design appropriate experiments that could validate the computational predictions and, more importantly, elucidate the structure and biological function of these proteins in human tissues and cells. To assist investigators, neXtProt has already introduced links to pre-computed and annually updated I-TASSER/COFACTOR predictions for proteins lacking function annotation as illustrated for JMJD7 (NX_P0C870) at

https://www.nextprot.org/entry/NX_P0C870/gh/zhanglabs/COFACTOR, where "NX_P0C870" can be replaced by neXtProt ID for the target of interest.

Chapter 6 Functions of Essential Genes and a Scale-free Protein Interaction Network Revealed by Structure-based Function and Interaction Prediction for a Minimal Genome[1]

6.1 Introduction

The question of what set of functionalities constitutes the minimal set necessary to enable life is one of the most important unanswered questions of contemporary biology [322-324]. While even the question of what constitutes "life" carries a vast range of philosophical difficulties [325,326], for the present purposes we define a living thing as an entity consisting of one or more membrane-bound cells, capable of separating itself from its surroundings, drawing energy from its environment, and using that energy to maintain (and possibly reproduce) itself. As the simplest organisms meeting this definition will be unicellular, and in all known cases such organisms make use of a DNA genome, investigations into the minimal basis for life have almost invariably focused on determining the minimal set of genetic components required to yield a living cell. Studies based on transposon knockout libraries or high-throughput targeted deletions substantially enhanced our ability to rationally design reduced genomes, by providing a high-throughput approach for identifying all genes that could not be individually knocked out [327-333]. Such knockout libraries cannot, however, provide all needed information for construction of a minimal genome, due to the presence of both positive and negative epistatic interactions that cannot be captured in a single pass using such approaches [324,328]. More targeted work [334] provided a window into the overall reducibility of microbial genomes by deleting all prophages

[1] This chapter was adapted from a manuscript under review by *Journal of Proteome Research*, entitled "Functions of Essential Genes and a Scale-free Protein Interaction Network Revealed by Structure-based Function and Interaction Prediction for a Minimal Genome" by C Zhang, W Zheng, M Cheng, GS Omenn, PL Freddolino, and Yang Zhang.

and mobile genetic elements from *E. coli* MG1655, yielding a genome that was reduced in size by ~15%; the reduced genome strain, MDS42, also showed several useful properties such as increased stability of cloned genes [335,336]. A new level of capability in the study of minimal genomes was achieved with the development of JCVI-syn1.0, a completely synthetic *Mycoplasma mycoides* derivative [337]. The subsequent inclusion of repeated cycles of transposon mutagenesis and a "design-build-test" cycle permitted comprehensive mapping of the genes that could not be complemented by any other gene in the original *Mycoplasma mycoides* genome, which we refer to as "essential". The cyclical genome reduction efforts described above yielded a well-defined list of 465 effectively essential genes for a minimal *Mycoplasma*, 438 of which encode proteins. The resulting organism, syn3.0, has a genome reduced in size by nearly 50%, and shows substantial differences in growth and cellular morphology from the *M. mycoides* parental strain [338], including a reduced growth rate, reduced colony sizes, and a filamentous and highly heterogeneous cellular morphology.

Simply knowing the identities of all genes needed in a minimal genome, however, does not permit resolution of the fundamental question of what functionalities are needed in a minimal cell. Upon the initial construction of syn3.0, researchers noted that ~⅓ of the protein coding genes in its genome could not be annotated by sequence homologs from characterized protein domain families[338]; more recent efforts to enable a complete metabolic reconstruction of syn3.0 still cannot assign a protein to all functions necessary in a minimal metabolic model [339]. Initial efforts to determine the functions and biological roles of the remainder of the syn3.0 proteome were based on sequence-based annotations and sequence-profile based protein family assignment [338,340], which have limited sensitivity when there are no close homology templates for annotation transfer. Later, Yang and Tsui attempted to annotate syn3.0 proteins by secondary structure

matching [341], which is developed to recognize templates with similar structure fold but not necessarily of related function. More recently, Antczak and colleagues applied a multi-pipeline approach to provide consensus predictions that added functional information for 66 of the proteins of unknown function in syn3.0, demonstrating a particular abundance of putative transporters and other transmembrane proteins [342].

We have recently shown that the inclusion of protein structural information, even from computationally predicted structures, can substantially enhance the accuracy of function predictions for difficult annotation targets [201,343]. To this end, we developed an I-TASSER/COFACTOR-based protocol that performs I-TASSER structure prediction followed by COFACTOR structure-based function annotation [303]. This pipeline has been shown to accurately assign functions for many proteins in microbes [201] and in humans [344], and is among the top predictors in the most recent Critical Assessment of Function Annotation round 3 (CAFA3) and CAFA PI competitions [304]. Moreover, the recent development of sequence-derived residue-residue contact prediction algorithms based on deep neural networks [232,345] has greatly enhanced the accuracy of protein structure assembly, which should in principle enhance the effectiveness of structure-based protein function prediction.

To have a complete understanding of the essential syn3.0 proteome, we developed and applied an enhanced C-I-TASSER/COFACTOR pipeline by the combination of contact map-based protein structure simulations with structure-based protein function annotation and protein-protein interaction (PPI) predictions. We found that high-confidence Molecular Function (MF) and Biological Process (BP) annotations from Gene Ontology (GO) can be provided for 86% and 88% of the syn3.0 proteome, respectively, while the utilization of deep neural-network contact-map information shows significant enhancements of both coverage and accuracy of protein

structure and functional models. Functions related to nutrient acquisition, microbe-host interactions, and nucleotide metabolism are enriched among the set of previously unannotated genes, likely indicating important and as-yet unresolved portions of syn3.0 physiology. Viewed at the level of the whole-cell protein-protein interaction network, we further note that the PPI network of syn3.0 follows the scale-free network architecture often noted in natural PPIs, but rare in randomly formed networks, suggesting that scale-free layouts persist even when an original, natural PPI network is artificially reduced to a minimal, essential form of itself.

6.2 Methods

6.2.1 Protein structure prediction

Structure models of all 438 proteins in the syn3.0 genome were predicted by C-I-TASSER [238], our most recent template-based protein structure prediction pipeline based on the I-TASSER structural assembly protocol [46] combined with deep learning-based residue-residue contact map predictions [232,345]. Briefly, C-I-TASSER first uses DeepMSA [240] to search the query protein sequence against three whole-genome and metagenome protein sequence databases, including Uniclust30 [222], UniRef90 [223], and Metaclust [185], to obtain a multiple sequence alignment (MSA). Next, residue-residue contacts are predicted from the MSA by the deep learning-based algorithms TripletRes/ResTriplet [232] and ResPRE [345]. Meanwhile, LOMETS threading [181] is performed to search the query protein sequence against the PDB database to align the query to template structures to extract continuous fragments. These fragments are finally assembled into the full length structures by a replica-exchange Monte Carlo (REMC) simulation, under the guidance of a composite force field consisting of the deep learning-predicted contacts, template-derived distance restraints, and knowledge-based energy terms calculated based on

statistics of PDB structures. The REMC simulation produces tens of thousands of "decoy" conformations, which are clustered by pairwise structure similarity [244]. The centroid of the largest cluster is refined at the atomic level [188] to obtain the final C-I-TASSER model.

As a control experiment to study the impact of deep learning predicted contacts on structure and function prediction, we also performed structure prediction for the same set of 438 proteins using the classical I-TASSER pipeline without contact prediction. Structure-based function annotations were separately performed for the top-ranked models produced by C-I-TASSER and I-TASSER for the same target protein, as detailed below.

6.2.2 Estimation of structure model quality

The global quality of structural models can be assessed by TM-score [247] between modeled and native structures of the target protein:

$$TM = \frac{1}{L}\sum_{i=1}^{L_{ali}} \frac{1}{1+(d_i/d_0)^2} \qquad (6.1)$$

where L is the number of residues in the target, d_i is the distance between the ith aligned residue pair, and $d_0 = 1.24\sqrt[3]{L-15} - 1.8$ is a length-dependent scaling factor. TM-score ranges between 0 and 1, with TM-score>0.5 meaning structure models of correct global topology [246].

Since the native structures of syn3.0 proteins are not available, we estimate the TM-score (eTM) of the C-I-TASSER models using a combination of threading alignment quality, contact satisfaction rate, and convergence of the structure assembly simulations:

$$eTM = c_0 + c_1 \cdot C + c_2 \cdot C^2 \qquad (6.2)$$

where the confidence score (C) is defined as:

$$C = w_1 \cdot \ln\left(\frac{M}{M_{total}} \cdot \frac{1}{<RMSD>}\right) + w_2 \cdot \sum_m \ln\left(\frac{Z(m)}{Z_0(m)}\right) + w_3 \cdot \ln\left(\frac{O(CM^{model}, CM^{pred})}{N(CM^{pred})}\right) \qquad (6.3)$$

c_0=0.79, c_1=0.1077, c_2=0.00098, w_1=0.77, w_2=1.36, and w_3=0.67 are free parameters obtained by maximizing the correlation between the estimated and actual TM-score on a separate set of 797 training protein domain structures from SCOPe database 40 version 2.06. M_{total} is the total number of decoy conformations used for clustering, while M is the number of decoys in the top cluster. $<RMSD>$ is the average RMSD among decoys in the same cluster. $Z(m)$ is the score of the top template by the mth threading method in LOMETS. $Z_0(m)$ is a cutoff above which templates are considered reliable. $N(CM^{pred})$ is the number of contacts predicted by deep learning and used for guiding the REMC simulation, while $O(CM^{native},CM^{pred})$ is the number of common contacts between the final model and the deep learning predicted contacts. For the (non-contact based) I-TASSER predicted structures, the estimated TM-score is calculated similarly, but with c_0=0.71, c_1=0.1300, c_2=0.00060, w_1=w_2=1, and w_3=0. The estimated TM-score was shown to highly correlate with actual TM-score, with a Pearson Correlation Coefficient (PCC) 0.91 on 300 test proteins that are non-homologous to the training proteins of I-TASSER [114].

6.2.3 Function annotation and enrichment analysis

Protein functions are predicted from the structure models by COFACTOR [201], which combines models from three complementary submodules based on structure, sequence, and PPI. In the structure-based submodule, the (C-)I-TASSER model is structurally aligned to function templates in the BioLiP database [346], where function annotations are obtained from the function templates identified by global and local structure similarity. In the sequence-based submodule, BLAST and PSI-BLAST [102] are used to search the query sequence against the UniProt Gene Ontology Annotation (UniProt-GOA) database [347] to obtain annotations from sequence homologs. Finally, the PPI-based submodule is ported from MetaGO [343], where the query

sequence is mapped to the PPI network of STRING [348], with the immediate neighbor (i.e. direct PPI partner) of the query searched against UniProt-GOA for function transfer. Function predictions from these three submodules are combined by weighted averaging to obtain the final prediction. Each predicted function has a confidence score (C-score^{Func}) ranging from 0 to 1, with C-score^{Func}>0.5 corresponding to a confident function prediction [201,344]. While COFACTOR predicts three categories of protein functions, namely Enzyme Commission (EC) numbers, Gene Ontology (GO) terms, and ligand binding sites (LBS), we do not separately discuss prediction of EC numbers because they can be mapped to MF GO terms [349].

Enrichment of GO terms in previously unannotated syn3.0 proteins (versus proteins with previous UniProt free-text annotation or UniProt-GOA GO term annotations) are quantified by a rate ratio test approach [350]. Briefly, for each GO term q, we compute the annotation rate (i.e. the number of proteins annotated with q divided by the total number of proteins) among UniProt-unannotated proteins, and that among UniProt-annotated proteins. We then test whether the ratio of the two rates is significantly different from 1. Some GO terms, such as GO:0005515 "protein binding", are too generic to suggest any specific function. Therefore, similar to our prior study [303], we discard any GO terms associated with >10% of annotated proteins in all steps of our analysis, including the definition of previously unannotated/annotated proteins and the rate ratio test of GO term enrichment.

6.2.4 PPI prediction

The PPI network of syn3.0 was predicted using the SPRING [351] dimer threading program. For a pair of query proteins, SPRING first searches the sequence of each protein chain to a monomeric template structure database by HHsearch [210]. The HHsearch aligned monomeric

templates are then structurally aligned to complexes in the PDB dimer template database by TM-align [227] to obtain the dimeric complex model. The final score of the dimeric complexes, SPRING-score, is a linear combination of three terms: the Z-score for HHsearch monomeric threading, TM-score of monomer-to-dimer structure alignment by TM-align, and a statistical energy potential for the dimer interface. The two query proteins are considered to interact with each other if there is a good complex hit with SPRING-score >2 and both of the monomer threading Z-scores >-2. The Z-score and SPRING-score cutoffs were trained to optimize the Matthews correlation coefficient (MCC) of classifying interacting versus non-interacting protein pairs, on a dataset consisting of 1,732 structurally characterized PPI pairs from the SPRING dimer template database and 4,117 pairs of non-interacting proteins from the Database of Interacting Proteins (DIP) [352]. Only hetero-dimeric interactions are considered in this study.

6.2.5 Data Availability

Protein sequences of syn3.0 were collected from NCBI accession CP014940.1. While the genome consists of 473 genes, this study only considered the 438 protein coding genes, as the other 35 genes encode non-coding RNAs with well-known functions such as tRNAs and rRNAs. The syn3.0 proteins are mapped to the closest UniProt 2019_09 entries from Mycoplasma mycoides reference proteomes UP000001016 and UP000011126. The GO annotations of these UniProt entries are collected from UniProt-GOA release 2019-09-17. All predicted structure models, functions and interactions are available at our public webserver at https://zhanglab.ccmb.med.umich.edu/C-I-TASSER/JCVI-syn3.0/, including a one sentence description of protein function generated using the most specific high confidence predicted GO term.

When we performed this study, a new version of the minimal genome, JCVI-syn3A (NCBI accession CP016816.2), was published [339], which includes 16 additional protein coding genes not included in the JCVI-syn3.0 genome. Although these new genes are not essential for the survival of the cell, they make the cell less fragile and have a more stable cellular morphology. For completeness, we have included these 16 new genes in our structure and function prediction as part of our online webserver, even though our main analysis focuses on the original JCVI-syn3.0 genome where all the genes are essential. To facilitate comparative study between JCVI-syn3.0 and JCVI-syn3A, the webserver displays the protein names and accessions for both genomes.

6.3 Results

6.3.1 Contact-assisted protein structure prediction and structure-based function prediction increase the coverage of function annotation

We began by investigating how many syn3.0 proteins can be assigned specific gene ontology (GO) term annotations, which were categorized by the original syn3.0 study 17 into 5 classes (Unknown, Generic, Putative, Probable, and Equivalog) in ascending order of function annotation confidence, based on a protein's match to TIGRfam protein family database [353]. Specifically, Unknown or Generic proteins lack functional homologs or do not have homologs with consistent function annotations, while Putative, Probably or Equivalog proteins can match homologous proteins with related functions in the same family. As shown in Figure 34A-E, for all five classes, the numbers of proteins for which GO terms can be assigned by the structure-based function annotation pipeline C-I-TASSER/COFACTOR are consistently greater than those in UniProt. Here, the UniProt terms in Figure 34A-E refer to the GO annotations from the

UniProt-GOA project [347]; all UniProt terms for the syn3.0 proteins in our study are from computational approaches such as UniRule and InterProScan [354] with evidence codes “Inferred from Electronic Annotation” (IEA) and “Inferred from Sequence or structural Similarity” (ISS). It is therefore fair to compare the coverage (i.e. the percentage of proteins that can be annotated) between UniProt annotations and C-I-TASSER/COFACTOR annotations, as both are computationally predicted GO terms. The broader coverage of C-I-TASSER/COFACTOR is particularly evident for the Unknown and Generic categories, which are considered uncharacterized in the original syn3.0 study 17. For example, C-I-TASSER/COFACTOR can annotate 49% and 45% of all Unknown proteins with specific MF and BP terms, respectively, which are 9 times more than UniProt for the same set of proteins (5% for both MF and BP) (Figure 34A). In both C-I-TASSER/COFACTOR and UniProt GO annotations, the number of proteins with specific Cellular Component (CC) terms is smaller than those with MF or BP terms. This is partly due to the simple cellular structure of syn3.0 (which has a single cell membrane and no cell wall or membrane-bound organelles), where most proteins localize to the cytoplasm or plasma membrane, instead of more specific subcellular locations.

The high sensitivity of our C-I-TASSER/COFACTOR pipeline can be attributed partly to the use of deep learning predicted contact maps in the structure predictions. Indeed, the confidence score of COFACTOR GO term prediction is consistently improved by using structure models from contact-assisted C-I-TASSER over the traditional I-TASSER approach for all three aspects of GO terms (Figure 34F-H). Accordingly, the quality of C-I-TASSER structure models in terms of average estimated TM-score (0.76) 38 is 8.6% higher than that of I-TASSER (0.70); 328 of the 434 proteins (76%) are estimated to have better structure model quality in C-I-TASSER than in I-TASSER (Figure 34I).

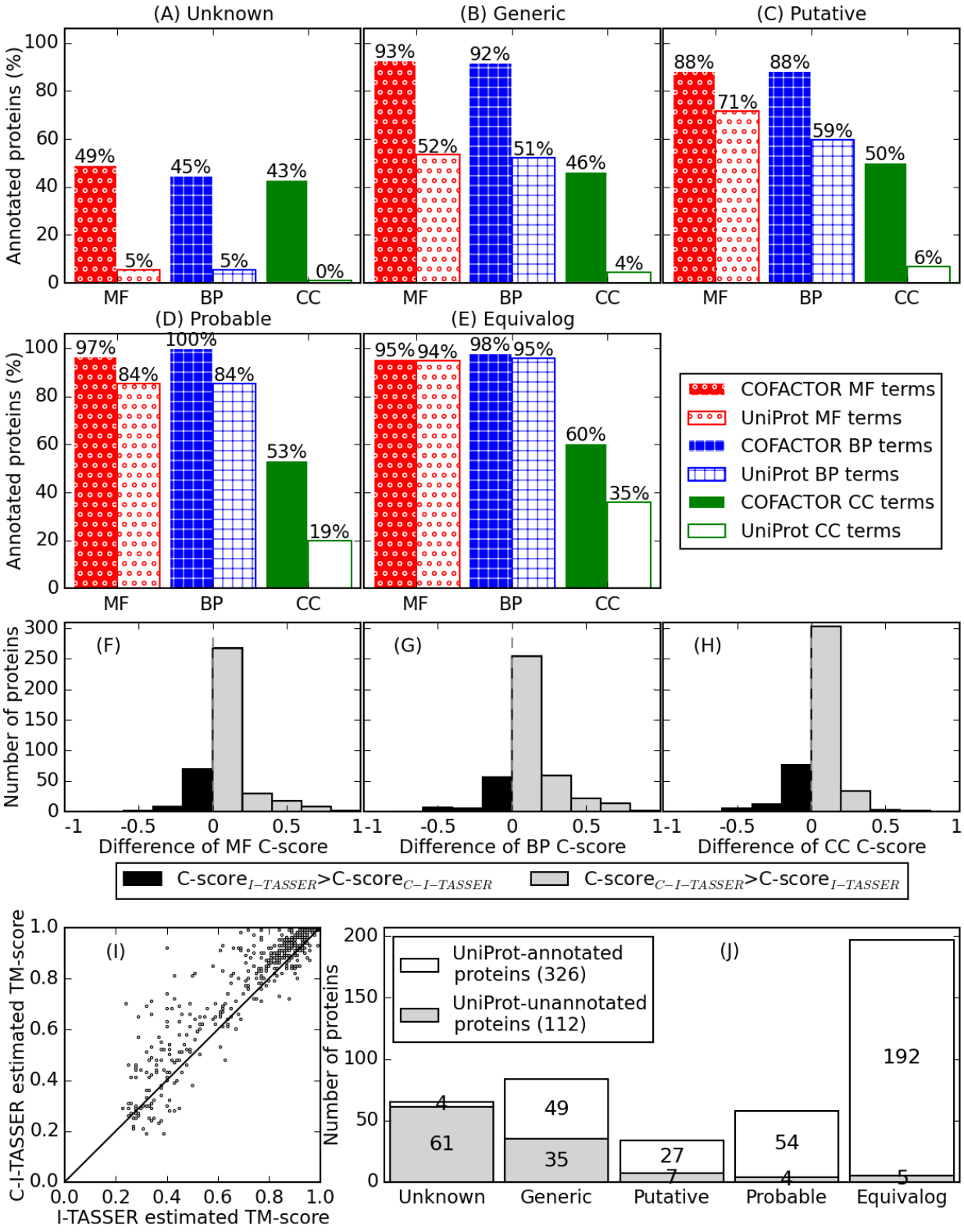


Figure 34. C-I-TASSER/COFACTOR improves coverage of protein function prediction (i.e. percentage of proteins with predicted function) for syn3.0. (A-E) Percentage of proteins that can be annotated with GO terms by C-I-TASSER/COFACTOR and by UniProt for the five categories of syn3.0 proteins classified in the original syn3.0 report, where "unknown" (A) and "generic" (B) proteins were considered unannotated. (F-H) Distribution of difference in confidence scores (C-scores) for COFACTOR GO term prediction using C-I-TASSER models

compared to those using I-TASSER models. For each protein, only GO terms predicted with C-score>0.5 in at least one of C-I-TASSER/COFACTOR and I-TASSER/COFACTOR are considered, and the average C-score difference for using C-I-TASSER compared to using I-TASSER for each protein is shown on the x-axis. The average C-score differences in structure-based GO term prediction using C-I-TASSER versus that using I-TASSER are +0.07, +0.11, and +0.06 for MF (F), BP (G), and CC (H), respectively. (I) Per-target comparison of estimated TM-score between I-TASSER (x-axis) and C-I-TASSER (y-axis). Points on the upper left triangle correspond to targets with better estimated quality in C-I-TASSER than in I-TASSER. J. Number of proteins with (white) and without (grey) function annotation (GO terms or free-text) in the five categories of syn3.0 proteins.

Despite the high sensitivity of the C-I-TASSER/COFACTOR pipeline, there are still 14% and 12% of the syn3.0 proteins that cannot be annotated with specific MF and BP terms, respectively, partly due to the high transmembrane contents for the targets (Figure 35), making them more difficult for experimental characterization and computational annotation.

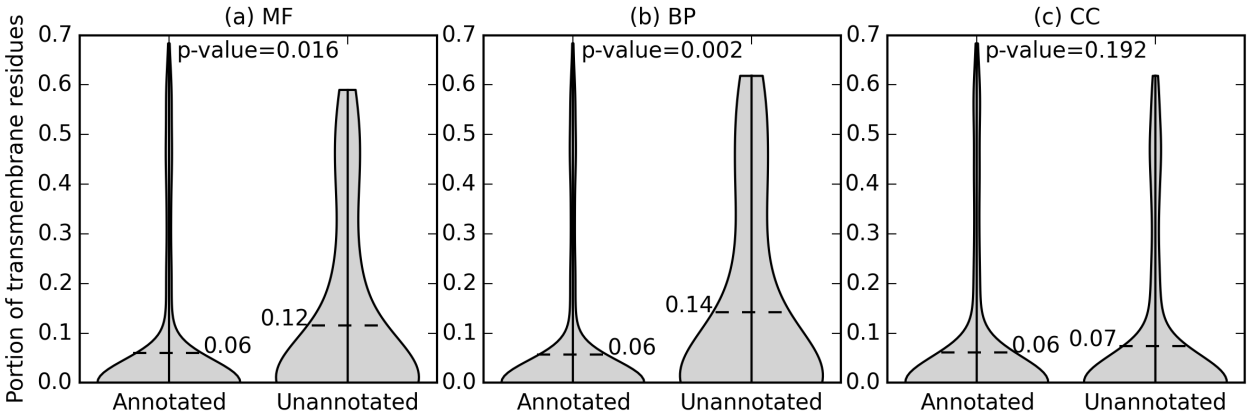


Figure 35. Violin plots for portions of residues predicted by TMHMM2.0 to be within transmembrane helices (y-axis) for JCVI-syn3.0 proteins that are annotated (left) versus unannotated (right) by C-I-TASSER/COFACTOR with C-score>0.5 for specific GO terms in the MF (A), BP (B) and CC (C) aspects. The p-value is calculated by single-tailed unpaired t-test to test if the average portion of transmembrane residues (dashed lines) for C-I-TASSER/COFACTOR annotated proteins is significantly smaller than that for unannotated proteins.

The original method for partitioning syn3.0 protein annotation status into 5 categories may not be sufficiently specific, as a protein not belonging to a characterized TIGRfam protein family can still be individually annotated. Thus, we re-classified annotated versus unannotated proteins based on whether their respective UniProt Gene Ontology Annotation (UniProt-GOA) [347] entries in the Mycoplasma mycoides proteome have specific GO term annotations, excluding overly general GO terms such as “protein binding” (see Methods). As shown in Figure 34J, 112

(26%) of the 438 proteins in syn3.0 are unannotated based on their UniProt entries. This is smaller than the number of proteins with unknown function (149 of 438 proteins) reported in previous studies [338,342], as some proteins previously reported to have unknown functions are now annotated as of UniProt release 2019_09. These inconsistencies could have resulted from either the difference in classifying annotated versus unannotated proteins, the recent improvement of the annotation pipeline used in UniProt, or both. For the sake of consistency with contemporary work [308], in later sections we use the term "unannotated proteins" to refer to proteins without UniProt annotation, regardless of their TIGRfam match.

6.3.2 Functions enriched in uncharacterized proteins highlight the dependency of syn3.0 on the environment

To obtain a more nearly complete understanding of the metabolism of syn3.0 and the nature of the required genes that it encodes, we applied a rate-ratio test approach (see Methods for details) to search for the GO terms that were enriched among previously unannotated proteins. Compared to previously annotated proteins, UniProt unannotated proteins are enriched for transporter activity and phosphatase activity for MF, and multi-cellular response for BP (Figure 36). This is consistent with a previous study that proposed some of the poorly characterized syn3.0 proteins are transporters [342]. Among the newly annotated proteins with "phosphatase activity" annotations, furthermore, at least half appear likely to act on nucleotide substrates, suggesting a particularly important role for these poorly annotated nucleotide phosphatases in syn3.0 for either signal transduction or metabolism. As case studies demonstrating the new information provided by the C-I-TASSER/COFACTOR pipeline, we select MMSYN1_0877 and MMSYN1_0440 (Figure 37) to discuss the derivation and

implication of their predicted functions “vitamin transporter” activity and “response to other organism”, respectively, which are the most significantly enriched terms for MF and BP, respectively.

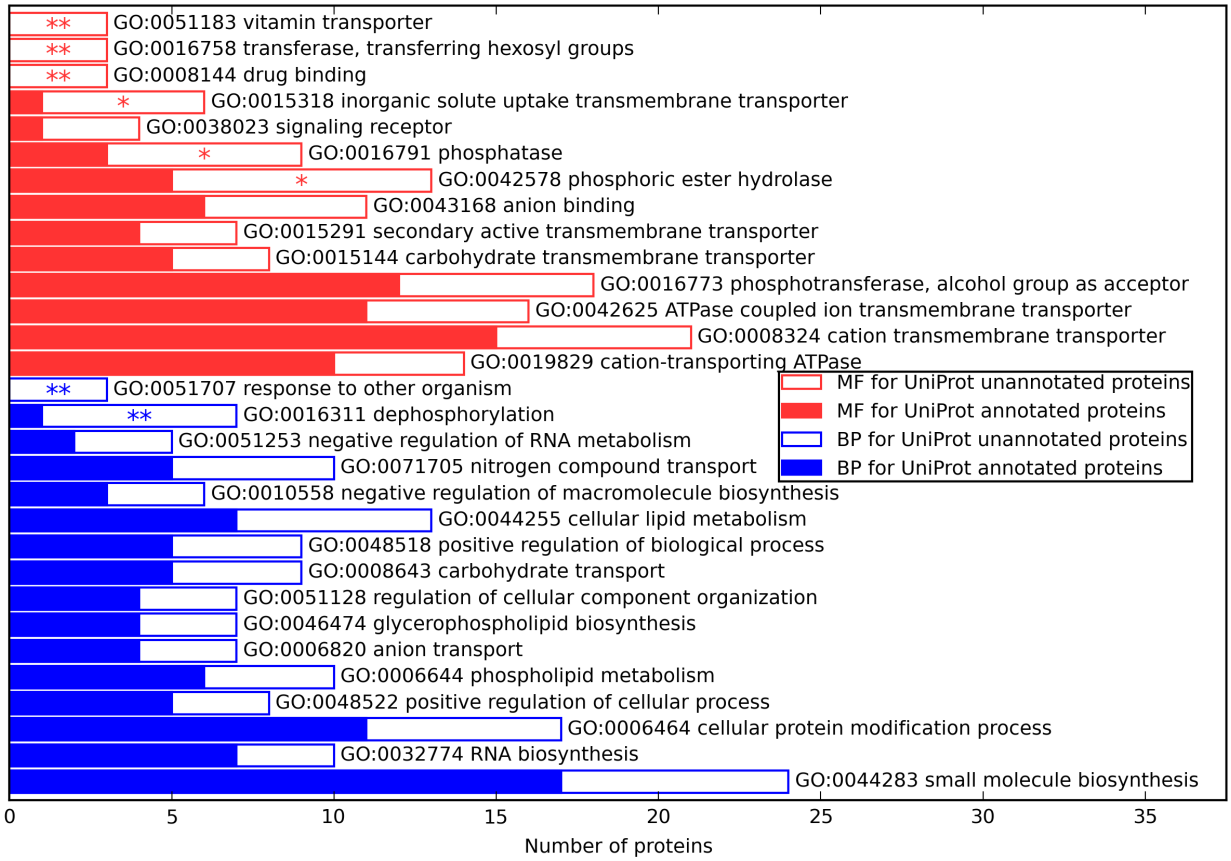


Figure 36. Enrichment of MF (upper half) and BP (lower half) GO terms predicted by C-I-TASSER/COFACTOR in proteins of unknown function (empty bars), compared to proteins of known function (solid bars). One asterisk is shown for significant enrichment of a GO term in the unknown function set ($p<0.05$ by rate ratio test) and two asterisks for significant enrichment after adjusting for multiple testing ($p<0.05$ with FDR correction). GO terms are ranked in descending order of ratio of annotations rate of a GO term in unannotated proteins versus that in annotated proteins.

Riboflavin transporter MMSYN1_0877

MMSYN1_0877 (Figure 37A) is an unannotated protein predicted to have “riboflavin transporter activity” and “vitamin transporter activity” with C-score=0.82 for MF by the C-I-TASSER/COFACTOR pipeline. The C-I-TASSER structure model exhibits a multi-pass transmembrane helix bundle with an estimated TM-score of 0.59 (indicating correct topological

fold [246]), with a riboflavin (i.e., vitamin B2) ligand recognized by COFACTOR. The protein is structurally similar to RibU, a Riboflavin uptake protein from *Staphylococcus aureus*, with TM-score=0.72 by TM-align 50. The presence of this putative transporter suggests that syn3.0 relies upon riboflavin uptake from the media for survival. Indeed, we find that *M. mycoides* have two Riboflavin kinase/FAD synthetase enzymes, ribC (UniProt ID: Q6MUC6) and ribF (UniProt ID: Q6MTQ9), which can make use of riboflavin to synthetize flavin mononucleotide or flavin adenine dinucleotide. However, *M. mycoides* lacks an identifiable pathway for de novo riboflavin biosynthesis, and thus presumably relies on uptake from the host or media (presumably via UniProt ID Q6MS70, the homolog of MMSYN1_0877). In the case of syn3.0, the ribC gene is also absent, apparently leaving riboflavin import via MMSYN1_0877 followed by RibF processing as the likely sole path for synthesis of riboflavin-containing compounds. The current lack of annotation of the *M. mycoides* homolog Q6MS70 is likely because our annotation prediction builds strongly on structural similarity to ECF-type riboflavin uptake proteins from *T. maratima* [355] and *S. aureus* [356], which have sub-2 Å RMSDs to the predicted MMSYN1_0877 structure, but amino acid sequence identities less than 22%.

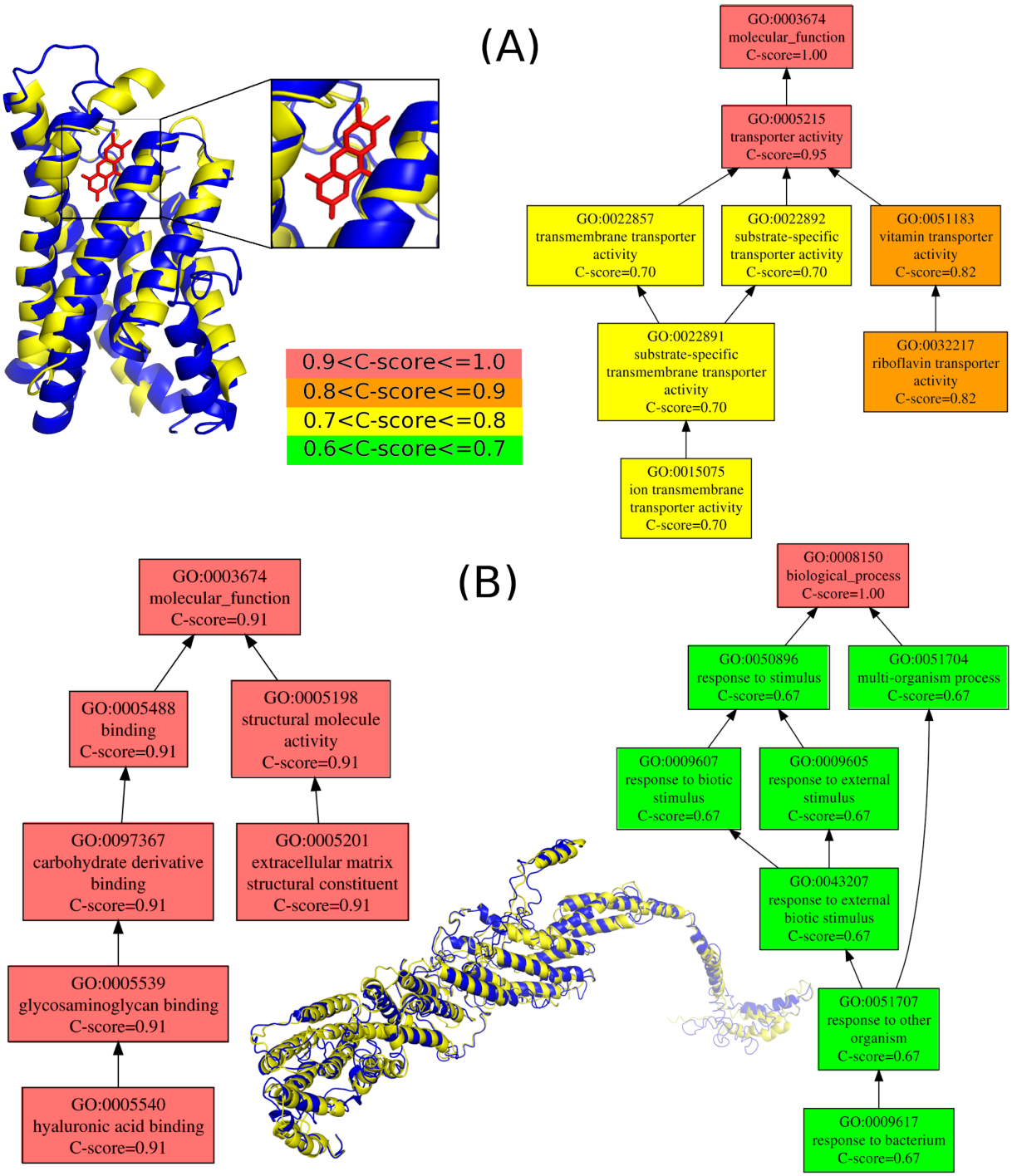


Figure 37. Exemplar proteins corresponding to GO terms that are highly abundant among the newly annotated set. (A) MMSYN1_0877, a protein with predicted "vitamin transporter" activity, and (B) MMSYN1_0440, a protein with predicted annotation of the "response to other organisms" GO term. (A) C-I-TASSER structure model (deep blue, estimated TM-score=0.59) of MMSYN1_0877 (NCBI accession: AMW76711.1) superposed to S. aureus riboflavin transporter RibU (light yellow, PDB ID: 3p5n chain A, TM-score=0.72) in complex with riboflavin (red stick). Top MF GO term predictions are shown on the right hand side directed acyclic graph, with different colors representing different ranges of COFACTOR C-scores for the predicted terms (center color map). (B) C-I-TASSER model (deep blue, estimated TM-score=0.33) of MMSYN1_0440 (NCBI accession: AMW76515.1) superposed to yeast exocyst complex component SEC8 (light yellow, PDB ID: 5yfp chain D with TM-score=0.84 but sequence identity 0.1). Top predicted MF and BP terms are shown in graphs on the left and right, respectively.

Hyaluronic acid binding protein MMSYN1_0440

Considering that syn3.0 can be cultured in vitro without the need to interact with other organisms, it is initially counter-intuitive that we observe several new annotations of the GO term "response to other organism". However, it must be noted that the ancestral *M. mycoides* is an obligate parasite of animal hosts, and the culture media used for syn3.0 contains a broad range of animal derivatives (beef heart infusion, peptones, and fetal bovine serum [338]); it is thus plausible that syn3.0 interacts with animal-derived media components for regulatory or mechanical purposes as well as nutritional purposes. As an example, the protein MMSYN1_0440 (Figure 37B) is predicted to be involved in "response to other organism" with C-score=0.57 for BP. This is substantiated by the predicted MF term "hyaluronic acid binding" with C-score=0.91, indicating likely interaction with animal-derived hyaluronic acid present in the culture media. The reason for the importance of this particular interaction for the viability of syn3.0 is not immediately clear. One possibility arises from the MMSYN1_0440 structural model, which shows good structural similarity to the yeast membrane tethering protein SEC8; MMSYN1_0440 may play an architectural role in maintaining membrane integrity or cell-cell contacts in syn3.0, likely interacting with hyaluronic acid polymers present in the media.

6.3.3 Whole-proteome dimeric threading reveals a scale-free PPI network

Given that many proteins perform their function by interacting with other proteins, we used SPRING, a dimeric threading approach 48, to investigate the organization of pairwise PPIs in the syn3.0 proteome. The interactome predicted by whole-proteome SPRING threading search is relatively sparse, with only 2.6% (2483) of all 95,703 protein pairs being predicted PPI

partners (Figure 38A). We initially speculated that, due to its simplicity, syn3.0 network structure might revert to a less ordered state instead of a scale-free layout typical of bacterial networks 57. However, we found that the PPI network is actually scale-free: $P(k)$, the fraction of proteins in the network having k partners, follows a power law distribution:

$$P(k) \sim k^{-\tau} \qquad (6.4)$$

A high goodness-of-fit is achieved with the parameter τ=1.40, resulting in the reduced chi-squared statistics and the coefficient of determination and approaching 0 and 1, respectively (Figure 38BC). This is significantly different from a randomly generated PPI network with the same number of positive (2483) and total (95703) protein pairs (Figure 39), where the number of PPI partners per protein fits poorly to the power law with the reduced chi-squared statistics and the coefficient of determination consistently greater than 1.5 and less than 0, respectively. This suggests that the scale-freeness of the SPRING-predicted PPI network is not coincidental. Scale-free networks were reported previously for naturally evolved biological networks: *E. coli*, for example, also has a scale-free PPI network [357] with τ=1.3 as estimated by our recent work (Gong W, Guerler A, Zhang C, et al. Submitted). On the other hand, the present study is the first time that a scale-free PPI network is observed for an artificial proteome, although genes are retained in the syn3.0 genome based solely on their essentiality without explicit consideration for the number of potential PPI. The unintentional retention of a scale-free PPI network in the deeply truncated syn3.0 proteome suggests the universal robustness of PPI network architecture, and the importance of the "hub" proteins (which regulate a large number of proteins with few PPI) for the overall viability of cells.

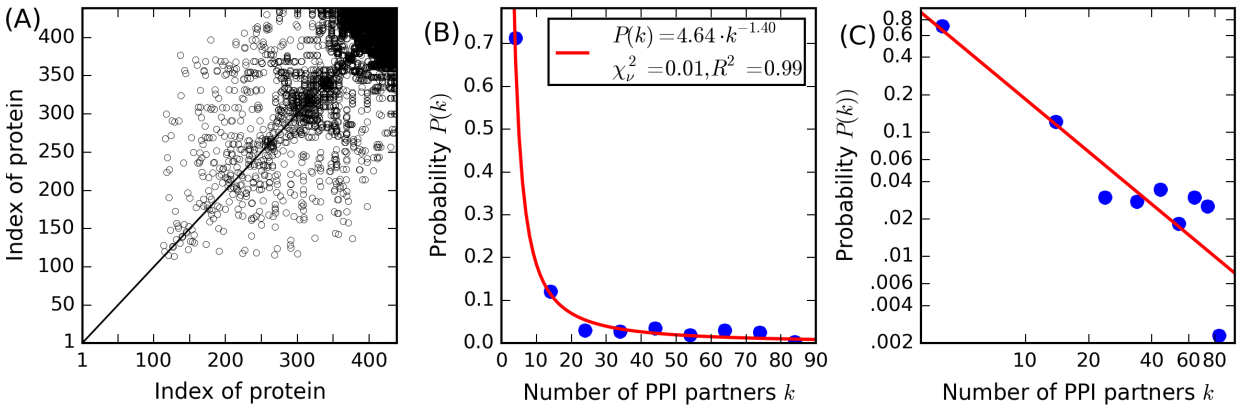


Figure 38. PPI predicted by SPRING. (A) Scatter plot of protein-protein interactions for all syn3.0 proteins ranked in ascending number of PPI partners, where a point means the protein pair is predicted to have a PPI. (B-C) Observed distribution (circles) for the number of PPI partners per protein in linear (B) and log (C) scale, and the power law fit (lines). In the inset, $\boldsymbol{\chi^2_\nu}$ is the reduced chi-squared statistic (lower values are better, with 0 being a perfect fit) and R^2 is the coefficient of determination (the higher the better, with 1 being a perfect fit), respectively, to quantify the goodness of fit.

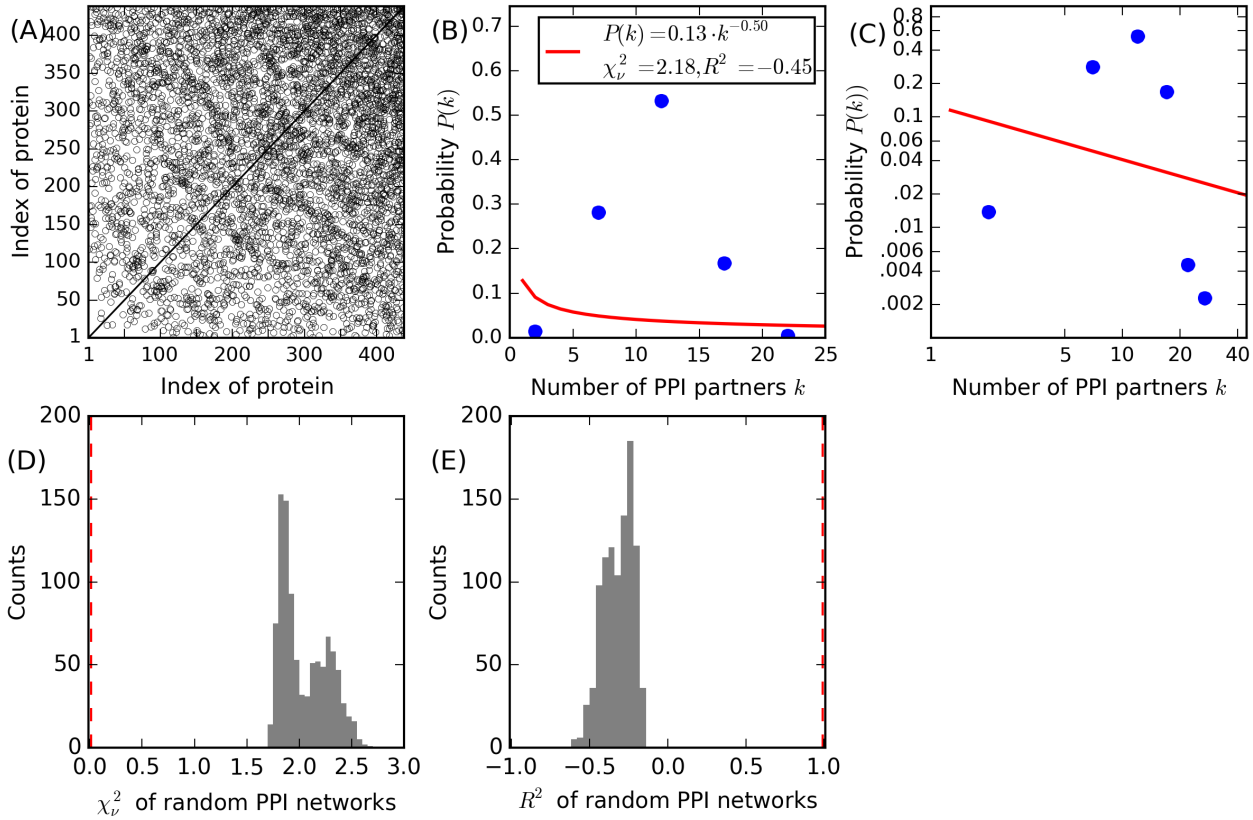


Figure 39. A random PPI network for syn3.0, where 2483 of all 95703 protein pairs are randomly selected as the positive PPI pairs. The number of positive pairs in this random network is therefore identical to the SPRING-predicted PPI network shown in Figure 38. (A) Scatter plot of PPIs for all syn3.0 proteins ranked in ascending number of PPI partners, where a point means the protein pair is predicted to have a PPI. (B-C) Observed distribution (circles) for the number of PPI partners per protein in linear (B) and log (C) scale, and the power law fit (lines). In the inset, $\boldsymbol{\chi^2_\nu}$ is the reduced chi-squared statistic (lower values are better, with 0 being a perfect fit) and R^2 is the coefficient of determination (the higher the better, with 1 being a perfect fit), respectively, to quantify the goodness of fit. Both metrics indicate that power law fits poorly to the distribution of the number of PPI partners per protein. (D-E) Histogram of $\boldsymbol{\chi^2_\nu}$ (D) and R^2 (E) values of 1000 randomly generated PPI networks for syn3.0 with the same

number of positive pairs as the SPRING-predicted network. The vertical dash lines to the left (D) or right (E) of the histograms indicates χ_v^2=0.01 and R^2=0.99, respectively, in the SPRING-predicted network (Figure 38B), which is consistently better fitted to a power law distribution to all 1000 randomly generated PPI networks.

6.4 Discussion and Conclusion

In this study, we extended a unified structure and function prediction pipeline for whole-genome function and PPI modeling of the syn3.0 minimal genome. This pipeline is able to assign function for 9 times more unknown proteins than existing UniProt annotations (Figure 34A), and substantially extends the reach of structure-based function prediction of poorly annotated proteins. These results further demonstrated the usefulness and impact of high-resolution protein structure simulations on large-scale proteome function annotations. In particular, the integration of deep neural network-based contact maps with the structural assembly simulations plays an essential role for not only improving the quality of structure models, but also increasing the coverage and reliability of functional predictions. We expect that the approach employed here will be of substantial utility for providing optimal computational structure/function predictions for other organisms, which are currently under progress in our laboratories.

The annotation efforts detailed here also provide a substantial boost to our ability to understand the biology of the reduced-genome syn3.0 strain, providing confident MF and BP models for 373 and 382 syn3.0 proteins, which represent, respectively, 86% and 88% of the proteome that were previously unannotated. Consistent with the findings of Antczak et al. [342], the spectrum of function annotations for these newly annotated proteins (Figure 36) places a strong emphasis on the importance of nutrient acquisition, demonstrating a broad range of uptake and metabolic pathways that had previously not been appreciated. Regulatory proteins comprise a substantial additional category of previously unannotated syn3.0 genes, with roles ranging from signaling receptors to nucleotide phosphatases (the latter of which likely play a role in second

messenger signaling, but may also be involved in nutrient assimilation). The importance of interactions with host tissue and host-derived molecules (including those present in the heavily animal-sourced syn3.0 growth media) is a common thread running throughout the newly identified annotations, ranging from uptake of host-derived nutrients (e.g., the riboflavin transporter shown in (Figure 37A) to architectural proteins binding host-derived glycans (Figure 37B). In the ongoing quest to develop a truly minimal genome, it will be intriguing to determine which of the syn3.0 genes represent simple metabolite uptake requirements (e.g., MMSYN1_0877) and which involve detection of host-derived substances that act as growth stimulators (as may be the case for some of the newly-annotated proteins bearing the "signaling receptor" and "response to other organism" GO terms); it is likely that the latter class of proteins may be dispensable if the downstream signaling paths can be elucidated, whereas the former likely cannot.

A somewhat unexpected discovery of this study is that the artificially reduced minimal syn3.0 genome retains a scale-free PPI network, similar to other naturally occurring PPI networks such as that of *E. coli*. Since the population of proteins with a high number of PPI partners is significantly enhanced in the scale-free networks in comparison with a random network (Figure 39) that follows a Gaussian distribution, the robustness of scale-free PPI network of the syn3.0 genome likely arises due to the biological importance of network hub proteins, which are unlikely to be removed over the course of genomic pruning and critically contribute to the successful generation of the genome. The scale-free behavior of biological networks should be an important consideration in future synthetic biology experiments.

Chapter 7 Conclusion

7.1 Overall Conclusion

This thesis presents an integrated protein structure and function modeling pipeline. The first pipeline component is the DeepMSA algorithm for generating a high quality MSA with deep and diverse alignment (Chapter 2). The MSA is used by deep learning to predict distance and orientations by deep learning to guide the D-QUARK protein folding simulation (Chapter 3). The predicted structure models are used for function template detection in the COFACTOR protocol, which combines structure, sequence and PPI for consensus protein function annotation (Chapter 4). The structure and function prediction pipeline is applied to several large-scale genome-wide annotation efforts, including the modeling of human uPE1 proteins (Chapter 5) and JCVI-syn3.0 minimal genome (Chapter 6).

7.2 Future Directions

Deep learning-based protein folding is an important direction in bioinformatics. Despite rapid progress in this field, there are still at least five open challenges in this field that are under-studied: real-value distance/orientation prediction, single sequence-based prediction, deep learning-based threading, end-to-end protein folding, and structure-based function annotation.

7.2.1 Real-value distance and orientation prediction

Most distance-based protein folding program, including RaptorX-Contact[171], DMPfold[172], AlphaFold[173], and trRosetta[196], and even D-QUARK, incorporates deep learning distance

prediction in the form of binned probability distribution. While predictors for distance bins are easier to develop as they can be extended from existing contact predictor, it is not without its inherent limit. Distance bins that are too wide can limit the resolution of predicted distance, while distance bins that are too narrow will result in a small number of training labels and therefore difficulty in training. Orientation bin prediction has a similar resolution limit.

A potential workaround is to predict real-value distance rather than probability of distance within bins. For example, GANProDist[168] trains a Generative Adversary Network (GAN) using an adversary loss to make the predicted distance map indistinguishable from native distance map. Unfortunately, GANProDist seems to perform poorly in CASP14 (as group ProdGAN_Gonglab) in both contact and distance prediction. It is still unclear whether this is caused by a flaw in the real-value distance label design[169] for GAN training, or the inherent unsuitability of GAN in distance/contact prediction. Meanwhile, PDNET[169] uses a more conventional ResNet architecture and a loss function that minimize the error for the prediction of the reciprocal of distance.

While these approaches can potentially address the real-value distance prediction problem, they also introduce another open question for how to incorporate them into protein folding simulation. To implement the distance restraint as an energy term in protein folding, either the upper/lower bounds or the probability distribution for the distance is required. This is straightforward for distance bin predictions, from which both the standard deviation and the probability distribution over the full range of distance can be easily derived, but not for predicted distance predicted with a single real-value.

7.2.2 Single sequence-based predictor

Currently, most contact and distance predictors critically depend on coevolution features derived from MSAs. Therefore, their accuracy relies on the availability of high quality MSAs with deep alignment depth and diverse sequence homologs. This reliance not only limits their application to targets with little to no sequence homologs such as designed proteins, but also dramatically increase the running time of an otherwise lightweight predictors. For example, on average, the TripletRes CASP server takes a few hours to construct the MSA for a target using DeepMSA, while the coevolution feature exactions and deep learning model evaluation only takes a few minutes.

A direct workaround to address this issue is to develop single sequence-based distance/contact predictors, where all features are derived from the target sequence alone. So far, single sequence-based contact predictors[358-360] are not yet able to achieve a similar accuracy as MSA-based predictors, although the single sequence-based predictors using sequence embedding features have shown promise.

7.2.3 Deep learning-based threading

Several threading programs have indirectly used deep learning by incorporating predicted secondary structures, contacts and distance as part of the scoring function[186,191,361] . However, very few threading programs directly apply deep learning to generate alignment. Actually, similar to contact/distance-map prediction, threading alignment can also be formulated as an image segmentation problem. The target-template alignment can be considered an asymmetric image, where a row and a column represent a query position and a template position, respectively. The pixel in the image represents the alignment score for aligning the

corresponding target residue to template residue. Currently, ThreaderAI [362] and SAdLSA[363] are probably the only deep learning threading programs based on this formulation, and have already reported to show improvements over more conventional threading programs.

7.2.4 End-to-end structure prediction

Most deep learning-based protein folding programs are not end-to-end. In other words, they must first generate contact or distance map prediction, and then the predicted contact/distance map to construct 3D structure by protein folding simulation, rather than directly generate tertiary structure from target sequence (or sequence profile) using a neural network. RGN[364] and NEMO[365] are two representative early attempts for end-to-end training of neural network for direct full length tertiary structure generation. In particular, RGN has received much academic and media attention due to its reported fast speed and high accuracy. Unfortunately, its performance in CASP13 is unimpressive; and follow up studies to reproduce the published performance have been largely unsuccessful[366,367]. One factor that may have limited the performance of these algorithms is their lack of coevolutionary features, which provide the critical long-range pairwise interaction information. Despite these setbacks, end-to-end protein folding represents a new avenue that deserves further research, and have shown to be useful for specific modeling tasks such as fragment generations[173].

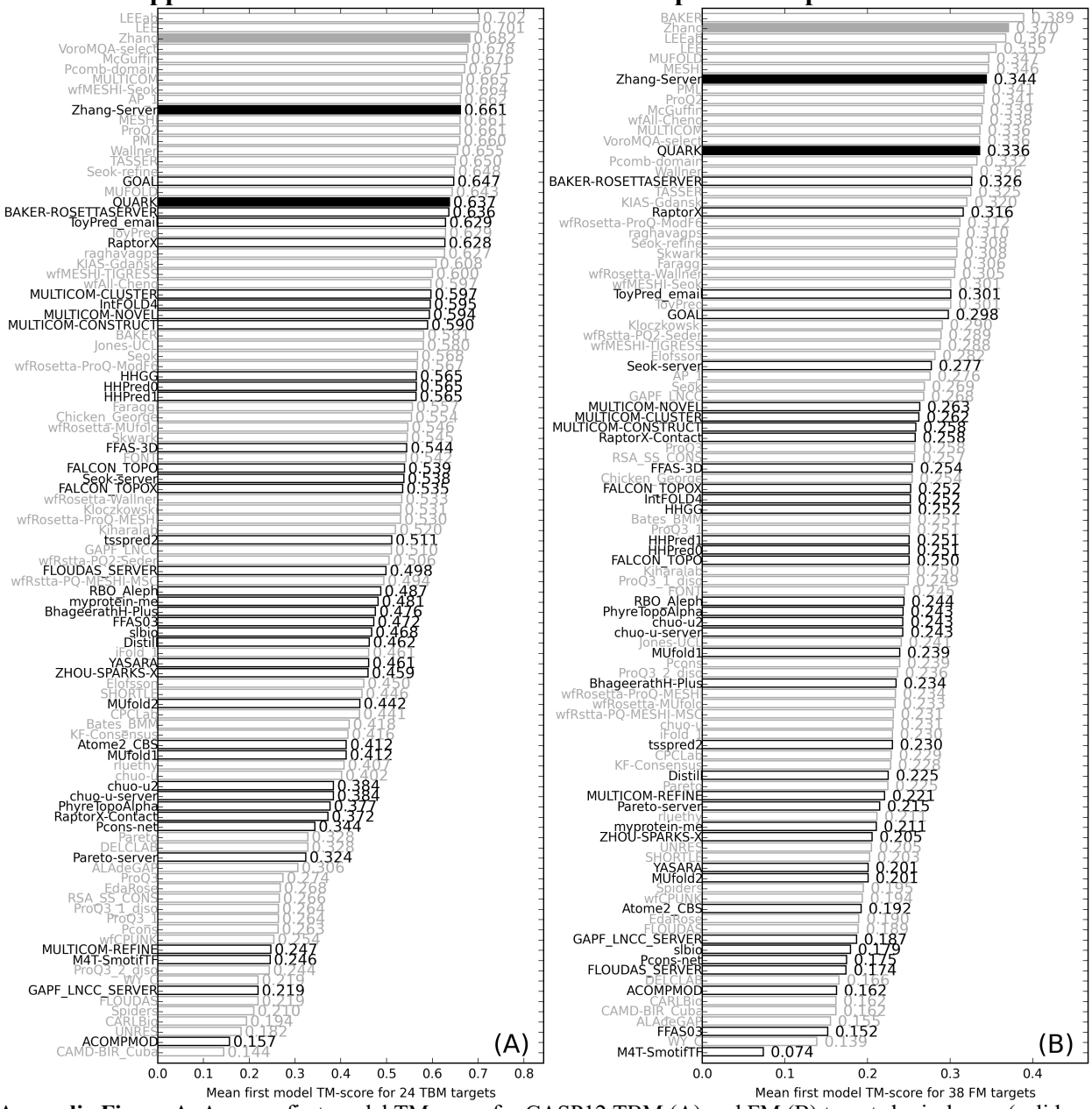
7.2.5 Structure-based *ab initio* function annotation

Ab initio annotation of protein function directly from structure is only recently proposed [202,203], as earlier structure-based function annotation algorithms are based on structure templates identified by global and local structure alignment [198-201]. In *ab initio* structure-based annotation

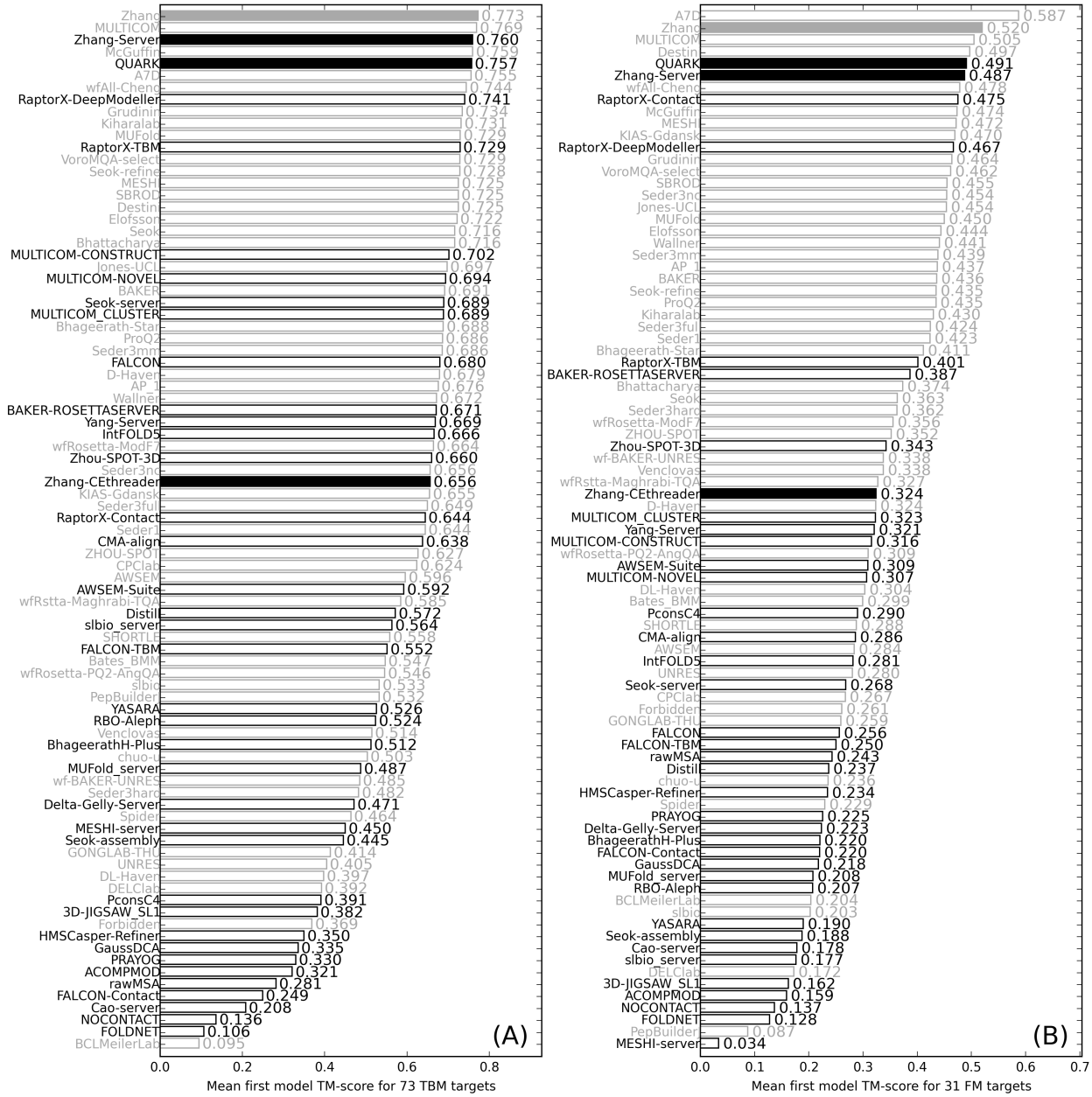
of protein-level functions such as Gene Ontology (GO) terms (as opposed to residue-level functions such as ligand binding sites), the protein structure is converted into a 3D density map[204] or a graph[202], where each graph edge represents an interaction between residues (nodes in the graph). Deep CNNs can then be trained on the converted structure to predict GO labels. Both conversions ignore the sequential connectivity of the protein, and do not use any sequence profile, which is known to be quite useful for function annotation tasks[201,267]. These approaches do not use contact or distance maps, which are more readily available than the tertiary structure model for uncharacterized proteins and inherently includes information of sequential relation between residues. Finally, a constant challenge of structure-based function annotation compared to other non-structure-based function predictor is the incompleteness of structure-function library, where at least two thirds of proteins with known function do not have experimental structure. This limitation could potentially be addressed by extensive data argumentation[368] and by training on predicted structure models.

Appendices

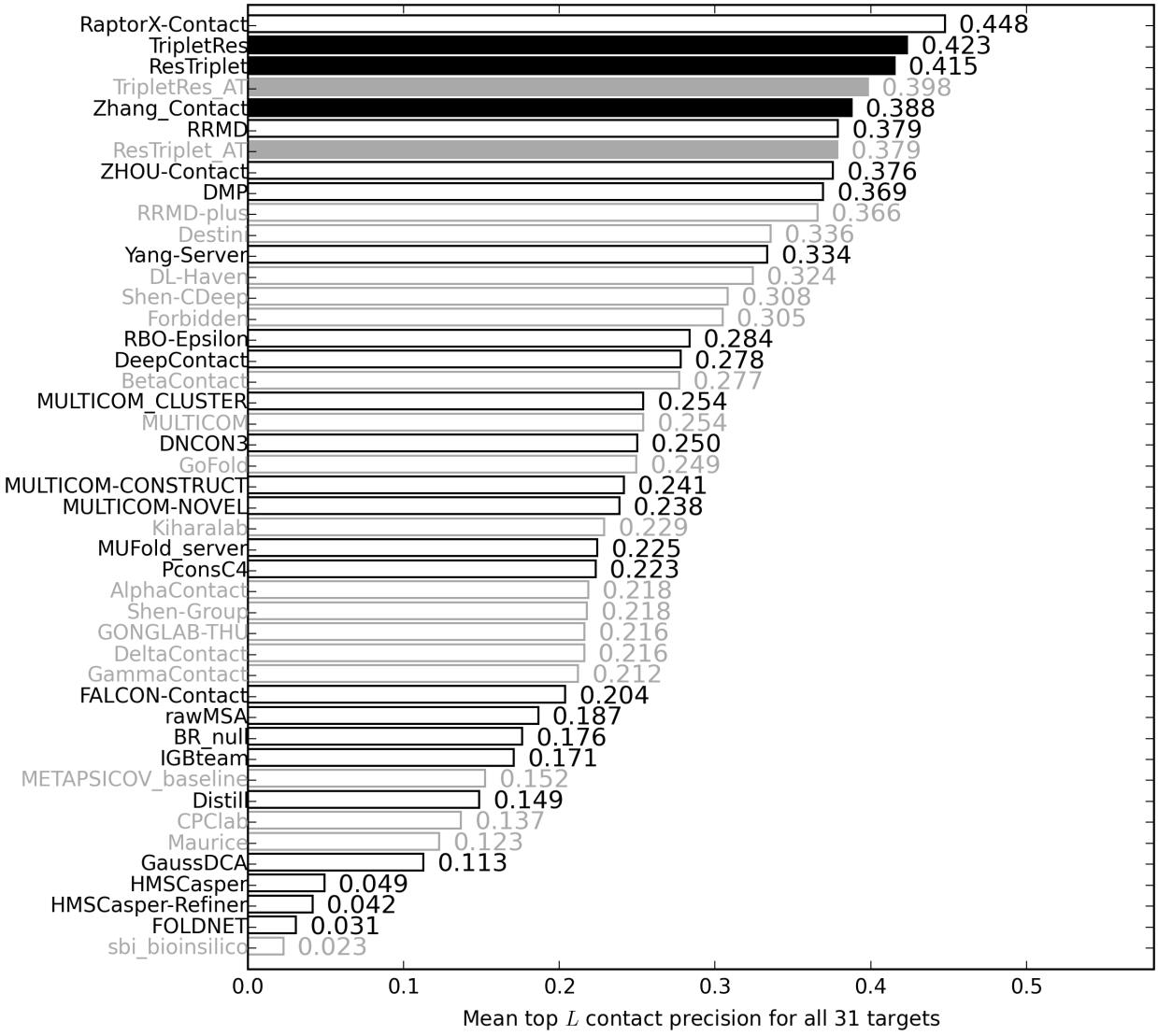
Appendix A Assessment of CASP12-CASP13 prediction performance



Appendix Figure A. Average first model TM-score for CASP12 TBM (A) and FM (B) targets by in-house (solid bars) and third-party (empty bars) CASP12 groups. Black and grey bars are server groups and human groups, respectively. It is not completely fair to compare human group performance against server performance, as human groups have much longer time than server groups and can use server results. Nonetheless, our in-house server groups ("Zhang-Server", i.e. I-TASSER, and "QUARK", i.e. C-QUARK) still outperform many human groups.



Appendix Figure B. Average first model TM-score for CASP13 TBM (A) and FM (B) targets by in-house (solid bars) and third-party (empty bars) CASP13 groups. Black and grey bars are server groups and human groups, respectively. It is not completely fair to compare human group performance against server performance, as human groups have much longer time than server groups and can use server results. Nonetheless, our in-house server groups ("Zhang-Server", i.e. C-I-TASSER, and "QUARK", i.e. C-QUARK) still outperform many human groups.



Appendix Figure C. Average top L long range contact by in-house (solid bars) and third-party (empty bars) CASP13 groups on the subset of 31 FM targets used in official CASP13 contact assessment. Black and grey bars are server groups and human groups, respectively.

Bibliography

1. Anfinsen CB. Principles That Govern Folding of Protein Chains. Science 1973;181(4096):223-230.
2. Bairoch A, Apweiler R, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang HZ, Lopez R, Magrane M, Martin MJ, Natale DA, O'Donovan C, Redaschi N, Yeh LSL. The universal protein resource (UniProt). Nucleic Acids Res 2005;33:D154-D159.
3. Glusker JP. X-ray crystallography of proteins. Methods Biochem Anal 1994;37:1-72.
4. Cavanaugh J., Fairbrother W. J., Palmer A.G., N. S. Protein NMR Spectroscopy: Principles and Practice: New York: Academic Press; 1996.
5. Cheng Y. Single-Particle Cryo-EM at Crystallographic Resolution. Cell 2015;161(3):450-457.
6. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-242.
7. Bairoch A, Bougueleret L, Altairac S, Amendolia V, Auchincloss A, Puy GA, Axelsen K, Baratin D, Blatter MC, Boeckmann B, Bollondi L, Boutet E, Quintaje SB, Breuza L, Bridge A, Saux VBL, deCastro E, Ciampina L, Coral D, Coudert E, Cusin I, David F, Delbard G, Dornevil D, Duek-Roggli P, Duvaud S, Estreicher A, Famiglietti L, Farriol-Mathis N, Ferro S, Feuermann M, Gasteiger E, Gateau A, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Hulo N, Innocenti A, James J, Jain E, Jimenez S, Jungo F, Junker V, Keller G, Lachaize C, Lane-Guermonprez L, Langendijk-Genevaux P, Lara V, Le Mercier P, Lieberherr D, Lima TD, Mangold V, Martin X, Michoud K, Moinat M, Morgat A, Nicolas M, Paesano S, Pedruzzi I, Perret D, Phan I, Pilbout S, Pillet V, Poux S, Pozzato M, Redaschi N, Reynaud S, Rivoire C, Roechert B, Sapsezian C, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Vitorello C, Yip L, Zuletta LF, Apweiler R, Alam-Faruque Y, Barrell D, Bower L, Browne P, Chan WM, Daugherty L, Donate ES, Eberhardt R, Fedotov A, Foulger R, Frigerio G, Garavelli J, Golin R, Horne A, Jacobsen J, Kleen M, Kersey P, Laiho K, Legge D, Magrane M, Martin MJ, Monteiro P, O'Donovan C, Orchard S, O'Rourke J, Patient S, Pruess M, Sitnov A, Whitefield E, Wieser D, Lin Q, Rynbeek M, di Martino G, Donnelly M, van Rensburg P, Wu C, Arighi C, Arminski L, Barker W, Chen YX, Crooks D, Hu ZZ, Hua HK, Huang HZ, Kahsay R, Mazumder R, McGarvey P, Natale D, Nikolskaya AN, Petrova N, Suzek B, Vasudevan S, Vinayaka CR, Yeh LS, Zhang J, Consortium U. The Universal Protein Resource (UniProt). Nucleic Acids Res 2008;36:D190-D195.
8. Levitt M, Warshel A. Computer-Simulation of Protein Folding. Nature 1975;253(5494):694-698.
9. Lewis PN, Momany FA, Scheraga HA. Folding of Polypeptide Chains in Proteins - Proposed Mechanism for Folding. P Natl Acad Sci USA 1971;68(9):2293-&.
10. Mccammon JA, Gelin BR, Karplus M. Dynamics of Folded Proteins. Nature 1977;267(5612):585-590.
11. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164-170.
12. Skolnick J, Kolinski A. Simulations of the Folding of a Globular Protein. Science 1990;250(4984):1121-1125.
13. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234(3):779-815.

14. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268(1):209-225.
15. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 2010;5(4):725-738.
16. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature Methods 2015;12(1):7-8.
17. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Protein structure determination using metagenome sequence data. Science 2017;355(6322):294-298.
18. Wodak SJ, Janin J. Computer analysis of protein-protein interaction. J Mol Biol 1978;124(2):323-342.
19. Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C, Vakser IA. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A 1992;89(6):2195-2199.
20. Kundrotas PJ, Zhu Z, Janin J, Vakser IA. Templates are available to model nearly all complexes of structurally characterized proteins. Proc Natl Acad Sci U S A 2012;109(24):9438-9441.
21. Chen R, Li L, Weng Z. ZDOCK: an initial-stage protein-docking algorithm. Proteins 2003;52(1):80-87.
22. Qin S, Zhou HX. A holistic approach to protein docking. Proteins 2007;69(4):743-749.
23. Comeau SR, Gatchell DW, Vajda S, Camacho CJ. ClusPro: an automated docking and discrimination method for the prediction of protein complexes. Bioinformatics 2004;20(1):45-50.
24. Moal IH, Chaleil RAG, Bates PA. Flexible Protein-Protein Docking with SwarmDock. Methods Mol Biol 2018;1764:413-428.
25. Fischer D, Eisenberg D. Assigning folds to the proteins encoded by the genome of Mycoplasma genitalium. Proc Natl Acad Sci U S A 1997;94(22):11929-11934.
26. Sanchez R, Sali A. Evaluation of comparative protein structure modeling by MODELLER-3. Proteins 1997;Suppl. 1:50-58.
27. Zhang Y, Skolnick J. Automated structure prediction of weakly homologous proteins on a genomic scale. Proc Natl Acad Sci USA 2004;101:7594-7599.
28. Lu L, Arakaki AK, Lu H, Skolnick J. Multimeric threading-based prediction of protein-protein interactions on a genomic scale: application to the Saccharomyces cerevisiae proteome. Genome Res 2003;13(6A):1146-1154.
29. Malmstrom L, Riffle M, Strauss CE, Chivian D, Davis TN, Bonneau R, Baker D. Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS biology 2007;5(4):e76.
30. Mukherjee S, Szilagyi A, Roy A, Zhang Y. Genome-wide protein structure prediction. In: Kolniski A, editor. Multiscale approaches to protein modeling: structure prediction, dynamics, thermodynamics and macromolecular assemblies: Springer-London; 2010. p 810-842.
31. Xu D, Zhang Y. Ab Initio structure prediction for Escherichia coli: towards genome-wide protein structure modeling and fold assignment. Sci Rep 2013;3:1895.
32. Zhang QC, Petrey D, Deng L, Qiang L, Shi Y, Thu CA, Bisikirska B, Lefebvre C, Accili D, Hunter T, Maniatis T, Califano A, Honig B. Structure-based prediction of protein-protein interactions on a genome-wide scale. Nature 2012;490(7421):556-560.
33. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 2004;32(Web Server issue):W526-531.
34. Källberg M, Wang H, Wang S, Peng J, Wang Z, Lu H, Xu J. Template-based protein structure modeling using the RaptorX web server. Nat Protocols 2012;7(8):1511-1522.
35. Kelley LA, Sternberg MJ. Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 2009;4(3):363-371.

36. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 2003;31(13):3381-3385.
37. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, Stuart AC, Mirkovic N, Rossi A, Marti-Renom MA, Fiser A, Webb B, Greenblatt D, Huang CC, Ferrin TE, Sali A. MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 2004;32(Database issue):D217-222.
38. Soding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 2005;33(Web Server issue):W244-248.
39. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846-856.
40. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19(8):1015-1018.
41. Wang Z, Eickholt J, Cheng J. MULTICOM: a multi-level combination approach to protein structure prediction and its assessments in CASP8. Bioinformatics 2010;26(7):882-888.
42. Pierce BG, Wiehe K, Hwang H, Kim BH, Vreven T, Weng Z. ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers. Bioinformatics 2014;30(12):1771-1773.
43. Tovchigrechko A, Vakser IA. GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 2006;34(Web Server issue):W310-314.
44. Macindoe G, Mavridis L, Venkatraman V, Devignes MD, Ritchie DW. HexServer: an FFT-based protein docking server powered by graphics processors. Nucleic Acids Res 2010;38(Web Server issue):W445-449.
45. Lyskov S, Gray JJ. The RosettaDock server for local protein-protein docking. Nucleic Acids Res 2008;36(Web Server issue):W233-238.
46. Zheng W, Zhang C, Bell EW, Zhang Y. I-TASSER gateway: A protein structure and function prediction server powered by XSEDE. Future Gener Comput Syst 2019;99:73-85.
47. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80(7):1715-1735.
48. Shakhnovich BE, Dokholyan NV, DeLisi C, Shakhnovich EI. Functional fingerprints of folds: evidence for correlated structure-function evolution. J Mol Biol 2003;326(1):1-9.
49. Watson JD, Sanderson S, Ezersky A, Savchenko A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM. Towards fully automated structure-based function prediction in structural genomics: a case study. Journal of molecular biology 2007;367(5):1511-1522.
50. Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, Raushel FM. Structure-based activity prediction for an enzyme of unknown function. Nature 2007;448(7155):775-779.
51. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. Protein function annotation by homology-based inference. Genome Biol 2009;10(2):207.
52. Baxter SM, Fetrow JS. Sequence- and structure-based protein function prediction from genomic information. Curr Opin Drug Discov Devel 2001;4(3):291-295.
53. Bonneau R, Tsai J, Ruczinski I, Baker D. Functional inferences from blind ab initio protein structure predictions. Journal of structural biology 2001;134(2-3):186-190.
54. Arakaki AK, Zhang Y, Skolnick J. Large scale assesment of the utility of low resolution protein structures for biochemical function assignment. Bioinformatics 2004;20:1087-1096.
55. Zhang Y, Thiele I, Weekes D, Li Z, Jaroszewski L, Ginalski K, Deacon AM, Wooley J, Lesley SA, Wilson IA, Palsson B, Osterman A, Godzik A. Three-dimensional structural view of the central metabolic network of Thermotoga maritima. Science 2009;325(5947):1544-1549.

56. Kinch LN, Grishin NV. Bioinformatics perspective on rhomboid intramembrane protease evolution and function. Biochim Biophys Acta 2013;1828(12):2937-2943.
57. Huang H, Zhao R, Dickson BM, Skeel RD, Post CB. alphaC helix as a switch in the conformational transition of Src/CDK-like kinase domains. J Phys Chem B 2012;116(15):4465-4475.
58. Cross TA, Dong H, Sharma M, Busath DD, Zhou HX. M2 protein from influenza A: from multiple structures to biophysical and functional insights. Curr Opin Virol 2012;2(2):128-133.
59. Cai XH, Jaroszewski L, Wooley J, Godzik A. Internal organization of large protein families: relationship between the sequence, structure, and function-based clustering. Proteins 2011;79(8):2389-2402.
60. Boyd A, Ciufo LF, Barclay JW, Graham ME, Haynes LP, Doherty MK, Riesen M, Burgoyne RD, Morgan A. A random mutagenesis approach to isolate dominant-negative yeast sec1 mutants reveals a functional role for domain 3a in yeast and mammalian Sec1/Munc18 proteins. Genetics 2008;180(1):165-178.
61. Capriotti E, Fariselli P, Casadio R. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 2005;33(Web Server issue):W306-310.
62. Tokuriki N, Tawfik DS. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol 2009;19(5):596-604.
63. Quan L, Lv Q, Zhang Y. STRUM: structure-based prediction of protein stability changes upon single-point mutation. Bioinformatics 2016;32(19):2936-2946.
64. Porta-Pardo E, Hrabe T, Godzik A. Cancer3D: understanding cancer mutations through protein structures. Nucleic Acids Res 2015;43(Database issue):D968-973.
65. Pires DE, Ascher DB, Blundell TL. mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics 2014;30(3):335-342.
66. Porta-Pardo E, Godzik A. Mutation Drivers of Immunological Responses to Cancer. Cancer Immunol Res 2016;4(9):789-798.
67. Evers A, Klebe G. Successful virtual screening for a submicromolar antagonist of the neurokinin-1 receptor based on a ligand-supported homology model. Journal of medicinal chemistry 2004;47(22):5381-5392.
68. Klebe G. Virtual ligand screening: strategies, perspectives and limitations. Drug discovery today 2006;11(13-14):580-594.
69. Zhou H, Skolnick J. FINDSITE(X): A Structure-Based, Small Molecule Virtual Screening Approach with Application to All Identified Human GPCRs. Mol Pharm 2012;9(6):1775-1784.
70. Roy A, Zhang Y. Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. Structure 2012;20(6):987-997.
71. Tseng YY, Dundas J, Liang J. Predicting protein function and binding profile via matching of local evolutionary and geometric surface patterns. J Mol Biol 2009;387(2):451-464.
72. Vajda S, Guarnieri F. Characterization of protein-ligand interaction sites using experimental and computational methods. Curr Opin Drug Disc 2006;9(3):354-362.
73. Drews J. Drug discovery: a historical perspective. Science 2000;287(5460):1960-1964.
74. Evers A, Klabunde T. Structure-based drug discovery using GPCR homology modeling: Successful virtual screening for antagonists of the Alpha1A adrenergic receptor. Journal of medicinal chemistry 2005;48(4):1088-1097.
75. Hubbard RE, editor. Structure-Based Drug Discovery. First edition ed: Royal Society of Chemistry; 2006.
76. Jacobson M, Sali A. Comparative protein structure modeling and its applications to drug discovery. Annu Rep Med Chem 2004;39:259-276.
77. Kuntz ID. Structure-based strategies for drug design and discovery. Science 1992;257(5073):1078-1082.

78. Whittle PJ, Blundell TL. Protein Structure-Based Drug Design. Annu Rev Bioph Biom 1994;23:349-375.
79. Ekins S, Mestres J, Testa B. In silico pharmacology for drug discovery: applications to targets and beyond. British journal of pharmacology 2007;152(1):21-37.
80. Becker OM, Dhanoa DS, Marantz Y, Chen D, Shacham S, Cheruku S, Heifetz A, Mohanty P, Fichman M, Sharadendu A, Nudelman R, Kauffman M, Noiman S. An integrated in silico 3D model-driven discovery of a novel, potent, and selective amidosulfonamide 5-HT1A agonist (PRX-00023) for the treatment of anxiety and depression. Journal of medicinal chemistry 2006;49(11):3116-3135.
81. Du J, Cross TA, Zhou HX. Recent progress in structure-based anti-influenza drug design. Drug discovery today 2012;17(19-20):1111-1120.
82. Archakov AI, Govorun VM, Dubanov AV, Ivanov YD, Veselovsky AV, Lewi P, Janssen P. Protein-protein interactions as a target for drugs in proteomics. Proteomics 2003;3(4):380-391.
83. Han X, Wang C, Qin C, Xiang W, Fernandez-Salas E, Yang CY, Wang M, Zhao L, Xu T, Chinnaswamy K, Delproposto J, Stuckey J, Wang S. Discovery of ARD-69 as a Highly Potent Proteolysis Targeting Chimera (PROTAC) Degrader of Androgen Receptor (AR) for the Treatment of Prostate Cancer. Journal of medicinal chemistry 2019.
84. Rost B. Twilight zone of protein sequence alignments. Protein Eng 1999;12(2):85-94.
85. Baker D, Sali A. Protein structure prediction and structural genomics. Science 2001;294(5540):93-96.
86. Kryshtafovych A, Monastyrskyy B, Fidelis K, Moult J, Schwede T, Tramontano A. Evaluation of the template-based modeling in CASP12. Proteins 2018;86 Suppl 1:321-334.
87. Dunbrack R. Template-based modeling assessment in CASP11. 2014; Riviera Maya, Mexico.
88. Zhang Y, Skolnick J. The protein structure prediction problem could be solved using the current PDB library. P Natl Acad Sci USA 2005;102(4):1029-1034.
89. Skolnick J, Zhou HY. Why Is There a Glass Ceiling for Threading Based Protein Structure Prediction Methods? J Phys Chem B 2017;121(15):3546-3554.
90. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins-Structure Function and Bioinformatics 2012;80(7):1715-1735.
91. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J. Critical assessment of methods of protein structure prediction (CASP)-Round XIII. Proteins 2019;87(12):1011-1020.
92. Moult J, Fidelis K, Kryshtafovych A, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)- Round IX. Proteins 2011;79:1-5.
93. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP) - round x. Proteins 2014;82:1-6.
94. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. Proteins 2016;84:4-14.
95. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A. Critical assessment of methods of protein structure prediction (CASP)Round XII. Proteins 2018;86:7-15.
96. Browne WJ, North ACT, Phillips DC. A Possible 3-Dimensional Structure of Bovine Alpha-Lactalbumin Based on That of Hens Egg-White Lysozyme. Journal of Molecular Biology 1969;42(1):65-&.
97. Needleman SB, Wunsch CD. A General Method Applicable to Search for Similarities in Amino Acid Sequence of 2 Proteins. Journal of Molecular Biology 1970;48(3):443-+.
98. Smith TF, Waterman MS. Identification of Common Molecular Subsequences. Journal of Molecular Biology 1981;147(1):195-197.

99. Wu ST, Zhang Y. MUSTER: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 2008;72(2):547-556.
100. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951-960.
101. Yang YD, Faraggi E, Zhao HY, Zhou YQ. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 2011;27(15):2076-2082.
102. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 1997;25(17):3389-3402.
103. Remmert M, Biegert A, Hauser A, Söding J. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods 2012;9(2):173-175.
104. Jo T, Hou J, Eickholt J, Cheng JL. Improving Protein Fold Recognition by Deep Learning Networks. Scientific Reports 2015;5.
105. Xia JQ, Peng ZL, Qi DW, Mu HB, Yang JY. An ensemble approach to protein fold classification by integration of template-based assignment and support vector machine classifier. Bioinformatics 2017;33(6):863-870.
106. Zhu JW, Zhang HC, Li SC, Wang C, Kong LP, Sun SW, Zheng WM, Bu DB. Improving protein fold recognition by extracting fold-specific features from predicted residue-residue contacts. Bioinformatics 2017;33(23):3749-3757.
107. Park H, DiMaio F, Baker D. The Origin of Consistent Protein Structure Refinement from Structural Averaging. Structure 2015;23(6):1123-1128.
108. Zhang Y. Progress and challenges in protein structure prediction. Curr Opin Struc Biol 2008;18(3):342-348.
109. Song YF, DiMaio F, Wang RYR, Kim D, Miles C, Brunette TJ, Thompson J, Baker D. High-Resolution Comparative Modeling with RosettaCM. Structure 2013;21(10):1735-1742.
110. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10(6):845-858.
111. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic acids research 2015;43(W1):W174-W181.
112. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nat Methods 2015;12(1):7.
113. Roy A, Kucukural A, Zhang Y. I-TASSER: a unified platform for automated protein structure and function prediction. Nature protocols 2010;5(4):725.
114. Zhang Y. I-TASSER server for protein 3D structure prediction. Bmc Bioinformatics 2008;9(1):40.
115. Wu ST, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. Bmc Biol 2007;5.
116. Levitt M, Lifson S. Refinement of Protein Conformations Using a Macromolecular Energy Minimization Procedure. Journal of Molecular Biology 1969;46(2):269-&.
117. Weiner SJ, Kollman PA, Case DA, Singh UC, Ghio C, Alagona G, Profeta S, Weiner P. A New Force-Field for Molecular Mechanical Simulation of Nucleic-Acids and Proteins. J Am Chem Soc 1984;106(3):765-784.
118. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules (vol 117, pg 5179, 1995). J Am Chem Soc 1996;118(9):2309-2309.
119. Duan Y, Kollman PA. Pathways to a protein folding intermediate observed in a 1-microsecond simulation in aqueous solution. Science 1998;282(5389):740-744.
120. Brooks BR, Bruccoleri RE, Olafson BD, States DJ, Swaminathan S, Karplus M. Charmm - a Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J Comput Chem 1983;4(2):187-217.

121. Neria E, Fischer S, Karplus M. Simulation of activation free energies in molecular systems. J Chem Phys 1996;105(5):1902-1921.
122. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. Journal of Physical Chemistry B 1998;102(18):3586-3616.
123. Jorgensen WL, Tiradorives J. The Opls Potential Functions for Proteins - Energy Minimizations for Crystals of Cyclic-Peptides and Crambin. J Am Chem Soc 1988;110(6):1657-1666.
124. Jorgensen WL, Maxwell DS, TiradoRives J. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. J Am Chem Soc 1996;118(45):11225-11236.
125. Scott WRP, Hunenberger PH, Tironi IG, Mark AE, Billeter SR, Fennen J, Torda AE, Huber T, Kruger P, van Gunsteren WF. The GROMOS biomolecular simulation program package. J Phys Chem A 1999;103(19):3596-3607.
126. Raval A, Piana S, Eastwood MP, Dror RO, Shaw DE. Refinement of protein structure homology models via long, all-atom molecular dynamics simulations. Proteins 2012;80(8):2071-2079.
127. Robustelli P, Piana S, Shaw DE. Developing a molecular dynamics force field for both folded and disordered protein states. P Natl Acad Sci USA 2018;115(21):E4758-E4766.
128. Lange OF, van der Spoel D, de Groot BL. Scrutinizing Molecular Mechanics Force Fields on the Submicrosecond Timescale with NMR Data. Biophys J 2010;99(2):647-655.
129. Lindorff-Larsen K, Maragakis P, Piana S, Eastwood MP, Dror RO, Shaw DE. Systematic Validation of Protein Force Fields against Experimental Data. Plos One 2012;7(2).
130. Beauchamp KA, Lin YS, Das R, Pande VS. Are Protein Force Fields Getting Better? A Systematic Benchmark on 524 Diverse NMR Measurements. J Chem Theory Comput 2012;8(4):1409-1414.
131. Lindorff-Larsen K, Piana S, Dror RO, Shaw DE. How Fast-Folding Proteins Fold. Science 2011;334(6055):517-520.
132. Mittal J, Best RB. Tackling Force-Field Bias in Protein Folding Simulations: Folding of Villin HP35 and Pin WW Domains in Explicit Water. Biophys J 2010;99(3):L26-L28.
133. Huang J, Rauscher S, Nawrocki G, Ran T, Feig M, de Groot BL, Grubmuller H, MacKerell AD. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. Nat Methods 2017;14(1):71-73.
134. Bowie JU, Eisenberg D. An Evolutionary Approach to Folding Small Alpha-Helical Proteins That Uses Sequence Information and an Empirical Guiding Fitness Function. P Natl Acad Sci USA 1994;91(10):4436-4440.
135. Simons KT, Kooperberg C, Huang E, Baker D. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. J Mol Biol 1997;268(1):209-225.
136. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using rosetta. Method Enzymol 2004;383:66-+.
137. Jones DT. Predicting novel protein folds by using FRAGFOLD. Proteins 2001:127-132.
138. Gobel U, Sander C, Schneider R, Valencia A. Correlated mutations and residue contacts in proteins. Proteins 1994;18(4):309-317.
139. Thomas DJ, Casari G, Sander C. The prediction of protein contacts from multiple sequence alignments. Protein Eng 1996;9(11):941-948.
140. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. P Natl Acad Sci USA 2009;106(1):67-72.

141. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, Zecchina R, Onuchic JN, Hwa T, Weigt M. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proceedings of the National Academy of Sciences of the United States of America 2011;108(49):E1293-E1301.
142. Chiu DKY, Kolodziejczak T. Inferring Consensus Structure from Nucleic-Acid Sequences. Comput Appl Biosci 1991;7(3):347-352.
143. Baldassi C, Zamparo M, Feinauer C, Procaccini A, Zecchina R, Weigt M, Pagnani A. Fast and Accurate Multivariate Gaussian Modeling of Protein Families: Predicting Residue Contacts and Protein-Interaction Partners. Plos One 2014;9(3).
144. Ekeberg M, Lovkvist C, Lan YH, Weigt M, Aurell E. Improved contact prediction in proteins: Using pseudolikelihoods to infer Potts models. Phys Rev E 2013;87(1).
145. Kamisetty H, Ovchinnikov S, Baker D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era (vol 110, pg 15674, 2013). P Natl Acad Sci USA 2013;110(46):18734-18734.
146. Seemayer S, Gruber M, Soding J. CCMpred-fast and precise prediction of protein residue-residue contacts from correlated mutations. Bioinformatics 2014;30(21):3128-3130.
147. Jones DT, Buchan DWA, Cozzetto D, Pontil M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. Bioinformatics 2012;28(2):184-190.
148. Li Y, Hu J, Zhang CX, Yu DJ, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics 2019;35(22):4647-4655.
149. Fariselli P, Casadio R. A neural network based predictor of residue contacts in proteins. Protein Eng 1999;12(1):15-21.
150. Fariselli P, Olmea O, Valencia A, Casadio R. Prediction of contact maps with neural networks and correlated mutations. Protein Eng 2001;14(11):835-843.
151. Xue B, Faraggi E, Zhou Y. Predicting residue-residue contact maps by a two-layer, integrated neural-network method. Proteins 2009;76(1):176-183.
152. Walsh I, Bau D, Martin AJ, Mooney C, Vullo A, Pollastri G. Ab initio and template-based prediction of multi-class distance maps by two-dimensional recursive neural networks. BMC Struct Biol 2009;9:5.
153. Ma J, Wang S, Wang Z, Xu J. Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. Bioinformatics 2015;31(21):3506-3513.
154. Tegge AN, Wang Z, Eickholt J, Cheng JL. NNcon: improved protein contact map prediction using 2D-recursive neural networks. Nucleic Acids Res 2009;37:W515-W518.
155. Cheng JL, Baldi P. Improved residue contact prediction using support vector machines and a large feature set. Bmc Bioinformatics 2007;8.
156. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics 2008;24(7):924-931.
157. Skwark MJ, Abdel-Rehim A, Elofsson A. PconsC: combination of direct information methods and alignments improves contact prediction. Bioinformatics 2013;29(14):1815-1816.
158. Jones DT, Singh T, Kosciolek T, Tetchner S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. Bioinformatics 2015;31(7):999-1006.
159. He BJ, Mortuza SM, Wang YT, Shen HB, Zhang Y. NeBcon: protein contact map prediction using neural network training coupled with naiive Bayes classifiers. Bioinformatics 2017;33(15):2296-2306.
160. Di Lena P, Nagata K, Baldi P. Deep architectures for protein contact map prediction. Bioinformatics 2012;28(19):2449-2457.
161. Eickholt J, Cheng JL. Predicting protein residue-residue contacts using deep networks and boosting. Bioinformatics 2012;28(23):3066-3072.

162. Wang S, Sun SQ, Li Z, Zhang RY, Xu JB. Accurate De Novo Prediction of Protein Contact Map by Ultra-Deep Learning Model. Plos Comput Biol 2017;13(1).
163. He KM, Zhang XY, Ren SQ, Sun J. Deep Residual Learning for Image Recognition. 2016 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr) 2016:770-778.
164. Kandathil SM, Greener JG, Jones DT. Prediction of interresidue contacts with DeepMetaPSICOV in CASP13. Proteins 2019;87(12):1092-1099.
165. Liu Y, Palmedo P, Ye Q, Berger B, Peng J. Enhancing Evolutionary Couplings with Deep Convolutional Neural Networks. Cell Syst 2018;6(1):65-74.
166. Hanson J, Peliwal K, Litfin T, Yang YD, Zhou YQ. Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. Bioinformatics 2018;34(23):4039-4045.
167. Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins 2019;87(12):1082-1091.
168. Ding WZ, Gong HP. Predicting the Real-Valued Inter-Residue Distances for Proteins. Adv Sci 2020.
169. Adhikari B. A fully open-source framework for deep learning protein real-valued distances. Scientific Reports 2020;10(1).
170. Xu D, Zhang Y. Toward optimal fragment generations for ab initio protein structure assembly. Proteins 2013;81(2):229-239.
171. Xu J. Distance-based protein folding powered by deep learning. Proc Natl Acad Sci U S A 2019;116(34):16856-16865.
172. Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nat Commun 2019;10.
173. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin CL, Z?dek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). Proteins-Structure Function and Bioinformatics 2019;87(12):1141-1148.
174. Zhou HY, Skolnick J. GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction. Biophysical Journal 2011;101(8):2043-2052.
175. Zhang J, Zhang Y. A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction. Plos One 2010;5(10).
176. Yang YD, Zhou YQ. Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 2008;72(2):793-803.
177. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences 2020:201914677.
178. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, Kamisetty H, Kyrpides NC, Baker D. Protein structure determination using metagenome sequence data. Science 2017;355(6322):294-297.
179. Wu Q, Peng Z, Anishchenko I, Cong Q, Baker D, Yang J. Protein contact prediction using metagenome sequence data and residual neural networks. Bioinformatics 2019;36(1):41-48.
180. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2019.

181. Zheng W, Zhang C, Wuyun Q, Pearce R, Li Y, Zhang Y. LOMETS2: improved meta-threading server for fold-recognition and structure-based function annotation for distant-homology proteins. Nucleic Acids Research 2019;47(W1):W429–W436.
182. Wang Y, Shi Q, Yang PS, Zhang CX, Mortuza SM, Xue ZD, Ning K, Zhang Y. Fueling ab initio folding with marine metagenomics enables structure and function predictions of new protein families. Genome Biol 2019;20(1).
183. Zheng W, Li Y, Zhang CX, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins-Structure Function and Bioinformatics 2019;87(12):1149-1164.
184. Xu JB, Wang S. Analysis of distance-based protein structure prediction by deep learning in CASP13. Proteins 2019;87(12):1069-1081.
185. Steinegger M, Soding J. Clustering huge protein sequence sets in linear time. Nat Commun 2018;9.
186. Zheng W, Wuyun QQG, Li Y, Mortuza SM, Zhang CX, Pearce R, Ruan JS, Zhang Y. Detecting distant-homology protein structures by aligning deep neural-network based contact maps. Plos Comput Biol 2019;15(10).
187. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. Journal of Computational Chemistry 2004;25(6):865-871.
188. Zhang J, Liang Y, Zhang Y. Atomic-level protein structure refinement using fragment-guided molecular dynamics conformation sampling. Structure 2011;19(12):1784-1795.
189. Kryshtafovych A, Monastyrskyy B, Fidelis K, Moult J, Schwede T, Tramontano A. Evaluation of the template-based modeling in CASP12. Proteins 2018;86:321-334.
190. Zhang CX, Mortuza SM, He BJ, Wang YT, Zhang Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins 2018;86:136-151.
191. Zhu JW, Wang S, Bu DB, Xu JB. Protein threading using residue co-variation and deep learning. Bioinformatics 2018;34(13):263-273.
192. Brunger AT. Version 1.2 of the Crystallography and NMR system. Nat Protoc 2007;2(11):2728-2733.
193. Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D 1998;54:905-921.
194. Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. Improved protein structure prediction using potentials from deep learning. Nature 2020.
195. Mao W, Ding W, Xing Y, Gong H. AmoebaContact and GDFold as a pipeline for rapid de novo protein structure prediction. Nature Machine Intelligence 2020;2(1):25-33.
196. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. Proceedings of the National Academy of Sciences of the United States of America 2020;117(3):1496-1503.
197. Nagano N, Orengo CA, Thornton JM. One fold with many functions: The evolutionary relationships between TIM barrel families based on their sequences, structures and functions. Journal of Molecular Biology 2002;321(5):741-765.
198. Yang J, Roy A, Zhang Y. Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 2013;29(20):2588-2595.
199. Wu Q, Peng ZL, Zhang Y, Yang JY. COACH-D: improved protein-ligand binding sites prediction with refined ligand-binding poses through molecular docking. Nucleic Acids Research 2018;46(W1):W438-W442.
200. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic acids research 2005;33(suppl_2):W89-W93.

201. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic acids research 2017;45(W1):W291-W299.
202. Gligorijevic V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, Xavier RJ, Knight R, Cho K, Bonneau R. Structure-Based Protein Function Prediction using Graph Convolutional Networks. bioRxiv 2020:786236.
203. Torng W, Altman RB. High precision protein functional site detection using 3D convolutional neural networks. Bioinformatics 2019;35(9):1503-1512.
204. Amidi A, Amidi S, Vlachakis D, Megalooikonomou V, Paragios N, Zacharaki EI. EnzyNet: enzyme classification using 3D convolutional neural networks on spatial representation. Peerj 2018;6.
205. Zhang C, Zheng W, Huang X, Bell EW, Zhou X, Zhang Y. Protein Structure and Sequence Reanalysis of 2019-nCoV Genome Refutes Snakes as Its Intermediate Host and the Unique Similarity between Its Spike Protein Insertions and HIV-1. Journal of Proteome Research 2020.
206. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. Journal of Molecular Biology 1999;292(2):195-202.
207. Wu S, Zhang Y. ANGLOR: A Composite Machine-Learning Algorithm for Protein Backbone Torsion Angle Prediction. Plos One 2008;3(10).
208. He B, Mortuza S, Wang Y, Shen H-B, Zhang Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. Bioinformatics 2017:btx164.
209. Adhikari B, Hou J, Cheng JL. DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. Bioinformatics 2018;34(9):1466-1472.
210. Soding J. Protein homology detection by HMM-HMM comparison. Bioinformatics 2005;21(7):951-960.
211. Zhang C, Mortuza SM, He B, Wang Y, Zhang Y. Template-based and free modeling of I-TASSER and QUARK pipelines using predicted contact maps in CASP12. Proteins 2018;86 Suppl 1:136-151.
212. Ovchinnikov S, Park H, Kim DE, DiMaio F, Baker D. Protein structure prediction using Rosetta in CASP12. Proteins 2018;86 Suppl 1:113-121.
213. Cozzetto D, Minneci F, Currant H, Jones DT. FFPred 3: feature-based function prediction for all Gene Ontology domains. Sci Rep 2016;6:31865.
214. Gil N, Fiser A. The choice of sequence homologs included in multiple sequence alignments has a dramatic impact on evolutionary conservation analysis. Bioinformatics 2018;35(1):12-19.
215. Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. Bioinformatics 2013;29(20):2588-2595.
216. Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. Bmc Bioinformatics 2019;20(1):473.
217. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755-763.
218. Schaarschmidt J, Monastyrskyy B, Kryshtafovych A, Bonvin A. Assessment of contact predictions in CASP12: Co-evolution and deep learning coming of age. Proteins 2018;86 Suppl 1:51-66.
219. Wu S, Szilagyi A, Zhang Y. Improving protein structure prediction using multiple sequence-based contact predictions. Structure 2011;19(8):1182-1191.
220. Buchan DWA, Jones DT. Improved protein contact predictions with the MetaPSICOV2 server in CASP12. Proteins 2018;86:78-83.
221. Wang Y, Shi Q, Yang P, Zhang C, Mortuza S, Xue Z, Ning K, Zhang Y. Fueling ab initio folding with marine microbiome enables structure and function predictions of new protein families. Genome biology 2019;In press.

222. Mirdita M, von den Driesch L, Galiez C, Martin MJ, Soding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. Nucleic Acids Res 2017;45(D1):D170-D176.
223. Suzek BE, Wang YQ, Huang HZ, McGarvey PB, Wu CH, Consortium U. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics 2015;31(6):926-932.
224. Hauser M, Mayer CE, Soding J. kClust: fast and sensitive clustering of large protein sequence databases. Bmc Bioinformatics 2013;14.
225. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li WZ, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 2011;7.
226. Hubbard TJP, Ailey B, Brenner SE, Murzin AG, Chothia C. SCOP, Structural Classification of Proteins Database: Applications to Evaluation of the Effectiveness of Sequence Alignment Methods and Statistics of Protein Structural Data. Acta Crystallographica 2010;54(6-1):1147-1154.
227. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302-2309.
228. Wu S, Zhang Y. LOMETS: a local meta-threading-server for protein structure prediction. Nucleic acids research 2007;35(10):3375-3382.
229. Seemayer S, Gruber M, Söding J. CCMpred—fast and precise prediction of protein residue–residue contacts from correlated mutations. Bioinformatics 2014;30(21):3128-3130.
230. Jones DT, Kandathil SM. High precision in protein contact prediction using fully convolutional neural networks and minimal sequence features. Bioinformatics 2018;34(19):3308-3315.
231. Michel M, Hurtado DM, Elofsson A. PconsC4: fast, free, easy, and accurate contact predictions. bioRxiv 2018:383133.
232. Li Y, Zhang C, Bell EW, Yu DJ, Zhang Y. Ensembling multiple raw coevolutionary features with deep residual neural networks for contact-map prediction in CASP13. Proteins 2019;87(12):1082–1091.
233. Wu S, Zhang Y. MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins 2008;72(2):547-556.
234. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57(4):702-710.
235. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Scientific Reports 2013;3:2619.
236. Zemla A, Venclovas C, Fidelis K, Rost B. A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. Proteins 1999;34(2):220-223.
237. Huang XQ, Zheng W, Pearce R, Zhang Y. SSIPe: accurately estimating protein-protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. Bioinformatics 2020;36(8):2429-2437.
238. Zheng W, Li Y, Zhang C, Pearce R, Mortuza SM, Zhang Y. Deep-learning contact-map guided protein structure prediction in CASP13. Proteins 2019;87(12):1149-1164.
239. Park J, Saitou K. ROTAS: a rotamer-dependent, atomic statistical potential for assessment and prediction of protein structures. Bmc Bioinformatics 2014;15.
240. Zhang C, Zheng W, Mortuza SM, Li Y, Zhang Y. DeepMSA: constructing deep multiple sequence alignment to improve contact prediction and fold-recognition for distant-homology proteins. Bioinformatics 2020;36(7):2105-2112.
241. He B, Mortuza SM, Wang Y, Shen H-B, Zhang Y. NeBcon: protein contact map prediction using neural network training coupled with naïve Bayes classifiers. Bioinformatics 2017;33(15):2296-2306.

242. Zhou HY, Zhou YQ. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 2002;11(11):2714-2726.
243. Nelder JA, Mead R. A Simplex-Method for Function Minimization. Comput J 1965;7(4):308-313.
244. Zhang Y, Skolnick J. SPICKER: A clustering approach to identify near-native protein folds. J Comput Chem 2004;25(6):865-871.
245. Xu D, Zhang Y. Improving the Physical Realism and Structural Accuracy of Protein Models by a Two-Step Atomic-Level Energy Minimization. Biophysical Journal 2011;101(10):2525-2534.
246. Xu J, Zhang Y. How significant is a protein structure similarity with TM-score= 0.5? Bioinformatics 2010;26(7):889-895.
247. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. Proteins 2004;57(4):702-710.
248. Bateman A, Martin MJ, O'Donovan C, Magrane M, Apweiler R, Alpi E, Antunes R, Ar-Ganiska J, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Chavali G, Cibrian-Uhalte E, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Gane P, Cas-Tro LG, Garmiri P, Hatton-Ellis E, Hieta R, Huntley R, Legge D, Liu WD, Luo J, MacDougall A, Mutowo P, Nightin-Gale A, Orchard S, Pichler K, Poggioli D, Pundir S, Pureza L, Qi GY, Rosanoff S, Saidi R, Sawford T, Shypitsyna A, Turner E, Volynkin V, Wardell T, Watkins X, Watkins, Cowley A, Figueira L, Li WZ, McWilliam H, Lopez R, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Aimo L, Argoud-Puy G, Auchincloss A, Axelsen K, Bansal P, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, Casal-Casas C, De Castro E, Coudert E, Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Jungo F, Keller G, Lara V, Lemercier P, Lieberherr D, Lombardot T, Martin X, Masson P, Morgat A, Neto T, Nouspikel N, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Chen CM, Chen YX, Garavelli JS, Huang HZ, Laiho KT, McGarvey P, Natale DA, Suzek BE, Vinayaka CR, Wang QH, Wang YQ, Yeh LS, Yerramalla MS, Zhang J, Consortium U. UniProt: a hub for protein information. Nucleic Acids Res 2015;43(D1):D204-D212.
249. Boutet E, Lieberherr D, Tognolli M, Schneider M, Bansal P, Bridge AJ, Poux S, Bougueleret L, Xenarios I. UniProtKB/Swiss-Prot, the Manually Annotated Section of the UniProt KnowledgeBase: How to Use the Entry View. Methods Mol Biol 2016;1374:23-54.
250. Jiang YX, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo DCE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SME, Martelli PL, Profiti G, Casadio R, Cao RZ, Zhong Z, Cheng JL, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Toronen P, Koskinen P, Holm L, Chen CT, Hsu WL, Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Dukka BKC, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent LC, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasidharan R, Yang HX, Paccanaro A, Gillis J, Sedeno-Cortes AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcet-Houben M, Supek F, Gong QT, Ning W, Zhou YP, Tian WD, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo M, Piovesan D, Tosatto SCE, del Pozo A, Fernandez JM, Maietta P, Valencia A, Tress ML, Benso A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, van Dijk ADJ, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida-e-Silva DC, Vencio

RZN, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJE, Wass MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC, Brenner SE, Linial M, Orengo CA, Rost B, Greene CS, Mooney SD, Friedberg I, Radivojac P. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biology 2016;17.
251. Roy A, Yang JY, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res 2012;40(W1):W471-W477.
252. Webb EC. Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes: Academic Press; 1992.
253. Zhang Y. Protein structure prediction: when is it useful? Curr Opin Struct Biol 2009;19(2):145-155.
254. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009-an integrated Gene Ontology Annotation resource. Nucleic Acids Res 2009;37:D396-D403.
255. Berman H, Henrick K, Nakamura H. Announcing the worldwide Protein Data Bank. Nat Struct Biol 2003;10(12):980-980.
256. Dyson HJ, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Bio 2005;6(3):197-208.
257. Yang JY, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic Acids Res 2013;41(D1):D1096-D1103.
258. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33(7):2302-2309.
259. Henikoff S, Henikoff JG. Amino-Acid Substitution Matrices from Protein Blocks. P Natl Acad Sci USA 1992;89(22):10915-10919.
260. Yu CG, Zavaljevski N, Desai V, Johnson S, Stevens FJ, Reifman J. The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. Bmc Bioinformatics 2008;9.
261. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. Journal of molecular biology 1990;215(3):403-410.
262. Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP, Kuhn M, Bork P, Jensen LJ, von Mering C. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res 2015;43(D1):D447-D452.
263. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. Nucleic acids research 2013;41(D1):D1096-1103.
264. Furnham N, Holliday GL, de Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM. The Catalytic Site Atlas 2.0: cataloging catalytic sites and residues identified in enzymes. Nucleic Acids Res 2014;42(D1):D485-D489.
265. Rogers DJ, Tanimoto TT. A computer program for classifying plants. Science (New York, NY) 1960;132(3434):1115-1118.
266. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, Graim K, Funk C, Verspoor K, Ben-Hur A, Pandey G, Yunes JM, Talwalkar AS, Repo S, Souza ML, Piovesan D, Casadio R, Wang Z, Cheng J, Fang H, Gough J, Koskinen P, Toronen P, Nokso-Koivisto J, Holm L, Cozzetto D, Buchan DW, Bryson K, Jones DT, Limaye B, Inamdar H, Datta A, Manjari SK, Joshi R, Chitale M, Kihara D, Lisewski AM, Erdin S, Venner E, Lichtarge O, Rentzsch R, Yang H, Romero AE, Bhat P, Paccanaro A, Hamp T, Kassner R, Seemayer S, Vicedo E, Schaefer C, Achten D, Auer F, Boehm A, Braun T, Hecht M, Heron M, Honigschmid P, Hopf TA, Kaufmann S, Kiening M, Krompass D, Landerer C, Mahlich Y, Roos M, Bjorne J, Salakoski T, Wong A, Shatkay H, Gatzmann F, Sommer I, Wass MN, Sternberg MJ, Skunca N, Supek F, Bosnjak M, Panov P, Dzeroski S, Smuc T, Kourmpetis YA, van Dijk AD, ter Braak CJ, Zhou Y, Gong Q, Dong X, Tian W, Falda M, Fontana P, Lavezzo E, Di Camillo B, Toppo S, Lan L, Djuric

N, Guo Y, Vucetic S, Bairoch A, Linial M, Babbitt PC, Brenner SE, Orengo C, Rost B, Mooney SD, Friedberg I. A large-scale evaluation of computational protein function prediction. Nat Methods 2013;10(3):221-227.
267. Gong QT, Ning W, Tian WD. GoFDR: A sequence alignment based method for predicting protein functions. Methods 2016;93:3-14.
268. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo da CE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SM, Martelli PL, Profiti G, Casadio R, Cao R, Zhong Z, Cheng J, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Toronen P, Koskinen P, Holm L, Chen CT, Hsu WL, Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Bkc D, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent LC, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasidharan R, Yang H, Paccanaro A, Gillis J, Sedeno-Cortes AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcet-Houben M, Supek F, Gong Q, Ning W, Zhou Y, Tian W, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo M, Piovesan D, Tosatto SC, Del Pozo A, Fernandez JM, Maietta P, Valencia A, Tress ML, Benso A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, van Dijk AD, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida ESDC, Vencio RZ, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJ, Wass MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC, Brenner SE, Linial M, Orengo CA, Rost B, Greene CS, Mooney SD, Friedberg I, Radivojac P. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 2016;17(1):184.
269. Capra JA, Laskowski RA, Thornton JM, Singh M, Funkhouser TA. Predicting Protein Ligand Binding Sites by Combining Evolutionary Sequence Conservation and 3D Structure. Plos Comput Biol 2009;5(12).
270. Brylinski M, Skolnick J. A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. P Natl Acad Sci USA 2008;105(1):129-134.
271. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 2008;9:40.
272. Xu D, Zhang Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. Proteins 2012;80(7):1715-1735.
273. Hildebrand A, Remmert M, Biegert A, Soding J. Fast and accurate automatic structure prediction with HHpred. Proteins 2009;77 Suppl 9:128-132.
274. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10(6):845-858.
275. Hanson RM, Prilusky J, Renjian Z, Nakane T, Sussman JL. JSmol and the Next-Generation Web-Based Representation of 3D Molecular Structure as Applied to Proteopedia. Isr J Chem 2013;53(3-4):207-216.
276. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martinez C, Fulcher C, Huerta AM, Kothari A, Krummenacker M, Latendresse M, Muniz-Rascado L, Ong Q, Paley S, Schroder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I, Karp PD. EcoCyc: fusing model organism databases with systems biology. Nucleic Acids Res 2013;41(D1):D605-D612.
277. Gaudet P, Michel PA, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, Duek PD, Gateau A, Gleizes A, Hinard V, de Laval VR, Lin JJ, Nikitin F, Schaeffer M, Teixeira D, Lane L, Bairoch A. The neXtProt knowledgebase on human proteins: 2017 update. Nucleic Acids Res 2017;45(D1):D177-D182.

278. Uhlen M, Fagerberg L, Hallstrom BM, Lindskog C, Oksvold P, Mardinoglu A, Sivertsson A, Kampf C, Sjostedt E, Asplund A, Olsson I, Edlund K, Lundberg E, Navani S, Szigyarto CA, Odeberg J, Djureinovic D, Takanen JO, Hober S, Alm T, Edqvist PH, Berling H, Tegel H, Mulder J, Rockberg J, Nilsson P, Schwenk JM, Hamsten M, von Feilitzen K, Forsberg M, Persson L, Johansson F, Zwahlen M, von Heijne G, Nielsen J, Ponten F. Tissue-based map of the human proteome. Science 2015;347(6220).
279. Carithers LJ, Ardlie K, Barcus M, Branton PA, Britton A, Buia SA, Compton CC, DeLuca DS, Peter-Demchok J, Gelfand ET, Guan P, Korzeniewski GE, Lockhart NC, Rabiner CA, Rao AK, Robinson KL, Roche NV, Sawyer SJ, Segre AV, Shive CE, Smith AM, Sobin LH, Undale AH, Valentino KM, Vaught J, Young TR, Moore HM, Consortium G. A Novel Approach to High-Quality Postmortem Tissue Procurement: The GTEx Project. Biopreserv Biobank 2015;13(5):311-319.
280. Yang JY, Yan RX, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. Nature Methods 2015;12(1):7-8.
281. Zhang CX, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein-protein interaction information. Nucleic Acids Res 2017;45(W1):W291-W299.
282. Schmidt T, Haas J, Cassarino TG, Schwede T. Assessment of ligand-binding residue predictions in CASP9. Proteins 2011;79:126-136.
283. Dong QW, Menon R, Omenn GS, Zhang Y. Structural Bioinformatics Inspection of neXtProt PE5 Proteins in the Human Proteome. J Proteome Res 2015;14(9):3750-3761.
284. Zhang C, Zheng W, Freddolino PL, Zhang Y. MetaGO: Predicting Gene Ontology of non-homologous proteins through low-resolution protein structure prediction and protein–protein network mapping. Journal of molecular biology 2018;430(15):2256-2265.
285. Zhang C, Rahimpour M, Freddolino PL, Zhang Y. Proteome-wide Structure-Based Function Prediction Reveals Roles of Proteins Responsible for *E. coli* Fitness. US HUPO 14th Annual Conference. Minneapolis, MN, USA; 2018.
286. Menon R, Panwar B, Eksi R, Kleer C, Guan YF, Omenn GS. Computational Inferences of the Functions of Alternative/Noncanonical Splice Isoforms Specific to HER2+/ER-/PR- Breast Cancers, a Chromosome 17 C-HPP Study. J Proteome Res 2015;14(9):3519-3529.
287. Gaudet P, Michel P-A, Zahn-Zabal M, Britan A, Cusin I, Domagalski M, Duek PD, Gateau A, Gleizes A, Hinard V. The neXtProt knowledgebase on human proteins: 2017 update. Nucleic acids research 2016;45(D1):D177-D182.
288. Duek P, Gateau A, Bairoch A, Lane L. Exploring the Uncharacterized Human Proteome Using neXtProt. Journal of Proteome Research 2018;17(12):4211-4226.
289. Paik YK, Overall CM, Corrales F, Deutsch EW, Lane L, Omenn GS. Toward Completion of the Human Proteome Parts List: Progress Uncovering Proteins That Are Missing or Have Unknown Function and Developing Analytical Methods. Journal of Proteome Research 2018;17(12):4023-4030.
290. Wu ST, Zhang Y. LOMETS: A local meta-threading-server for protein structure prediction. Nucleic Acids Res 2007;35(10):3375-3382.
291. Xu Y, Xu D. Protein threading using PROSPECT: Design and evaluation. Proteins-Structure Function and Genetics 2000;40(3):343-354.
292. Yan RX, Xu D, Yang JY, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Scientific Reports 2013;3:2619.
293. Jaroszewski L, Rychlewski L, Li ZW, Li WZ, Godzik A. FFAS03: a server for profile-profile sequence alignments. Nucleic Acids Res 2005;33:W284-W288.
294. Madera M. Profile Comparer: a program for scoring and aligning profile hidden Markov models. Bioinformatics 2008;24(22):2630-2631.

295. Lobley A, Sadowski MI, Jones DT. pGenTHREADER and pDomTHREADER: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 2009;25(14):1761-1767.
296. Xu D, Jaroszewski L, Li ZW, Godzik A. FFAS-3D: improving fold recognition by including optimized structural features and template re-ranking. Bioinformatics 2014;30(5):660-667.
297. Zhou HY, Zhou YQ. Single-body residue-level knowledge-based energy score combined with sequence-profile and secondary structure information for fold recognition. Proteins 2004;55(4):1005-1013.
298. Zhang Y. I-TASSER server for protein 3D structure prediction. Bmc Bioinformatics 2008;9(1):1.
299. Skolnick J, Gao M. Interplay of physics and evolution in the likely origin of protein biochemical function. P Natl Acad Sci USA 2013;110(23):9344-9349.
300. Bateman A, Martin MJ, O'Donovan C, Magrane M, Alpi E, Antunes R, Bely B, Bingley M, Bonilla C, Britto R, Bursteinas B, Bye-A-Jee H, Cowley A, Da Silva A, De Giorgi M, Dogan T, Fazzini F, Castro LG, Figueira L, Garmiri P, Georghiou G, Gonzalez D, Hatton-Ellis E, Li WZ, Liu WD, Lopez R, Luo J, Lussi Y, MacDougall A, Nightingale A, Palka B, Pichler K, Poggioli D, Pundir S, Pureza L, Qi GY, Rosanoff S, Saidi R, Sawford T, Shypitsyna A, Speretta E, Turner E, Tyagi N, Volynkin V, Wardell T, Warner K, Watkins X, Zaru R, Zellner H, Xenarios I, Bougueleret L, Bridge A, Poux S, Redaschi N, Aimo L, Argoud-Puy G, Auchincloss A, Axelsen K, Bansal P, Baratin D, Blatter MC, Boeckmann B, Bolleman J, Boutet E, Breuza L, Casal-Casas C, de Castro E, Coudert E, Cuche B, Doche M, Dornevil D, Duvaud S, Estreicher A, Famiglietti L, Feuermann M, Gasteiger E, Gehant S, Gerritsen V, Gos A, Gruaz-Gumowski N, Hinz U, Hulo C, Jungo F, Keller G, Lara V, Lemercier P, Lieberherr D, Lombardot T, Martin X, Masson P, Morgat A, Neto T, Nouspikel N, Paesano S, Pedruzzi I, Pilbout S, Pozzato M, Pruess M, Rivoire C, Roechert B, Schneider M, Sigrist C, Sonesson K, Staehli S, Stutz A, Sundaram S, Tognolli M, Verbregue L, Veuthey AL, Wu CH, Arighi CN, Arminski L, Chen CM, Chen YX, Garavelli JS, Huang HZ, Laiho K, McGarvey P, Natale DA, Ross K, Vinayaka CR, Wang QH, Wang YQ, Yeh LS, Zhang J, Consortium U. UniProt: the universal protein knowledgebase. Nucleic Acids Res 2017;45(D1):D158-D169.
301. Belfort GM, Kandror KV. Cellugyrin and synaptogyrin facilitate targeting of synaptophysin to a ubiquitous synaptic vesicle-sized compartment in PC12 cells. Journal of Biological Chemistry 2003;278(48):47971-47978.
302. Belfort GM, Bakirtzi K, Kandror KV. Cellugyrin induces biogenesis of synaptic-like microvesicles in PC12 cells. Journal of Biological Chemistry 2005;280(8):7262-7272.
303. Zhang C, Wei X, Omenn GS, Zhang Y. Structure and Protein Interaction-based Gene Ontology Annotations Reveal Likely Functions of Uncharacterized Proteins on Human Chromosome 17. Journal of proteome research 2018;17(12):4186-4196.
304. Zhou NH, Jiang YX, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, Lewis KA, Georghiou G, Nguyen HN, Hamid MN, Davis L, Dogan T, Atalay V, Rifaioglu AS, Dalkiran A, Atalay RC, Zhang CX, Hurto RL, Freddolino PL, Zhang Y, Bhat P, Supek F, Fernandez JM, Gemovic B, Perovic VR, Davidovic RS, Sumonja N, Veljkovic N, Asgari E, Mofrad MRK, Profiti G, Savojardo C, Martelli PL, Casadio R, Boecker F, Schoof H, Kahanda I, Thurlby N, McHardy AC, Renaux A, Saidi R, Gough J, Freitas AA, Antczak M, Fabris F, Wass MN, Hou J, Cheng JL, Wang Z, Romero AE, Paccanaro A, Yang HX, Goldberg T, Zhao CG, Holm L, Toronen P, Medlar AJ, Zosa E, Borukhov I, Novikov I, Wilkins A, Lichtarge O, Chi PH, Tseng WC, Linial M, Rose PW, Dessimoz C, Vidulin V, Dzeroski S, Sillitoe I, Das S, Lees JG, Jones DT, Wan C, Cozzetto D, Fa R, Torres M, Vesztrocy AW, Rodriguez JM, Tress ML, Frasca M, Notaro M, Grossi G, Petrini A, Re M, Valentini G, Mesiti M, Roche DB, Reeb J, Ritchie DW, Aridhi S, Alborzi SZ, Devignes MD, Koo DE, Bonneau R, Gligorijevic V, Barot M, Fang H, Toppo S, Lavezzo E, Falda M, Berselli M, Tosatto SCE, Carraro M, Piovesan D, Rehman HU, Mao QZ, Zhang SS, Vucetic S, Black GS, Jo DE, Suh E, Dayton JB, Larsen DJ, Omdahl AR,

McGuffin LJ, Brackenridge DA, Babbitt PC, Yunes JM, Fontana P, Zhang F, Zhu SF, You RH, Zhang ZH, Dai SY, Yao SW, Tian WD, Cao RZ, Chandler C, Amezola M, Johnson D, Chang JM, Liao WH, Liu YW, Pascarelli S, Frank Y, Hoehndorf R, Kulmanov M, Boudellioua I, Politano G, Di Carlo S, Benso A, Hakala K, Ginter F, Mehryary F, Kaewphan S, Bjorne J, Moen H, Tolvanen MEE, Salakoski T, Kihara D, Jain A, Smuc T, Altenhoff A, Ben-Hur A, Rost B, Brenner SE, Orengo CA, Jeffery CJ, Bosco G, Hogan DA, Martin MJ, O'Donovan C, Mooney SD, Greene CS, Radivojac P, Friedberg I. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biology 2019;20(1).

305. Sali A, Blundell TL. Comparative Protein Modeling by Satisfaction of Spatial Restraints. Journal of Molecular Biology 1993;234(3):779-815.
306. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome biology 2016;17(1):184.
307. SciPy developers, scipy.stats.pearsonr. https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html (retrieved Aug 5 2019).
308. Paik YK, Lane L, Kawamura T, Chen YJ, Cho JY, LaBaer J, Yoo JS, Domont G, Corrales F, Omenn GS, Archakov A, Encarnacion-Guevara S, Lui SQ, Salekdeh GH, Cho JY, Kim CY, Overall CM. Launching the C-HPP neXt-CP50 Pilot Project for Functional Characterization of Identified Proteins with No Known Function. Journal of Proteome Research 2018;17(12):4042-4050.
309. Martin DMA, Berriman M, Barton GJ. GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. Bmc Bioinformatics 2004;5(1):178.
310. Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics 2019;in press:bioRxiv 615260.
311. Glogowska A, Kunanuvat U, Stetefeld J, Patel TR, Thanasupawat T, Krcek J, Weber E, Wong GW, Del Bigio MR, Hoang-Vu C, Hombach-Klonisch S, Klonisch T. C1q-tumour necrosis factor-related protein 8 (CTRP8) is a novel interaction partner of relaxin receptor RXFP1 in human brain cancer cells. J Pathol 2013;231(4):466-479.
312. Markolovic S, Zhuang QQ, Wilkins SE, Eaton CD, Abboud MI, Katz MJ, McNeil HE, Lesniak RK, Hall C, Struwe WB, Konietzny R, Davis S, Yang M, Ge W, Benesch JLP, Kessler BM, Ratcliffe PJ, Cockman ME, Fischer R, Wappner P, Chowdhury R, Coleman ML, Schofield CJ. The Jumonji-C oxygenase JMJD7 catalyzes (3S)-lysyl hydroxylation of TRAFAC GTPases. Nat Chem Biol 2018;14(7):688-+.
313. Liu HL, Wang C, Lee S, Deng Y, Wither M, Oh S, Ning FK, Dege C, Zhang QQ, Liu XJ, Johnson AM, Zang JY, Chen ZZ, Janknecht R, Hansen K, Marrack P, Li CY, Kappler JW, Hagman J, Zhang GY. Clipping of arginine-methylated histone tails by JMJD5 and JMJD7. Proceedings of the National Academy of Sciences of the United States of America 2017;114(37):E7717-E7726.
314. Hsieh J, Koseki M, Molusky MM, Yakushiji E, Ichi I, Westerterp M, Iqbal J, Chan RB, Abramowicz S, Tascau L, Takiguchi S, Yamashita S, Welch CL, Di Paolo G, Hussain MM, Lefkowitch JH, Rader DJ, Tall AR. TTC39B deficiency stabilizes LXR reducing both atherosclerosis and steatohepatitis. Nature 2016;535(7611):303-U282.
315. Moua P, Checketts M, Xu LG, Shu HB, Reyland ME, Cusick JK. RELT family members activate p38 and induce apoptosis by a mechanism distinct from TNFR1. Biochem Bioph Res Co 2017;491(1):25-32.
316. Cusick JK, Mustian A, Goldberg K, Reyland ME. RELT induces cellular death in HEK 293 epithelial cells. Cell Immunol 2010;261(1):1-8.
317. Perland E, Lekholm E, Eriksson MM, Bagchi S, Arapi V, Fredriksson R. The Putative SLC Transporters Mfsd5 and Mfsd11 Are Abundantly Expressed in the Mouse Brain and Have a Potential Role in Energy Homeostasis. Plos One 2016;11(6):e0156912.

318. Zjablovskaja P, Kardosova M, Danek P, Angelisova P, Benoukraf T, Wurm AA, Kalina T, Sian S, Balastik M, Delwel R, Brdicka T, Tenen DG, Behre G, Fiore F, Malissen B, Horejsi V, Alberich-Jorda M. EVI2B is a C/EBP alpha target gene required for granulocytic differentiation and functionality of hematopoietic progenitors. Cell Death Differ 2017;24(4):705-716.
319. Ji YX, Huang Z, Yang X, Wang XZ, Zhao LP, Wang PX, Zhang XJ, Alves-Bezerra M, Cai L, Zhang P, Lu YX, Bai L, Gao MM, Zhao H, Tian S, Wang Y, Huang ZX, Zhu XY, Zhang Y, Gong J, She ZG, Li F, Cohen DE, Li HL. The deubiquitinating enzyme cylindromatosis mitigates nonalcoholic steatohepatitis. Nat Med 2018;24(2):213-+.
320. Catalan-Dibene J, Vazquez MI, Luu VP, Nuccio SP, Karimzadeh A, Kastenschmidt JM, Villalta SA, Ushach I, Pone EJ, Casali P, Raffatellu M, Burkhardt AM, Hernandez-Ruiz M, Heller G, Hevezi PA, Zlotnik A. Identification of IL-40, a Novel B Cell-Associated Cytokine. J Immunol 2017;199(9):3326-3335.
321. Jiang YX, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, Funk CS, Kahanda I, Verspoor KM, Ben-Hur A, Koo DCE, Penfold-Brown D, Shasha D, Youngs N, Bonneau R, Lin A, Sahraeian SME, Martelli PL, Profiti G, Casadio R, Cao RZ, Zhong Z, Cheng JL, Altenhoff A, Skunca N, Dessimoz C, Dogan T, Hakala K, Kaewphan S, Mehryary F, Salakoski T, Ginter F, Fang H, Smithers B, Oates M, Gough J, Toronen P, Koskinen P, Holm L, Chen CT, Hsu WL, Bryson K, Cozzetto D, Minneci F, Jones DT, Chapman S, Dukka BKC, Khan IK, Kihara D, Ofer D, Rappoport N, Stern A, Cibrian-Uhalte E, Denny P, Foulger RE, Hieta R, Legge D, Lovering RC, Magrane M, Melidoni AN, Mutowo-Meullenet P, Pichler K, Shypitsyna A, Li B, Zakeri P, ElShal S, Tranchevent LC, Das S, Dawson NL, Lee D, Lees JG, Sillitoe I, Bhat P, Nepusz T, Romero AE, Sasidharan R, Yang HX, Paccanaro A, Gillis J, Sedeno-Cortes AE, Pavlidis P, Feng S, Cejuela JM, Goldberg T, Hamp T, Richter L, Salamov A, Gabaldon T, Marcet-Houben M, Supek F, Gong QT, Ning W, Zhou YP, Tian WD, Falda M, Fontana P, Lavezzo E, Toppo S, Ferrari C, Giollo M, Piovesan D, Tosatto SCE, del Pozo A, Fernandez JM, Maietta P, Valencia A, Tress ML, Benso A, Di Carlo S, Politano G, Savino A, Rehman HU, Re M, Mesiti M, Valentini G, Bargsten JW, van Dijk ADJ, Gemovic B, Glisic S, Perovic V, Veljkovic V, Veljkovic N, Almeida-e-Silva DC, Vencio RZN, Sharan M, Vogel J, Kansakar L, Zhang S, Vucetic S, Wang Z, Sternberg MJE, Wass MN, Huntley RP, Martin MJ, O'Donovan C, Robinson PN, Moreau Y, Tramontano A, Babbitt PC, Brenner SE, Linial M, Orengo CA, Rost B, Greene CS, Mooney SD, Friedberg I, Radivojac P. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol 2016;17(1):184.
322. Mushegian AR, Koonin EV. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proceedings of the National Academy of Sciences of the United States of America 1996;93(19):10268-10273.
323. Koonin EV. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol 2003;1(2):127-136.
324. Smalley DJ, Whiteley M, Conway T. In search of the minimal Escherichia coli genome. Trends Microbiol 2003;11(1):6-8.
325. Koshland DE. The seven pillars of life. Science 2002;295(5563):2215-2216.
326. Cleland CE, Chyba CF. Defining 'life'. Origins Life Evol B 2002;32(4):387-393.
327. Akerley BJ, Rubin EJ, Camilli A, Lampe DJ, Robertson HM, Mekalanos JJ. Systematic identification of essential genes by in vitro mariner mutagenesis. Proceedings of the National Academy of Sciences of the United States of America 1998;95(15):8927-8932.
328. Yu BJ, Sung BH, Koob MD, Lee CH, Lee JH, Lee WS, Kim MS, Kim SC. Minimization of the Escherichia coli genome using a Tn5-targeted Cre/loxP excision system. Nat Biotechnol 2002;20(10):1018-1023.
329. Hutchison CA, Peterson SN, Gill SR, Cline RT, White O, Fraser CM, Smith HO, Venter JC. Global transposon mutagenesis and a minimal mycoplasma genome. Science 1999;286(5447):2165-2169.

330. Sassetti CM, Boyd DH, Rubin EJ. Comprehensive identification of conditionally essential genes in mycobacteria. Proceedings of the National Academy of Sciences of the United States of America 2001;98(22):12712-12717.
331. Akerley BJ, Rubin EJ, Novick VL, Amaya K, Judson N, Mekalanos JJ. A genome-scale analysis for identification of genes required for growth or survival of Haemophilus influenzae. Proceedings of the National Academy of Sciences of the United States of America 2002;99(2):966-971.
332. Judson N, Mekalanos JJ. TnAraOut, A transposon-based approach to identify and characterize essential bacterial genes. Nat Biotechnol 2000;18(7):740-745.
333. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El Bakkoury M, Bangham R, Benito R, Brachat S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo CY, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelmy J, Winzeler EA, Yang YH, Yen G, Youngman E, Yu KX, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M. Functional profiling of the Saccharomyces cerevisiae genome. Nature 2002;418(6896):387-391.
334. Posfai G, Plunkett G, Feher T, Frisch D, Keil GM, Umenhoffer K, Kolisnychenko V, Stahl B, Sharma SS, de Arruda M, Burland V, Harcum SW, Blattner FR. Emergent properties of reduced-genome Escherichia coli. Science 2006;312(5776):1044-1046.
335. Umenhoffer K, Feher T, Baliko G, Ayaydin F, Posfai J, Blattner FR, Posfai G. Reduced evolvability of Escherichia coli MDS42, an IS-less cellular chassis for molecular and synthetic biology applications. Microb Cell Fact 2010;9.
336. Csorgo B, Feher T, Timar E, Blattner FR, Posfai G. Low-mutation-rate, reduced-genome Escherichia coli: an improved host for faithful maintenance of engineered genetic constructs. Microb Cell Fact 2012;11.
337. Gibson DG, Glass JI, Lartigue C, Noskov VN, Chuang RY, Algire MA, Benders GA, Montague MG, Ma L, Moodie MM, Merryman C, Vashee S, Krishnakumar R, Assad-Garcia N, Andrews-Pfannkoch C, Denisova EA, Young L, Qi ZQ, Segall-Shapiro TH, Calvey CH, Parmar PP, Hutchison CA, Smith HO, Venter JC. Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. Science 2010;329(5987):52-56.
338. Hutchison CA, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L, Pelletier JF, Qi ZQ, Richter RA, Strychalski EA, Sun LJ, Suzuki Y, Tsvetanova B, Wise KS, Smith HO, Glass JI, Merryman C, Gibson DG, Venter JC. Design and synthesis of a minimal bacterial genome. Science 2016;351(6280).
339. Breuer M, Earnest TM, Merryman C, Wise KS, Sun LJ, Lynottn MR, Hutchison CA, Smith HO, Lapek JD, Gonzalez DJ, De Crecy-Lagard V, Haas D, Hanson AD, Labhsetwar P, Glass JI, Luthey-Schulten Z. Essential metabolism for a minimal cell. Elife 2019;8.
340. Danchin A, Fang G. Unknown unknowns: essential genes in quest for function. Microb Biotechnol 2016;9(5):530-540.
341. Yang ZY, Tsui SKW. Functional Annotation of Proteins Encoded by the Minimal Bacterial Genome Based on Secondary Structure Element Alignment. Journal of Proteome Research 2018;17(7):2511-2520.
342. Antczak M, Michaelis M, Wass MN. Environmental conditions shape the nature of a minimal bacterial genome. Nat Commun 2019;10.
343. Zhang C, Zheng W, Freddolino PL, Zhang Y. MetaGO: Predicting Gene Ontology of Non-homologous Proteins Through Low-Resolution Protein Structure Prediction and Protein Protein Network Mapping. Journal of Molecular Biology 2018;430(15):2256-2265.

344. Zhang C, Lane L, Omenn GS, Zhang Y. Blinded Testing of Function Annotation for uPE1 Proteins by I-TASSER/COFACTOR Pipeline Using the 2018–2019 Additions to neXtProt and the CAFA3 Challenge. Journal of proteome research 2019;18(12):4154-4166.
345. Li Y, Hu J, Zhang C, Yu D-J, Zhang Y. ResPRE: high-accuracy protein contact prediction by coupling precision matrix with deep residual neural networks. Bioinformatics 2019:btz291.
346. Yang J, Roy A, Zhang Y. BioLiP: a semi-manually curated database for biologically relevant ligand–protein interactions. Nucleic acids research 2012;41(D1):D1096-D1103.
347. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, O'Donovan C. The GOA database: Gene Ontology annotation updates for 2015. Nucleic Acids Research 2015;43(D1):D1057-D1063.
348. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering C. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Research 2019;47(D1):D607-D613.
349. Hill DP, Davis AP, Richardson JE, Corradi JP, Ringwald M, Eppig JT, Blake JA. Strategies for biological annotation of mammalian systems: Implementing gene ontologies in mouse genome informatics. Genomics 2001;74(1):121-128.
350. Wei X, Zhang C, Freddolino PL, Zhang Y. Detecting Gene Ontology misannotations using taxon-specific rate ratio comparisons. Bioinformatics 2020.
351. Guerler A, Govindarajoo B, Zhang Y. Mapping Monomeric Threading to Protein-Protein Structure Prediction. J Chem Inf Model 2013;53(3):717-725.
352. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Research 2004;32:D449-D451.
353. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Research 2003;31(1):371-373.
354. Jones P, Binns D, Chang HY, Fraser M, Li WZ, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. InterProScan 5: genome-scale protein function classification. Bioinformatics 2014;30(9):1236-1240.
355. Karpowich NK, Song JM, Wang DN. An Aromatic Cap Seals the Substrate Binding Site in an ECF-Type S Subunit for Riboflavin. Journal of Molecular Biology 2016;428(15):3118-3130.
356. Zhang P, Wang JW, Shi YG. Structure and mechanism of the S component of a bacterial ECF transporter. Nature 2010;468(7324):717-U148.
357. Rajagopala SV, Sikorski P, Kumar A, Mosca R, Vlasblom J, Arnold R, Franca-Koh J, Pakala SB, Phanse S, Ceol A, Hauser R, Siszler G, Wuchty S, Emili A, Babu M, Aloy P, Pieper R, Uetz P. The binary protein-protein interaction landscape of Escherichia coli. Nat Biotechnol 2014;32(3):285-290.
358. Chen MC, Li Y, Zhu YH, Ge F, Yu DJ. SSCpred: Single-Sequence-Based Protein Contact Prediction Using Deep Fully Convolutional Network. J Chem Inf Model 2020;60(6):3295-3303.
359. Rives A, Goyal S, Meier J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. bioRxiv 2019:622803.
360. Rao RS, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, Abbeel P, Song YS. Evaluating Protein Transfer Learning with TAPE. Adv Neur In 2019;32.
361. Du ZY, Pan S, Wu Q, Peng ZL, Yang JY. CATHER: a novel threading algorithm with predicted contacts. Bioinformatics 2020;36(7):2119-2125.
362. Zhang H, Shen Y. Template-based prediction of protein structure with deep learning. bioRxiv 2020.
363. Gao M, Skolnick J. A novel sequence alignment algorithm based on deep learning of the protein folding code. Bioinformatics 2020.

364. AlQuraishi M. End-to-End Differentiable Learning of Protein Structure. Cell Syst 2019;8(4):292-+.
365. Ingraham J, Riesselman AJ, Sander C, Marks DS. Learning Protein Structure with a Differentiable Simulator. 2019.
366. Li J. Universal Transforming Geometric Network. arXiv preprint arXiv:190800723 2019.
367. Derevyanko G, Lamoureux G. TorchProteinLibrary: A computationally efficient, differentiable representation of protein structure. arXiv preprint arXiv:181201108 2018.
368. Wan C, Jones DT. Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. Nature Machine Intelligence 2020.