

# The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens

Naihui Zhou<sup>1,2</sup>, Yuxiang Jiang<sup>3</sup>, Timothy R Bergquist<sup>4</sup>, Alexandra J Lee<sup>5</sup>, Balint Z Kacsóh<sup>6,7</sup>, Alex W Crocker<sup>8</sup>, Kimberley A Lewis<sup>8</sup>, George Georghiou<sup>9</sup>, Huy N Nguyen<sup>1,10</sup>, Md Nafiz Hamid<sup>1,2</sup>, Larry Davis<sup>2</sup>, Tunca Dogan<sup>12,13</sup>, Volkan Atalay<sup>14</sup>, Ahmet S Rifaioglu<sup>14,16</sup>, Alperen Dalkiran<sup>17</sup>, Rengul Cetin-Atalay<sup>18</sup>, Chengxin Zhang<sup>19</sup>, Rebecca L Hurto<sup>20</sup>, Peter L Freddolino<sup>21,22</sup>, Yang Zhang<sup>23,24</sup>, Prajwal Bhat<sup>25</sup>, Fran Supek<sup>26,27</sup>, José M Fernández<sup>28,29</sup>, Branislava Gemovic<sup>30</sup>, Vladimir R Perovic<sup>31</sup>, Radoslav S Davidovic<sup>30</sup>, Neven Sumonja<sup>30</sup>, Nevena Veljkovic<sup>30</sup>, Ehsaneddin Asgari<sup>32,33</sup>, Mohammad RK Mofrad<sup>34</sup>, Giuseppe Profiti<sup>35,36</sup>, Castrense Savojarado<sup>37</sup>, Pier Luigi Martelli<sup>37</sup>, Rita Casadio<sup>37</sup>, Florian Boecker<sup>38</sup>, Indika Kahanda<sup>39</sup>, Natalie Thurlby<sup>40</sup>, Alice C McHardy<sup>41,42</sup>, Alexandre Renaux<sup>43,44</sup>, Rabie Saidi<sup>45</sup>, Julian Gough<sup>46</sup>, Alex A Freitas<sup>47</sup>, Magdalena Antczak<sup>48</sup>, Fabio Fabris<sup>47</sup>, Mark N Wass<sup>49</sup>, Jie Hou<sup>50,51</sup>, Jianlin Cheng<sup>51</sup>, Jie Hou<sup>50,51</sup>, Zheng Wang<sup>52</sup>, Alfonso E Romero<sup>53</sup>, Alberto Paccanaro<sup>54</sup>, Haixuan Yang<sup>55</sup>, Tatyana Goldberg<sup>56</sup>, Chenguang Zhao<sup>57</sup>, Liisa Holm<sup>58</sup>, Petri Törönen<sup>58</sup>, Alan J Medlar<sup>58</sup>, Elaine Zosa<sup>58</sup>, Itamar Borukhov<sup>59</sup>, Ilya Novikov<sup>60</sup>, Angela Wilkins<sup>61</sup>, Olivier Lichtarge<sup>61</sup>, Po-Han Chi<sup>62</sup>, Wei-Cheng Tseng<sup>63</sup>, Michal Linial<sup>64</sup>, Peter W Rose<sup>65</sup>, Christophe Dessimoz<sup>66,67</sup>, Vedrana Vidulin<sup>68</sup>, Saso Dzeroski<sup>69,70</sup>, Ian Sillitoe<sup>71</sup>, Sayoni Das<sup>72</sup>, Jonathan Gill Lees<sup>73,74</sup>, David T Jones<sup>75,76</sup>, Cen Wan<sup>75,76</sup>, Domenico Cozzetto<sup>75,76</sup>, Rui Fa<sup>75,76</sup>, Mateo Torres<sup>53</sup>, Alex Wiarwick Vesztröcy<sup>77,78</sup>, Jose Manuel Rodriguez<sup>79</sup>, Michael L Tress<sup>80</sup>, Marco Frasca<sup>81</sup>, Marco Notaro<sup>81</sup>, Giuliano Grossi<sup>81</sup>, Alessandro Petrini<sup>81</sup>, Matteo Re<sup>81</sup>, Giorgio Valentini<sup>81</sup>, Marco Mesiti<sup>81</sup>, Daniel B Roche<sup>82</sup>, Jonas Reeb<sup>83</sup>, David W Ritchie<sup>84</sup>, Sabeur Aridhi<sup>84</sup>, Seyed Ziaeddin Alborzi<sup>85,86</sup>, Marie-Dominique Devignes<sup>85,87</sup>, Da Chen Emily Koo<sup>88</sup>, Richard Bonneau<sup>89,90</sup>, Vladimir Gligorijević<sup>91</sup>, Meet Barot<sup>92</sup>, Hai Fang<sup>93</sup>, Stefano Toppo<sup>94</sup>, Enrico Lavezzo<sup>94</sup>, Marco Falda<sup>95</sup>, Michele Berselli<sup>94</sup>, Silvio CE Tosatto<sup>96,97</sup>, Marco Carraro<sup>98</sup>, Damiano Piovesan<sup>99</sup>, Hafeez Ur Rehman<sup>100</sup>, Qizhong Mao<sup>101,102</sup>, Shanshan Zhang<sup>103</sup>, Slobodan Vucetic<sup>104</sup>, Gage S Black<sup>105,106</sup>, Dane Jo<sup>105,106</sup>, Dallas J Larsen<sup>105,106</sup>, Ashton R Omdahl<sup>105,106</sup>, Luke W Sagers<sup>105,106</sup>, Erica Suh<sup>105,106</sup>, Jonathan B Dayton<sup>105,106</sup>, Liam J McGuffin<sup>107</sup>, Danielle A Brackenridge<sup>107</sup>, Patricia C Babbitt<sup>108,109</sup>, Jeffrey M Yunes<sup>110,111</sup>, Paolo Fontana<sup>112</sup>, Feng Zhang<sup>113,114</sup>, Shanfeng Zhu<sup>115</sup>, Ronghui You<sup>115</sup>, Zihan Zhang<sup>115</sup>, Suyang Dai<sup>116</sup>, Shuwei Yao<sup>117</sup>, Weidong Tian<sup>113,114</sup>, Renzhi Cao<sup>118</sup>, Caleb Chandler<sup>118</sup>, Miguel Amezola<sup>118</sup>, Devon Johnson<sup>118</sup>, Jia-Ming Chang<sup>119</sup>, Wen-Hung Liao<sup>119</sup>, Yi-Wei Liu<sup>119</sup>, Stefano Pascarelli<sup>120</sup>, Yotam Frank<sup>121</sup>, Robert Hoehndorf<sup>122</sup>, Maxat Kulmanov<sup>123</sup>, Imane Boudelloua<sup>124,125</sup>, Gianfranco Politano<sup>126</sup>, Stefano Di Carlo<sup>126</sup>, Alfredo Benso<sup>126</sup>, Kai Hakala<sup>127,128</sup>, Filip Ginter<sup>127,129</sup>, Farrokh Mehryary<sup>127,128</sup>, Suwisa Kaewphan<sup>130,131</sup>, Jari Björne<sup>132,133</sup>, Hans Moen<sup>134</sup>, Martti E E Tolvanen<sup>135</sup>, Tapio Salakoski<sup>132,133</sup>, Daisuke Kihara<sup>136,137</sup>, Aashish Jain<sup>138</sup>, Tomislav Šmuc<sup>139</sup>, Adrian Altenhoff<sup>140,141</sup>, Asa Ben-Hur<sup>142</sup>, Burkhard Rost<sup>143,144</sup>, Steven E Brenner<sup>145</sup>, Christine A Orengo<sup>72</sup>, Constance J Jeffery<sup>146</sup>, Giovanni Bosco<sup>147</sup>, Deborah A Hogan<sup>8</sup>, Maria J Martin<sup>9</sup>, Claire O'Donovan<sup>9</sup>, Sean D Mooney<sup>4</sup>, Casey S Greene<sup>148,149</sup>, Predrag Radivojac<sup>150</sup>, and Iddo Friedberg<sup>1,2</sup>

<sup>1</sup> *Veterinary Microbiology and Preventive Medicine, Iowa State University*

<sup>2</sup> *Program in Bioinformatics and Computational Biology, Iowa State University, Ames, IA, USA*

<sup>3</sup> *Indiana University Bloomington, Bloomington, Indiana, USA*

<sup>4</sup> *Department of Biomedical Informatics and Medical Education, University of Washington, Seattle, WA, USA*

<sup>5</sup> *Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA, USA*

<sup>6</sup> *Department of Molecular and Systems Biology, Geisel School of Medicine at Dartmouth*

<sup>7</sup> *Department of Molecular and Systems Biology, Hanover, NH, USA*

<sup>8</sup> *Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, USA*

<sup>9</sup> *European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, United Kingdom*

<sup>10</sup> *Program in Computer Science, Iowa State University, Ames, IA, USA*

- <sup>11</sup> Program in Bioinformatics and Computational Biology, Ames, IA, USA  
<sup>12</sup> Graduate School of Informatics, Middle East Technical University (METU)  
<sup>13</sup> European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI)  
<sup>14</sup> Department of Computer Engineering, Middle East Technical University (METU)  
<sup>16</sup> Department of Computer Engineering, Iskenderun Technical University, Hatay, Turkey, Ankara, Turkey  
<sup>17</sup> Department of Computer Engineering, Middle East Technical University (METU), Ankara, Turkey  
<sup>18</sup> CanSyL, Graduate School of Informatics, Middle East Technical University (METU), Ankara, Select a State or Province, Turkey  
<sup>19</sup> Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA  
<sup>20</sup> Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA  
<sup>21</sup> Department of Biological Chemistry, University of Michigan  
<sup>22</sup> Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA  
<sup>23</sup> Department of Computational Medicine and Bioinformatics, University of Michigan  
<sup>24</sup> Department of Biological Chemistry, University of Michigan, Ann Arbor, MI, USA  
<sup>25</sup> Achira Labs, Bangalore, India  
<sup>26</sup> Institute for Research in Biomedicine (IRB Barcelona)  
<sup>27</sup> Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain  
<sup>28</sup> INB Coordination Unit, Life Sciences Department, Barcelona Supercomputing Center  
<sup>29</sup> (former) INB GN2, Structural and Computational Biology Programme, Spanish National Cancer Research Centre, Barcelona, Catalonia, Spain  
<sup>30</sup> Laboratory for Bioinformatics and Computational Chemistry, Institute of Nuclear Sciences VINCA, University of Belgrade, Belgrade, Serbia  
<sup>31</sup> Laboratory for Bioinformatics and Computational Chemistry, Institute of Nuclear Sciences VINCA, University of Belgrade, Belgrade, Serbia  
<sup>32</sup> Molecular Cell Biomechanics Laboratory, Departments of Bioengineering, University of California Berkeley  
<sup>33</sup> Computational Biology of Infection Research, Helmholtz Centre for Infection Research, Berkeley, CA, USA  
<sup>34</sup> Departments of Bioengineering and Mechanical Engineering, Berkeley, CA, USA  
<sup>35</sup> Bologna Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy  
<sup>36</sup> National Research Council, IBIOM, Bologna, Italy  
<sup>37</sup> Bologna Biocomputing Group, Department of Pharmacy and Biotechnology, University of Bologna, Italy, Bologna, Italy  
<sup>38</sup> University of Bonn: INRES Crop Bioinformatics, Bonn, North Rhine-Westphalia, Germany  
<sup>39</sup> Gianforte School of Computing, Montana State University, Bozeman, Montana, USA  
<sup>40</sup> University of Bristol, Computer Science, Bristol, Bristol, United Kingdom  
<sup>41</sup> Computational Biology of Infection Research, Helmholtz Centre for Infection Research  
<sup>42</sup> RESIST, DFG Cluster of Excellence 2155, Brunswick, Germany  
<sup>43</sup> Interuniversity Institute of Bioinformatics in Brussels, Université libre de Bruxelles - Vrije Universiteit Brussel  
<sup>44</sup> Machine Learning Group, Artificial Intelligence lab, Vrije Universiteit Brussel, Brussels, Belgium  
<sup>45</sup> European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Cambridge, UK  
<sup>46</sup> MRC Laboratory of Molecular Biology, Cambridge, United Kingdom  
<sup>47</sup> University of Kent, School of Computing, Canterbury, United Kingdom  
<sup>48</sup> School of Biosciences, University of Kent, Canterbury, United Kingdom  
<sup>49</sup> School of Biosciences, University of Kent, Canterbury, Kent, United Kingdom  
<sup>50</sup> University of Missouri, Computer Science, Columbia, Missouri, USA  
<sup>51</sup> Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, Missouri, USA  
<sup>52</sup> University of Miami, Coral Gables, Florida, USA  
<sup>53</sup> Centre for Systems and Synthetic Biology, Department of Computer Science, Royal Holloway, University of London, Egham, Surrey, United Kingdom  
<sup>54</sup> Centre for Systems and Synthetic Biology, Department of Computer Science, Royal Holloway, University of London, Egham, United Kingdom  
<sup>55</sup> School of Mathematics, Statistics and Applied Mathematics. National University of Ireland, Galway, Galway, Ireland  
<sup>56</sup> Department of Informatics, Bioinformatics & Computational Biology, Technical University of Munich, Germany, Munich, Germany  
<sup>57</sup> School of Computing Sciences and Computer Engineering, University of Southern Mississippi, Hattiesburg, Mississippi, USA  
<sup>58</sup> Institute of Biotechnology, University of Helsinki, Helsinki, Finland  
<sup>59</sup> Compugen Ltd., Holon, Israel  
<sup>60</sup> Baylor College of Medicine, Department of Biochemistry and Molecular Biology, Houston, TX, USA  
<sup>61</sup> Baylor College of Medicine, Department of Molecular and Human Genetics, Houston, TX, USA

- <sup>62</sup> National TsingHua University, Hsinchu, Taiwan
- <sup>63</sup> Department of Electrical Engineering in National Tsing Hua University, Hsinchu City, Taiwan
- <sup>64</sup> The Hebrew University of Jerusalem , Jerusalem, Israel
- <sup>65</sup> University of California San Diego, San Diego Supercomputer Center, La Jolla, California, USA
- <sup>66</sup> Department of Computational Biology and Center for Integrative Genomics, University of Lausanne, Switzerland
- <sup>67</sup> Department of Genetics, Evolution & Environment, and Department of Computer Science, University College London, UK, Lausanne, Switzerland
- <sup>68</sup> Department of Knowledge Technologies, Jozef Stefan Institute, Ljubljana, Slovenia
- <sup>69</sup> Jozef Stefan Institute
- <sup>70</sup> Jozef Stefan International Postgraduate School, Ljubljana, Slovenia
- <sup>71</sup> Research Dept. of Structural and Molecular Biology, University College London, London, England
- <sup>72</sup> Research Dept. of Structural and Molecular Biology, University College London, London, United Kingdom
- <sup>73</sup> Research Dept. of Structural and Molecular Biology, University College London
- <sup>74</sup> Oxford Brookes University, Department of Health and Life Sciences, Oxford, UK
- <sup>75</sup> University College London, Department of Computer Science
- <sup>76</sup> The Francis Crick Institute, Biomedical Data Science Laboratory, London, United Kingdom
- <sup>77</sup> Department of Genetics, Evolution and Environment, University College London, Gower Street, London, WC1E 6BT, United Kingdom
- <sup>78</sup> SIB Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland, London, United Kingdom
- <sup>79</sup> Cardiovascular Proteomics Laboratory, Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), Madrid, Spain
- <sup>80</sup> Bioinformatics Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain
- <sup>81</sup> Università degli Studi di Milano - Computer Science Dept. - AnacletoLab, Milan, Milan, Italy
- <sup>82</sup> Institut de Biologie Computationnelle, LIRMM, CNRS-UMR 5506, Université de Montpellier, Montpellier, France
- <sup>83</sup> Department of Informatics, Chair of Bioinformatics and Computational Biology, Technical University of Munich, Germany, Munich, Germany
- <sup>84</sup> University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France, Nancy, France
- <sup>85</sup> University of Lorraine, CNRS, Inria, LORIA, 54000 Nancy, France
- <sup>86</sup> University of Lorraine, Nancy, Lorraine, France
- <sup>87</sup> Inria, Nancy, France
- <sup>88</sup> Department of Biology, New York University, New York, NY, USA
- <sup>89</sup> NYU Center for Data Science, New York NY 10010
- <sup>90</sup> Flatiron Institute, CCB, NY NY 10010, New York, NY, USA
- <sup>91</sup> Center for Computational Biology (CCB), Flatiron Institute, Simons Foundation, New York, NY, USA
- <sup>92</sup> Center for Data Science, New York University, New York, NY 10011, USA, New York, NY, USA
- <sup>93</sup> Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK
- <sup>94</sup> University of Padova, Department of Molecular Medicine, Padova, Italy
- <sup>95</sup> Dept. of Biology - University of Padova, Padova, Italy
- <sup>96</sup> Department of Biomedical Sciences, University of Padua
- <sup>97</sup> CNR Institute of Neuroscience, Padova, Italy
- <sup>98</sup> Department of Biomedical Sciences, University of Padua, Padova, Padova, Italy
- <sup>99</sup> Department of Biomedical Sciences, University of Padua, Padova, Italy
- <sup>100</sup> Department of Computer Science, National University of Computer and Emerging Sciences, Peshawar, Pakistan., Peshawar, Khyber Pakhtoonkhwa, Pakistan
- <sup>101</sup> Temple University
- <sup>102</sup> University of California, Riverside, Philadelphia, PA, USA
- <sup>103</sup> Temple University, Philadelphia, PA, USA
- <sup>104</sup> Temple University, Department of Computer and Information Sciences, Philadelphia, PA, USA
- <sup>105</sup> Department of Biology, Brigham Young University
- <sup>106</sup> Bioinformatics Research Group, Provo, UT, USA
- <sup>107</sup> School of Biological Sciences, University of Reading, Reading, England, United Kingdom
- <sup>108</sup> Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco
- <sup>109</sup> Department of Pharmaceutical Chemistry, University of California, San Francisco, San Francisco, CA, USA
- <sup>110</sup> UC Berkeley - UCSF Graduate Program in Bioengineering, University of California, San Francisco, CA 94158, USA
- <sup>111</sup> Department of Bioengineering and Therapeutic Sciences, University of California, San Francisco, CA 94158, USA, San Francisco, California, USA
- <sup>112</sup> Research and Innovation Center, Edmund Mach Foundation, 38010S. Michele all'Adige, Italy, San Michele all'Adige, Italy
- <sup>113</sup> State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, School

- of Life Sciences, Fudan University*
- <sup>114</sup>*Department of Pediatrics, Brain Tumor Center, Division of Experimental Hematology and Cancer Biology, Shanghai, Shanghai, China*
- <sup>115</sup>*School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China*
- <sup>116</sup>*School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China*
- <sup>117</sup>*School of Computer Science and Shanghai Key Lab of Intelligent Information Processing, Fudan University, Shanghai, China*
- <sup>118</sup>*Pacific Lutheran University, Department of Computer Science, Tacoma, WA, USA*
- <sup>119</sup>*Department of Computer Science, National Chengchi University, Taipei, Taiwan*
- <sup>120</sup>*Okinawa Institute of Science and Technology, Tancha, Okinawa, Japan*
- <sup>121</sup>*Tel Aviv University, Tel Aviv, Israel*
- <sup>122</sup>*Computer, Electrical and Mathematical Sciences & Engineering Division, Computational Bioscience Research Center, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*
- <sup>123</sup>*King Abdullah University of Science and Technology, Computational Bioscience Research Center, Thuwal, Jeddah, Saudi Arabia*
- <sup>124</sup>*Computational Bioscience Research Center (CBRC), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia*
- <sup>125</sup>*Computer, Electrical and Mathematical Sciences Engineering Division (CEMSE), King Abdullah University of Science and Technology, Thuwal, Saudi Arabia, Thuwal, Saudi Arabia*
- <sup>126</sup>*Politecnico di Torino, Control and Computer Engineering Department, Torino, TO, Italy*
- <sup>127</sup>*University of Turku, Department of Future Technologies, Turku NLP Group*
- <sup>128</sup>*University of Turku Graduate School (UTUGS), Turku, Finland*
- <sup>129</sup>*University of Turku, Turku, Finland*
- <sup>130</sup>*Turku Centre for Computer Science (TUCS)*
- <sup>131</sup>*University of Turku, Department of Future Technologies, Turku, Finland*
- <sup>132</sup>*Department of Future Technologies, Faculty of Science and Engineering, University of Turku, FI-20014, Turku, Finland*
- <sup>133</sup>*Turku Centre for Computer Science (TUCS), Agora, Vesilinnantie 3, FI-20500 TURKU, Turku, Finland*
- <sup>134</sup>*University of Turku, Faculty of Science and Engineering, Department of Future Technologies, Turku, Finland*
- <sup>135</sup>*University of Turku, Department of Future Technologies, Turku, Finland*
- <sup>136</sup>*Department of Biological Sciences, Department of Computer Science, Purdue University, West Lafayette, IN, 47907, USA*
- <sup>137</sup>*Department of Pediatrics, University of Cincinnati, Cincinnati, OH, 45229, USA, West Lafayette, IN, USA*
- <sup>138</sup>*Department of Computer Science, Purdue University, West Lafayette, IN, USA*
- <sup>139</sup>*Division of Electronics, Rudjer Boskovic Institute, Zagreb, Croatia*
- <sup>140</sup>*Department of Computer Science, ETH Zurich*
- <sup>141</sup>*SIB Swiss Institute of Bioinformatics, Zurich, Switzerland*
- <sup>142</sup>*Department of Computer Science, Colorado State University, Fort Collins, CO, USA*
- <sup>143</sup>*Department of Informatics, Technical University of Munich, Germany*
- <sup>144</sup>*Institute for Food and Plant Sciences WZW, Technical University of Munich, Weihenstephan, Germany, Munich, Germany*
- <sup>145</sup>*University of California, Berkeley, Berkeley, CA, USA*
- <sup>146</sup>*Biological Sciences, University of Illinois at Chicago, Chicago, Illinois, USA*
- <sup>147</sup>*Department of Microbiology and Immunology, Geisel School of Medicine at Dartmouth, Hanover, NH, US*
- <sup>148</sup>*Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania*
- <sup>149</sup>*Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation, Philadelphia, Pennsylvania, USA*
- <sup>150</sup>*Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA*

## Abstract

The Critical Assessment of Functional Annotation (CAFA) is an ongoing, global, community-driven effort to evaluate and improve the computational annotation of protein function. Here we report on the results of the third CAFA challenge, CAFA3, that featured an expanded analysis over the previous CAFA rounds, both in terms of volume of data analyzed and the types of analysis performed. In a novel and major new development, computational predictions and assessment goals drove some of the experimental assays, resulting in new functional annotations for more than 1000 genes. Specifically, we performed experimental whole-genome mutation screening in *Candida albicans* and *Pseudomonas aureginosa* genomes, which provided us with genome-wide experimental data for genes associated with biofilm formation and motility (*P. aureginosa* only). We further performed targeted assays on selected genes in *Drosophila melanogaster*, which we suspected of being involved in long-term memory. We conclude that, while predictions of the molecular function and biological process annotations have slightly improved over time, those of the cellular component have not. Term-centric prediction of experimental annotations remains equally challenging; although the performance of the top methods is significantly better than expectations set by baseline methods in *C. albicans* and *D. melanogaster*, it leaves considerable room and need for improvement. We finally report that the CAFA community now involves a broad range of participants with expertise in bioinformatics, biological experimentation, biocuration, and bio-ontologies, working together to improve functional annotation, computational function prediction, and our ability to manage big data in the era of large experimental screens.

## 1 Introduction

High-throughput nucleic acid sequencing (1) and mass-spectrometry proteomics (2) have provided us with a deluge of data for DNA, RNA, and proteins in diverse species. However, extracting detailed functional information from such data remains one of the recalcitrant challenges in the life sciences and biomedicine. Low-throughput biological experiments often provide highly informative empirical data related to various functional aspects of a gene product, but these experiments are limited by time and cost. At the same time, high-throughput experiments, while providing large amounts of data, often provide information that is not specific enough to be useful (3). For these reasons, it is important to explore computational strategies for transferring functional information from the group of functionally characterized macromolecules to others that have not been studied for particular activities (4, 5, 6, 7, 8, 9).

To address the growing gap between high-throughput data and deep biological insight, a variety of computational methods that predict protein function have been developed over the years (10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24). This explosion in the number of methods is accompanied by the need to understand how well they perform, and what improvements are needed to satisfy the needs of the life sciences community. The Critical Assessment of Functional Annotation (CAFA) is a community challenge

16 that seeks to bridge the gap between the ever-expanding pool of molecular data and the limited resources  
17 available to understand protein function (25, 26, 27).

18 The first two CAFA challenges were carried out in 2010-2011 (25) and 2013-2014 (26). In CAFA1 we  
19 adopted a time-delayed evaluation method, where protein sequences that lacked experimentally verified  
20 annotations, or *targets*, were released for prediction. After the submission deadline for predictions, a subset  
21 of these targets accumulated experimental annotations over time, either as a consequence of new publications  
22 about these proteins or the biocuration work updating the annotation databases. The members of this set of  
23 proteins were used as *benchmarks* for evaluating the participating computational methods, as the function  
24 was revealed only after the prediction deadline.

25 CAFA2 expanded the challenge founded in CAFA1. The expansion included the number of ontologies  
26 used for predictions, the number of target and benchmark proteins, and the introduction of new assessment  
27 metrics that mitigate the problems with functional similarity calculation over concept hierarchies such as  
28 Gene Ontology (28). Importantly, we provided evidence that the top-scoring methods in CAFA2 outper-  
29 formed the top scoring methods in CAFA1, highlighting that methods participating in CAFA improved over  
30 the three year period. Much of this improvement came as a consequence of novel methodologies with some  
31 effect of the expanded annotation databases (26). Both CAFA1 and CAFA2 have shown that computa-  
32 tional methods designed to perform function prediction outperform a conventional function transfer through  
33 sequence similarity (25, 26).

34 In CAFA3 (2016-2017) we continued with all types of evaluations from the first two challenges and  
35 additionally performed experimental screens to identify genes associated with specific functions. This allowed  
36 us to provide unbiased evaluation of the term-centric performance based on a unique set of benchmarks  
37 obtained by assaying *Candida albicans*, *Pseudomonas aeruginosa* and *Drosophila melanogaster*. We also  
38 held a challenge following CAFA3, dubbed CAFA- $\pi$ , to provide the participating teams another opportunity  
39 to develop or modify prediction models. The genome-wide screens on *C. albicans* identified 240 genes  
40 previously not known to be involved in biofilm formation, whereas the screens on *P. aeruginosa* identified  
41 532 new genes involved in biofilm formation and 403 genes involved in motility. Finally, we used CAFA  
42 predictions to select genes from *D. melanogaster* and assay them for long-term memory involvement. This  
43 experiment allowed us to both evaluate prediction methods and identify eleven new fly genes involved in this

44 biological process (29). Here we present the outcomes of the CAFA3 challenge, as well as the accompanying  
45 challenge CAFA- $\pi$ , and discusses further directions for the community interested in the function of biological  
46 macromolecules.

## 47 **2 Results**

### 48 **2.1 Top methods have slightly improved since CAFA2**

49 One of CAFA's major goals is to quantify the progress in function prediction over time. We therefore  
50 conducted comparative evaluation of top CAFA1, CAFA2, and CAFA3 methods according to their ability  
51 to predict Gene Ontology (28) terms on a set of common benchmark proteins. This benchmark set was  
52 created as an intersection of CAFA3 benchmarks (proteins that gained experimental annotation after the  
53 CAFA3 prediction submission deadline), and CAFA1 and CAFA2 target proteins. Overall, this set contained  
54 377 protein sequences with annotations in the Molecular Function Ontology (MFO), 717 sequences in the  
55 Biological Process Ontology (BPO) and 548 sequences in the Cellular Component Ontology (CCO), which  
56 allowed for a direct comparison of all methods that have participated in the challenges so far. The head-  
57 to-head comparisons in MFO, BPO, and CCO between top five CAFA3 and CAFA2 methods are shown in  
58 Figure 1. CAFA3 and CAFA1 comparisons are shown in Figure S1 in the Supplemental Materials.

59 We first observe that, in effect, the performance of baseline methods (25, 26) has not improved since  
60 CAFA2. The Naïve method, which uses the term frequency in the existing annotation database as prediction  
61 score for every input protein, has the same  $F_{\max}$  performance using both annotation database in 2014 (when  
62 CAFA2 was held) and in 2017 (when CAFA3 was held), which suggests little change in term frequencies in the  
63 annotation database since 2014. On the other hand, BLAST-based annotation transfer, tells a contrasting  
64 tale between ontologies. In MFO, the BLAST method based on the existing annotations in 2017 is slightly  
65 but significantly better than the BLAST method based on 2014 training data. In BPO and CCO, however,  
66 the BLAST based on the later database has not outperformed its earlier counterpart, although the changes  
67 in effect size (absolute change in  $F_{\max}$ ) in both ontologies are small.

68 When surveying all three CAFA challenges, the performance of both baseline methods has been relatively  
69 stable, with some fluctuations of BLAST. Such performance of direct sequence-based function transfer is

70 surprising, given the steady growth of annotations in UniProt-GOA (30); i.e., there were 259,785 experimental  
71 annotations in 2011, 341,938 in 2014 and 434,973 in 2017, but there does not seem to be a definitive trend  
72 with the BLAST method, as they go up and down in  $F_{\max}$  across ontologies. We conclude from these  
73 observations on the baseline methods that first, the ontologies are in different annotation states and should  
74 not be treated as a whole. Second, methods that perform direct function transfer based on sequence similarity  
75 do not necessarily benefit from a larger training dataset. Although the performance observed in our work is  
76 also dependent on the benchmark set, it appears that the annotation databases remain sparsely populated to  
77 effectively exploit function transfer by sequence similarity, thus justifying the need for advanced methodology  
78 development for this problem.

79 [Figure 1 about here.]

80 Head-to-head comparisons of the top five CAFA3 methods against top five CAFA2 methods show mixed  
81 results. In MFO, the top CAFA3 method, GOLabeler (23) outperformed all CAFA2 methods by a consid-  
82 erable margin, as shown in Figure 2. The rest of the four CAFA3 top methods did not perform as well as  
83 the top two methods of CAFA2, although only to a limited extent, with little change in  $F_{\max}$ . Of the top 12  
84 methods ranked in MFO, seven are from CAFA3, five are from CAFA2 and none are from CAFA1. Despite  
85 the increase in database size, the majority of function prediction methods do not seem to have improved  
86 in predicting protein function in MFO since 2014, except for one method that stood out. In BPO, the top  
87 three methods in CAFA3 outperformed their CAFA2 counterparts, but with very small margins. Out of the  
88 top 12 methods in BPO, eight are from CAFA3, four are from CAFA2 and none are from CAFA1. Finally,  
89 in CCO, although 8 out of top 12 methods over all CAFA challenges come from CAFA3, the top method is  
90 from CAFA2. The differences between the top performing methods are small, as in the case of BPO.

91 The performance of top methods in CAFA2 was significantly better than of those in CAFA1, and it is  
92 interesting to note that this trend has not continued in CAFA3. This could be due to many reasons, such as  
93 the quality of the benchmark sets, the overall quality of the annotation database, the quality of ontologies  
94 or a relatively short period of time between challenges.

95 [Figure 2 about here.]



## 96 2.2 Protein-centric evaluation

97 The *protein-centric* evaluation measures the accuracy of assigning GO terms to a protein. This performance  
98 is shown in Figures 3, 4 and 5.

99 [Figure 3 about here.]

100 [Figure 4 about here.]

101 [Figure 5 about here.]

102 We observe that all top methods outperform the baselines with the patterns of performance consistent  
103 with CAFA1 and CAFA2 findings. Predictions of MFO terms achieved the highest  $F_{\max}$  compared with  
104 predictions in the other two ontologies. BLAST outperforms Naïve in predictions in MFO, but not in BPO  
105 or CCO. This is because sequence similarity based methods such as BLAST tend to perform best when  
106 transferring basic biochemical annotations such as enzymatic activity. Functions in biological process, such  
107 as pathways, may not be as preserved by sequence similarity, hence the poor BLAST performance in BPO.  
108 The reasons behind the difference among the three ontologies include the structure and complexity of the  
109 ontology as well as the state of the annotation database, as discussed previously (26, 31). It is less clear why  
110 the performance in CCO is weak, although it might be hypothesized that such performance is related to the  
111 structure of the ontology itself (31).

112 The top performing method in MFO did not have as high an advantage over others when evaluated  
113 using the  $S_{\min}$  metric. The  $S_{\min}$  metric weights GO terms by conditional information content, since the  
114 prediction of more informative terms are more desirable than less informative, more general, terms. This  
115 could potentially explain the smaller gap between the top predictor and the rest of the pack in  $S_{\min}$ . The  
116 weighted  $F_{\max}$  and normalized  $S_{\min}$  evaluations can be found in Figures S4 and S5.

## 117 2.3 Species-specific categories

118 The benchmarks in each species were evaluated individually as long as there were at least 15 proteins per  
119 species. Here we present results on both eukaryotic and prokaryotic species (Figure 6). We observed that  
120 different methods could perform differently on different species. As shown in Figure 14, bacterial proteins

121 make up a small portion of all benchmark sequences, so their effects on the performances of the methods  
122 are often masked. Species-specific analyses are thus meaningful to researchers studying certain organisms.  
123 Evaluation results on individual species including human (Figure S6), *Arabidopsis thaliana* (Figure S7) and  
124 *Escherichia coli* (Figure S10) can be found in Supplemental Materials (Figures S6-S14).

125 [Figure 6 about here.]

## 126 2.4 Diversity of methods

127 It was suggested in the analysis of CAFA2 that ensemble methods that integrate data from different sources  
128 have the potential of improving prediction accuracy (32). Multiple data sources, including sequence, struc-  
129 ture, expression profile and so on are all potentially predictive of the function of the protein. Therefore,  
130 methods that take advantage of these rich sources as well as existing techniques from other research groups  
131 might see improved performance. Indeed, the one method that stood out from the rest in CAFA3 and per-  
132 formed significantly better than all methods across three challenges, is a machine learning based ensemble  
133 method (23). Therefore, it is important to analyze what information sources and prediction algorithms are  
134 better at predicting function. Moreover, the similarity of the methods might explain the limited improvement  
135 in the rest of the methods in CAFA3.

136 [Figure 7 about here.]

137 The top CAFA2 and CAFA3 methods are very similar in performance, but that could be a result of ag-  
138 gregating predictions of different proteins to one metric. When computing the similarity of each pair of  
139 methods as the reciprocal of the Euclidean distance of prediction scores (Figure 7), we are not interested  
140 whether these predictions are correct according to the benchmarks, but simply whether they are similar to  
141 one another. Top CAFA2 and CAFA3 methods are more similar than with CAFA1 models. It is clear that  
142 some top methods are heavily based on the Naïve and BLAST baseline methods. It is interesting to note  
143 that the top two best methods in BPO are not similar to any other top methods. The same pattern was  
144 observed for CAFA2 methods.

145 [Figure 8 about here.]

146 Participating teams also provided keywords that describe their approach to function prediction with their  
147 submissions. A list of keywords was given to the participants, listed in Page 24 of Supplementary Materials.  
148 Figure 8 shows the frequency of each keyword. In addition, we have weighted the frequency of the keywords  
149 with the prediction accuracy of the specific method. Machine learning and sequence alignment remain  
150 the most-used approach by scientists predicting in all three ontologies. By raw count, machine learning is  
151 more popular than sequence alignment, but once adjusted by performance, they are almost identical. This  
152 indicates that methods that use sequence alignments are more helpful in predicting the correct function than  
153 the popularity of their use suggests.

## 154 **2.5 Evaluation via molecular screening**

155 Databases with proteins annotated by biocuration, such as UniProt knowledge base, have been the primary  
156 source of benchmarks in the CAFA challenges. New to CAFA3, we also evaluated the extent to which methods  
157 participating in CAFA could predict the results of genetic screens in model organisms done specifically for this  
158 project. Predicting GO terms for a protein (protein-centric) and predicting which proteins are associated  
159 with a given function (term-centric) are related but different computational problems: the former is a  
160 multi-label classification problem with a structured output, while the latter is a binary classification task.  
161 Predicting the results of a genome-wide screen for a single or a small number of functions fits the term-centric  
162 formulation. To see how well all participating CAFA methods perform term-centric predictions, we mapped  
163 results from the protein-centric CAFA3 methods onto these terms. In addition we held a separate CAFA  
164 challenge, CAFA- $\pi$  whose purpose was to attract additional submissions from algorithms that specialize in  
165 term-centric tasks.

166 We performed screens for three functions in three species, which we then used to assess protein function  
167 prediction. In the bacterium *Pseudomonas aeruginosa* and the fungus *Candida albicans* we performed  
168 genome-wide screens capable of uncovering genes with two functions, biofilm formation (GO:0042710) and  
169 motility (for *P. aeruginosa* only) (GO:0001539), as described in Methods. In *Drosophila melanogaster* we  
170 performed targeted assays, guided by previous CAFA submissions, of a selected set of genes and assessed  
171 whether or not they affected long-term memory (GO:0007616).

172 We discuss the prediction results for each function below in detail. The performance, as assessed by the

173 genome-wide screens, was generally lower than in the protein-centric evaluations that were curation driven.  
174 We hypothesize that it may simply be more difficult to perform term-centric prediction for broad activities  
175 such as biofilm formation and motility. For *P. aeruginosa*, an existing compendium of gene expression  
176 data was already available (33). We used the Pearson correlation over this collection of data to provide  
177 a complementary baseline to the standard BLAST approach used throughout CAFA. We found that an  
178 expression-based method outperformed the CAFA participants, suggesting that success on certain term-  
179 centric challenges will require the use of different types of data. On the other hand, the performance of the  
180 methods in predicting long-term memory in the *Drosophila* genome was relatively accurate.

### 181 **2.5.1 Biofilm formation**

182 In March 2018, there were 3019 annotations to biofilm formation (GO:0042710) and its descendent terms  
183 across all species, of which 325 used experimental evidence codes. These experimentally annotated proteins  
184 included 131 from the *Candida* Genome Database (34) for *C. albicans* and 29 for *P. aeruginosa*, the two  
185 organisms that we screened.

186 Of the 2746 genes we screened in the *Candida albicans* colony biofilm assay, 245 were required for the  
187 formation of wrinkled colony biofilm formation (Table 1). Of these, only five were already annotated in  
188 UniProt: *MOB*, *EED1* (*DEF1*), and *YAK1*, which encode proteins involved in hyphal growth, an important  
189 trait for biofilm formation (35, 36, 37, 38). Also, *NUP85*, a nuclear pore protein involved in early phase  
190 arrest of biofilm formation (39) and *VPS1*, which contributes to protease secretion, filamentation, and biofilm  
191 formation (40). Of the 2063 proteins that we did not find to be associated with biofilm formation, 29 were  
192 annotated to the term in the GOA database. Some of the proteins in this category highlight the need for  
193 additional information to GO term annotation. For example, *Wor1* and the pheromone receptor are key  
194 for biofilm formation in strains under conditions in which the mating pheromone is produced (41), but not  
195 required in the monocultures of the commonly studied  $a/\alpha$  mating type strain used here.

196 No method in CAFA- $\pi$  or CAFA3 (not shown) exceeded an AUC of 0.60 on this term-centric challenge  
197 (Figure 9) for either species. Performance for the best methods slightly exceeded a BLAST-based baselines.  
198 In the past, we have found that predicting BPO terms, such as biofilm formation, resulted in poorer method  
199 performance than predicting MFO terms. Many CAFA methods use sequence alignment as their primary

			GOA annotations	
			Unannotated	Annotated
<i>C. albicans</i>	Total: 2308			
	CAFA experiments	False	2034	29
		True	240	5
<i>P. aeruginosa</i>	Total: 4056			
	CAFA experiments	False	3491	25
		True	532	9

Table 1: Number of proteins in *Candida albicans* and *Pseudomonas aeruginosa* associated with function Biofilm formation (GO:0042710) in the GOA databases versus experimental results.

200 source of information (Section 2.4). For *Pseudomonas aeruginosa* a pre-built expression compendium was  
201 available from prior work (33). Where the compendium was available, simple gene-expression based baselines  
202 were the best performing approaches. This suggests that successful term-centric prediction of biological  
203 processes may need to rely more heavily on information that is not sequence-based, and, as previously  
204 reported, may require methods that use broad collections of gene expression data (42, 43).

205 [Figure 9 about here.]

## 206 2.5.2 Motility

207 In March 2018 there were 302,121 annotations for proteins with the GO term: cilium or flagellum-dependent  
208 cell motility (GO:0001539) and its descendent terms, which included cell motility in all eukaryotic (GO:0060285),  
209 bacterial (GO:0071973) and archael (GO:0097590) organisms. Of these, 187 had experimental evidence codes  
210 and the most common organism to have annotations was *P. aeruginosa*, on which our screen was performed  
211 (Table S2).

212 As expected, mutants defective in the flagellum or its motor were defective in motility (*fliC* and other  
213 *fli* and *flg* genes). For some of the genes that were expected, but not detected, the annotation was based  
214 on experiments performed in a medium different from what was used in these assays. For example, PhoB  
215 regulates motility but only when phosphate concentration is low (44). Among the genes that were scored  
216 as defective in motility, some are known to have decreased motility due to over production of carbohydrate  
217 matrix material (*bifA*) (45), or the absence of directional swimming due to absence of chemotaxis functions  
218 (e.g., *cheW*, *cheA*) and others likely showed this phenotype because of a medium specific requirement such  
219 as biotin (*bioA*, *bioC*, and *bioD*) (46). Table 2 shows the contingency table for number of proteins that are

Total: 3630	GOA annotations		
	Unannotated	Annotated	
CAFA experiments	False	3195	12
	True	403	21

Table 2: Number of proteins in *Pseudomonas aeruginosa* associated with function Motility (GO:0001539) in the GOA databases versus experimental results.

220 detected by our experiment versus GOA annotations.

221 The results from this evaluation were consistent with what we observed for biofilm formation. Many  
222 of the genes annotated as being involved in biofilm formation were identified in the screen. Others that  
223 were annotated as being involved in biofilm formation did not show up in the screen because the strain  
224 background used here, strain PA14, uses the exopolysaccharide matrix carbohydrate Pel (47) in contrast to  
225 the Psl carbohydrate used by another well characterized strain, strain PAO1 (48, 49). The *psl* genes were  
226 known to be dispensable for biofilm formation in the strain PA14 background and this nuance highlights the  
227 need for more information to be taken into account when making predictions.

228 The CAFA- $\pi$  methods outperformed our BLAST-based baselines but failed to outperform expression-  
229 based baselines. Transferred methods from CAFA3 also did not outperform these baselines. It is important to  
230 note this consistency across terms, reinforcing the finding that term-centric prediction of biological processes  
231 is likely to require non-sequence information to be included.

232 [Figure 10 about here.]

### 233 2.5.3 Long-term memory in *D. melanogaster*

234 Prior to our experiments, there were 1901 annotations made in long-term memory, including 283 experimental  
235 annotations. *Drosophila melanogaster* had the most annotated proteins of long-term memory with 217, while  
236 human has 7, as shown in Table S3.

237 We performed RNAi experiments in *Drosophila melanogaster* to assess whether 29 target genes were  
238 associated with long-term memory (GO:0007616); for details on target selection, see (29). None of the  
239 29 genes had an existing annotation in the GOA database. Because no genome-wide screen results were  
240 available, we did not release this as part of CAFA- $\pi$  and instead relied only on the transfer of methods that

241 predicted “long-term memory” at least once in *D. melanogaster* from CAFA3. Results from this assessment  
242 were more promising than our findings from the genome-wide screens in microbes (Figure 11). Certain  
243 methods performed well, substantially exceeding the baselines.

244 [Figure 11 about here.]

## 245 **2.6 Participation Growth**

246 The CAFA challenge has seen growth in participation, as shown in Figure 12. To cope with the increasingly  
247 large data size, CAFA3 utilized the Synapse (50) online platform for submission. Synapse allowed for easier  
248 access for participants, as well as easier data collection for the organizers. The results were also released to  
249 the individual teams via this online platform. During the submission process, the online platform also allows  
250 for customized format checkers to ensure the quality of the submission.

251 [Figure 12 about here.]

## 252 **3 Methods**

### 253 **3.1 Benchmark collection**

254 In CAFA3, we adopted the same benchmark generation methods as CAFA1 and CAFA2, with a similar time-  
255 line (Figure 13). The crux of a time-delayed challenge is the annotation growth period between time  $t_0$  and  
256  $t_1$ . All target proteins that have gained experimental annotation during this period are taken as benchmarks  
257 in all three ontologies. “No-knowledge” (NK, no prior experimental annotations) and “Limited-knowledge”  
258 (LK, partial prior experimental annotations) benchmarks were also distinguished based on whether the  
259 newly-gained experimental annotation is in an ontology that already have experimental annotations or not.  
260 Evaluation results in Figures 3, 4, and 5 are made using the No-knowledge benchmarks. Evaluation results  
261 on the Limited-knowledge benchmarks are shown in Figure S3 in the Supplemental Materials. For more  
262 information regarding NK and LK designations, please refer to the Supplemental Materials and the CAFA2  
263 paper (26).

264 [Figure 13 about here.]

265 After collecting these benchmarks, we performed two major deletions from the benchmark data. Upon  
266 inspecting the taxonomic distribution of the benchmarks, we noticed a large number of new experimental  
267 annotations from *Candida albicans*. After consulting with UniProt-GOA, we determined these annotations  
268 have already existed in the Candida Genome Database long before 2018, but were only recently migrated to  
269 GOA. Since these annotations were already in the public domain before the CAFA3 submission deadline, we  
270 have deleted any annotation from *Candida albicans* with an assigned date prior to our CAFA3 submission  
271 deadline. Another major change is the deletion of any proteins with only a protein-binding (GO:0005515)  
272 annotation. Protein-binding is a highly generalized function description, does not provide more specific  
273 information about the actual function of a protein, and in many cases may indicate a non-functional, non-  
274 specific binding. If it is the only annotation that a protein has gained, then it is hardly an advance in our  
275 understanding of that protein, therefore we deleted these annotations from our benchmark set. Annotations  
276 with a depth of 3 make up almost half of all annotations in MFO before the removal (Figure S15b). After  
277 the removal, the most frequent annotations became of depth 5 (Figure S15a). In BPO, the most frequent  
278 annotations are of depth 5 or more, indicating a healthy increase of specific GO terms being added to our  
279 annotation database. In CCO, however, most new annotations in our benchmark set are of depth 3, 4 and  
280 5 (Figure S15). This difference could partially explain why the same computational methods perform very  
281 differently in different ontologies, and benchmark sets. We have also calculated total information content  
282 per protein for the benchmark sets shown in Figure S16. Taxonomic distributions of the proteins in our final  
283 benchmark set are shown in Figure 14.

284 [Figure 14 about here.]

285 Additional analyses were performed to assess the characteristics of the benchmark set, including the overall  
286 information content of the terms being annotated.

### 287 **3.2 Protein-centric evaluation**

288 Two main evaluation metrics were used in CAFA3, the  $F_{\max}$  and the  $S_{\min}$ . The  $F_{\max}$  based on the precision-  
289 recall curve, while the  $S_{\min}$  is based the RU-MI curve. Mathematical definitions of these metrics are shown  
290 in pages 22 and 23 of Supplemental Materials. The RU-MI curve (51) takes into account the information



291 content of each GO term in addition to counting the number of true positives, false positives, etc. See  
292 Supplemental Materials for their mathematical definitions. The information theory based evaluation metrics  
293 counters the high-throughput low-information annotations such as protein binding, but down-weighting these  
294 terms according to their information content, as the ability to predict such non-specific functions are not as  
295 desirable and useful and the ability to predict more specific functions.

296 The two assessment modes from CAFA2 were also used in CAFA3. In the partial mode, predictions were  
297 evaluated only on those benchmarks for which a model made at least one prediction. The full evaluation  
298 mode evaluates all benchmark proteins and methods were penalized for not making predictions. Evaluation  
299 results in Figures 3, 4, and 5 are made using the full evaluation mode. Evaluation results using the partial  
300 mode are shown in Figure S2 in the Supplemental Materials.

301 Two baseline models were also computed for these evaluations. The Naïve method assigns the term  
302 frequency as the prediction score for any protein, regardless of any protein-specific properties. BLAST  
303 was based on results using the Basic Local Alignment Search Tool (BLAST) software against the training  
304 database (52). A term will be predicted as the highest local alignment sequence identity among all BLAST  
305 hits annotated from the training database. Both of these methods were trained on the experimentally  
306 annotated proteins and their sequences in Swiss-Prot (53) at time  $t_0$ .

### 307 **3.3 Microbe screens**

308 To assess matrix production, we used mutants from the PA14 NR collection (54). Mutants were transferred  
309 from the  $-80^{\circ}\text{C}$  freezer stock using a sterile 48-pin multiprong device into 200 $\mu\text{l}$  LB in a 96-well plate. The  
310 cultures were incubated overnight at  $37^{\circ}\text{C}$ , and their OD600 was measured to assess growth. Mutants were  
311 then transferred to tryptone agar with 15g of tryptone and 15g of agar in 1L amended with Congo red  
312 (Aldrich, 860956) and Coomassie brilliant blue (J.T. Baker Chemical Co., F789-3). Plates were incubated  
313 at  $37^{\circ}\text{C}$  overnight followed by four day incubation at room temperature on allow the wrinkly phenotype to  
314 develop. Colonies were imaged and scored on Day 5. To assess motility, mutants were revived from freezer  
315 stocks as described above. After overnight growth, a sterile 48-pin multiprong transfer device with a pin  
316 diameter of 1.58 mm was used to stamp the mutants from the overnight plates into the center of swim  
317 agar made with M63 medium with 0.2% glucose and casamino acids and 0.3% agar). Care was taken to

318 avoid touching the bottom of the plate. Swim plates were incubated at room temperature (19-22°C) for  
319 approximately 17 hours before imaging and scoring. Experimental procedures in *P. aeruginosa* to determine  
320 proteins that are associated with the two functions in CAFA- $\pi$  are shown in Figure 15.

321 [Figure 15 about here.]

322 Biofilm formation in *Candida albicans* was assessed in single gene mutants from the Noble (55) and  
323 GRACE (56) collections. In the Noble Collection, mutants of *C. albicans* have had both copies of the  
324 candidate gene deleted. Most of the mutants were created in biological duplicate. From this collection,  
325 1274 strains corresponding to 653 unique genes were screened. The GRACE collection provided mutants  
326 with one copy of each gene deleted and the other copy placed under the control of a doxycycline-repressible  
327 promoter. To assay these strains, we used medium supplemented with 100 $\mu$ g/ml doxycycline strains, when  
328 rendered them functional null mutants. We screened 2348 mutants from the GRACE collection, 255 of  
329 which overlapped with mutants in the Noble collection, for 2746 total unique mutants screened in total. To  
330 assess defects in biofilm formation or biofilm-related traits, we performed two assays: (1) colony morphology  
331 on agar medium and (2) biofilm formation on a plastic surface (Figure 16). For both of these assays we  
332 used Spider medium, which was designed to induce hyphal growth in *C. albicans* (57), and which promotes  
333 biofilm formation (39). Strains were first replicated from frozen 96 well plates to YPD agar plates. Strains  
334 were then replicated from YPD agar to YPD broth, and grown overnight at 30°C. From YPD broth, strains  
335 were introduced onto Spider agar plates and into 96 well plates of Spider broth. When strains from the  
336 GRACE collection were assayed, 100 $\mu$ g/ml doxycycline was included in the agar and broth, and aluminium  
337 foil was used to protect the media from light. Spider agar plates inoculated with *C. albicans* mutants  
338 were incubated at 37°C for two days before colony morphologies were scored. Strains in Spider Broth were  
339 shaken at 225 rpm at 37°C for three days, and then assayed for biofilm formation at the air-liquid interface  
340 as follows. First, broth was removed by slowly tilting plates and pulling liquid away by running a gloved  
341 hand over the surface. Biofilms were stained by adding 100 $\mu$ l of 0.1 percent crystal violet dye in water to  
342 each well of the plate. After 15 minutes, plates were gently washed in three baths of water to remove dye  
343 without disturbing biofilms. To score biofilm formation for agar plates, colonies were scored by eye as either  
344 smooth, intermediate, or wrinkled. A wild-type colony would score wrinkled, and mutants with intermediate

345 or smooth appearance were considered defective in colony biofilm formation. For biofilm formation on a  
 346 plastic surface, the presence of a ring of cell material in the well indicated normal biofilm formation, while  
 347 low or no ring formation mutants were considered defective. Genes whose mutations resulted defects in both  
 348 or either assay were considered True for biofilm function. A complete list of the mutants identified in the  
 349 screens is available in Table S1.

350 [Figure 16 about here.]

351 A protein is considered True in the biofilm function, if its mutant phenotype is smooth or intermediate under  
 352 Doxycycline.

### 353 3.4 Term-centric evaluation

354 The evaluations of the CAFA- $\pi$  methods were based on the experimental results in Section 3.3. We adopted  
 355 both  $F_{\max}$  based on precision-recall curves and area under ROC curves. There are a total of six baseline  
 356 methods, as described in Table 3.

	Model Number	Training data	Score assignment
expression	1	Gene expression compendium for <i>P. aeruginosa PAO1</i>	Highest correlation score out of all pairwise correlations
	2		Top 10 average correlation score
blast	1	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot	Highest sequence identity out of all pairwise BLASTp hits
	2	All experimental annotation in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	
blastcomp	1	All experimental <b>and</b> computational annotations in UniProt-GOA. Sequences from Swiss-Prot	
	2	All experimental <b>and</b> computational annotations in UniProt-GOA. Sequences from Swiss-Prot and TrEMBL	

Table 3: Baseline methods in term-centric evaluation of protein function prediction.

## 357 4 Discussion

358 Since 2010, the CAFA community has been a home to a growing group of scientists across the globe sharing  
359 the goal of improving computational function prediction. CAFA has been advancing this goal in three ways.  
360 First, through independent evaluation of computational methods against the set of benchmark proteins, thus  
361 providing a direct comparison of the methods' reliability and performance at a given time point. Second, the  
362 challenge assesses the quality of the current state of the annotations, whether they are made computationally  
363 or not, and is set up to reliably track it over time. Finally, as described in this work, CAFA has started  
364 to drive the creation of new experimental annotations by facilitating synergies between different groups of  
365 researchers interested in function of biological macromolecules. These annotations not only represent new  
366 biological discoveries, but simultaneously serve to provide benchmark data for rigorous method evaluation.

367 CAFA3 and CAFA- $\pi$  feature the latest advances in the CAFA series to create advanced and accurate  
368 methods for protein function prediction. We use the repeated nature of the CAFA project to identify certain  
369 trends via historical assessments. The analysis revealed that the performance of CAFA methods improved  
370 dramatically between CAFA1 and CAFA2. However, the protein-centric results for CAFA3 are mixed when  
371 compared to historical methods. Though the best performing CAFA3 method outperformed the top CAFA2  
372 methods (Figure 1), this was not consistently true for other rankings. Among all three CAFA challenges,  
373 CAFA2 and CAFA3 methods inhabit the top 12 places in MFO and BPO. Between CAFA2 and CAFA3  
374 the performance increase is more subtle. Based on the annotations of methods (Supplementary Materials),  
375 many of the top-ranking methods are improved versions of methods that have been evaluated in CAFA2.  
376 Interestingly, the top performing CAFA3 method, which consistently outperformed methods from all past  
377 CAFAs in the major categories, was a novel contribution (Zhu lab).

378 For this iteration of CAFA we performed genome-wide screens of phenotypes in *P. aeruginosa* and  
379 *C. albicans* as well as a targeted screen in *D. melanogaster*. This not only allowed us to assess the accuracy  
380 with which methods predict genes associated with select biological processes, but also to use CAFA as  
381 an additional driver for new biological discovery. In short, our experimental work identified more than a  
382 thousand of new functional annotations in three highly divergent species. Though all screens have certain  
383 limitations, the genome-wide screens also bypass questions of biases in curation. This evaluation provides

384 key insights: CAFA3 methods did not generalize well to selected terms. Because of that, we ran a second  
385 effort, CAFA- $\pi$ , in which participants focused solely on predicting the results of these targeted assays. This  
386 targeted effort led to improved performance, suggesting that when the goal is to identify genes associated  
387 with a specific phenotype, tuning methods may be required.

388 For CAFA evaluations, we have included both Naïve and sequence-based (BLAST) baseline methods.  
389 For the evaluation of *P. aeruginosa* screen results, we were also able to include a gene expression baseline  
390 from a previously published compendium (33). Intriguingly, the expression-based predictions outperformed  
391 existing methods for this task. In future CAFA efforts, we will include this type of baseline expression-based  
392 method across evaluations to continue to assess the extent to which this data modality informs gene function  
393 prediction. The results from the CAFA3 effort suggest that gene expression may be particularly important  
394 for successfully predicting term-centric biological process annotations.

395 The primary takeaways from CAFA3 are: (1) Genome-wide screens complement annotation-based efforts  
396 to provide a richer picture of protein function prediction; (2) The best performing method was a new method,  
397 instead of a light retooling of an existing approach; (3) Gene expression, and more broadly, systems data  
398 may provide key information to unlocking biological process predictions, and (4) Performance of the best  
399 methods has continued to improve. The results of the screens released as part of CAFA3 can lead to a  
400 re-examination of approaches which we hope will lead to improved performance in CAFA4.

## 401 5 Acknowledgements

402 Will be provided with the final manuscript

## 403 6 Data and Software

404 Data are available on figshare: [https://figshare.com/articles/Supplementary\\_data/8135393](https://figshare.com/articles/Supplementary_data/8135393)

405 The assessment software used in this paper is available under GNU-GPLv3 license at: [https://github.](https://github.com/ashleyzhou972/CAFA_assessment_tool)

406 [com/ashleyzhou972/CAFA\\_assessment\\_tool](https://github.com/ashleyzhou972/CAFA_assessment_tool)

## 407 **7 Funding**

408 The work of IF was funded, in part, by National Science Foundation award DBI-1458359. The work of CSG  
409 and AJL was funded, in part, by National Science Foundation award DBI-1458390 and GBMF 4552 from the  
410 Gordon and Betty Moore Foundation. The work of DAH and KAL was funded, in part, by National Science  
411 Foundation award DBI-1458390, National Institutes of Health NIGMS P20 GM113132, and the Cystic  
412 Fibrosis Foundation CFRDP STANTO19R0. The work of AP, HY, AR and MT was funded by BBSRC grants  
413 BB/K004131/1, BB/F00964X/1 and BB/M025047/1, Consejo Nacional de Ciencia y Tecnología Paraguay  
414 (CONACyT) grants 14-INV-088 and PINV15-315, and NSF Advances in Bio Informatics grant 1660648.  
415 DK acknowledges supports from the National Institutes of Health (R01GM123055) and the National Science  
416 Foundation (DMS1614777, CMMI1825941). PB acknowledges support from National Institutes of Health  
417 (R01GM60595). GB and BZK acknowledge support from the National Science Foundation (NSF 1458390)  
418 and NIH DP1MH110234. FS was funded by the ERC StG 757700 "HYPER-INSIGHT" and by the Spanish  
419 Ministry of Science, Innovation and Universities grant BFU2017-89833-P. FS further acknowledges funding  
420 from the Severo Ochoa award to the IRB Barcelona. The work of SK was funded by ATT Tieto käyttöön grant  
421 and Academy of Finland. TB and SM were funded by NIH awards UL1 TR002319 and U24 TR002306. The  
422 work of CZ and ZW was funded by National Institutes of Health R15GM120650 to ZW and start-up funding  
423 from the University of Miami to ZW. PR acknowledges NSF grant DBI-1458477. PT acknowledges support  
424 from Helsinki Institute for Life Sciences. The work of FZ and WT was funded by the National Natural Science  
425 Foundation of China (31671367, 31471245, 91631301) and the National Key Research and Development  
426 Program of China (2016YFC1000505, 2017YFC0908402]. CS acknowledges support by the Italian Ministry  
427 of Education, University and Research (MIUR) PRIN 2017 project 2017483NH8. SZ is supported by National  
428 Natural Science Foundation of China (No. 61872094 and No. 61572139) and Shanghai Municipal Science  
429 and Technology Major Project (No. 2017SHZDZX01). PLF and RLH were supported by the National  
430 Institutes of Health NIH R35-GM128637 and R00-GM097033. DTJ, CW, DC and RF were supported by  
431 the UK Biotechnology and Biological Sciences Research Council (BB/L020505/1 and BB/L002817/1) and  
432 Elsevier. The work of YZ and CZ was funded in part by the National Institutes of Health award GM083107,  
433 GM116960, AI134678, the National Science Foundation award DBI1564756, and the Extreme Science and

434 Engineering Discovery Environment (XSEDE) award MCB160101 and MCB160124. The work of BG, VP,  
435 RD, NS and NV was funded by the Ministry of Education, Science and Technological Development of the  
436 Republic of Serbia, Project No. 173001. The work of YWL, WHL, JMC was funded by the Taiwan Ministry  
437 of Science and Technology (106-2221-E-004-011-MY2). YWL, WHL, JMC further acknowledge support from  
438 “the Human Project from Mind, Brain and Learning” of the NCCU Higher Education Sprout Project by  
439 the Taiwan Ministry of Education and the National Center for High-performance Computing for computer  
440 time and facilities. The work of IK and AB was funded by Montana State University and NSF Advances  
441 in Biological Informatics program through grant number 0965768. BR, TG and JR are supported by the  
442 Bavarian Ministry for Education through funding to the TUM. The work of RB, VG, MB, and DCEK was  
443 supported by the Simons Foundation and NIH NINDS grant number 1R21NS103831-01.

## 444 References

- 445 [1] S. Goodwin, J. D. McPherson, and W. R. McCombie. Coming of age: ten years of next-generation  
446 sequencing technologies. *Nat Rev Genet*, 17(6):333–351, 2016.
- 447 [2] R. Aebersold and M. Mann. Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207, 2003.
- 448 [3] A. M. Schnoes, D. C. Ream, A. W. Thorman, P. C. Babbitt, and I. Friedberg. Biases in the experimental  
449 annotations of protein function and their effect on our understanding of protein function space. *PLoS*  
450 *Comput Biol*, 9(5):e1003063, 2013.
- 451 [4] B. Rost, J. Liu, R. Nair, K. O. Wrzeszczynski, and Y. Ofran. Automatic prediction of protein function.  
452 *Cell Mol Life Sci*, 60(12):2637–2650, 2003.
- 453 [5] I. Friedberg. Automated protein function prediction—the genomic challenge. *Brief Bioinform*, 7(3):225–  
454 242, 2006.
- 455 [6] R. Sharan, I. Ulitsky, and R. Shamir. Network-based prediction of protein function. *Mol Syst Biol*,  
456 3:88, 2007.
- 457 [7] R. Rentsch and C. A. Orengo. Protein function prediction—the power of multiplicity. *Trends Biotechnol*,  
458 27(4):210–219, 2009.
- 459 [8] A. Shehu, D. Barbara, and K. Molloy. *A survey of computational methods for protein function predic-*  
460 *tions*, pages 225–298. Springer, 2016.
- 461 [9] D. Cozzetto and D. T. Jones. Computational methods for annotation transfers from sequence. *Methods*  
462 *Mol Biol*, 1446:55–67, 2017.
- 463 [10] M. Pellegrini, E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates. Assigning protein  
464 functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci USA*,  
465 96(8):4285–4288, 1999.
- 466 [11] L. J. Jensen, R. Gupta, N. Blom, D. Devos, J. Tamames, C. Kesmir, H. Nielsen, H. H. Staerfeldt,  
467 K. Rapacki, C. Workman, C. A. Andersen, S. Knudsen, A. Krogh, A. Valencia, and S. Brunak. Prediction



- 468 of human protein function from post-translational modifications and localization features. *J Mol Biol*,  
469 319(5):1257–1265, 2002.
- 470 [12] M. Deng, K. Zhang, S. Mehta, T. Chen, and F. Sun. Prediction of protein function using protein-protein  
471 interaction data. *J Comput Biol*, 10(6):947–960, 2003.
- 472 [13] F. Pazos and M. J. Sternberg. Automated prediction of protein function and detection of functional  
473 sites from structure. *Proc Natl Acad Sci USA*, 101(41):14754–14759, 2004.
- 474 [14] E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein  
475 function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–310, 2005.
- 476 [15] B. E. Engelhardt, M. I. Jordan, K. E. Muratore, and S. E. Brenner. Protein molecular function prediction  
477 by Bayesian phylogenomics. *PLoS Comput Biol*, 1(5):e45, 2005.
- 478 [16] F. Enault, K. Suhre, and J. M. Claverie. Phydbac “Gene Function Predictor”: a gene annotation tool  
479 based on genomic context analysis. *BMC Bioinformatics*, 6:247, 2005.
- 480 [17] T. Hawkins, S. Luban, and D. Kihara. Enhanced automated function prediction using distantly related  
481 sequences and contextual association by PFP. *Protein Sci*, 15(6):1550–1556, 2006.
- 482 [18] M. N. Wass and M. J. Sternberg. Confunc–functional annotation in the twilight zone. *Bioinformatics*,  
483 24(6):798–806, 2008.
- 484 [19] S. Mostafavi, D. Ray, D. Warde-Farley, C. Grouios, and Q. Morris. GeneMANIA: a real-time multiple  
485 association network integration algorithm for predicting gene function. *Genome Biol*, 9(Suppl 1):S4,  
486 2008.
- 487 [20] A. Sokolov and A. Ben-Hur. Hierarchical classification of gene ontology terms using the GOstruct  
488 method. *J Bioinform Comput Biol*, 8(2):357–376, 2010.
- 489 [21] W. T. Clark and P. Radivojac. Analysis of protein function and its prediction from amino acid sequence.  
490 *Proteins*, 79(7):2086–2096, 2011.

- 491 [22] D. Piovesan, M. Giollo, E. Leonardi, C. Ferrari, and S. C. E. Tosatto. INGA: protein function predic-  
492 tion combining interaction networks, domain assignments and sequence similarity. *Nucleic Acids Res*,  
493 43(W1):W134–W140, 2015.
- 494 [23] R. You, Z. Zhang, Y. Xiong, F. Sun, H. Mamitsuka, and S. Zhu. GOLabeler: improving sequence-based  
495 large-scale protein function prediction by learning to rank. *Bioinformatics*, 34(14):2465–2473, 2018.
- 496 [24] R. Fa, D. Cozzetto, C. Wan, and D. T. Jones. Predicting human protein function with multi-task deep  
497 neural networks. *PLoS One*, 13(6):e0198216, 2018.
- 498 [25] P. Radivojac, W. T. Clark, T. R. Oron, A. M. Schnoes, T. Wittkop, A. Sokolov, K. Graim, C. Funk,  
499 K. Verspoor, A. Ben-Hur, G. Pandey, J. M. Yunes, A. S. Talwalkar, S. Repo, M. L. Souza, D. Piovesan,  
500 R. Casadio, Z. Wang, J. Cheng, H. Fang, J. Gough, P. Koskinen, P. Toronen, J. Nokso-Koivisto,  
501 L. Holm, D. Cozzetto, D. W. Buchan, K. Bryson, D. T. Jones, B. Limaye, H. Inamdar, A. Datta,  
502 S. K. Manjari, R. Joshi, M. Chitale, D. Kihara, A. M. Lisewski, S. Erdin, E. Venner, O. Lichtarge,  
503 R. Rentzsch, H. Yang, A. E. Romero, P. Bhat, A. Paccanaro, T. Hamp, R. Kassner, S. Seemayer,  
504 E. Vicedo, C. Schaefer, D. Achten, F. Auer, A. Boehm, T. Braun, M. Hecht, M. Heron, P. Honigschmid,  
505 T. A. Hopf, S. Kaufmann, M. Kiening, D. Krompass, C. Landerer, Y. Mahlich, M. Roos, J. Bjorne,  
506 T. Salakoski, A. Wong, H. Shatkay, F. Gatzmann, I. Sommer, M. N. Wass, M. J. Sternberg, N. Skunca,  
507 F. Supek, M. Bosnjak, P. Panov, S. Dzeroski, T. Smuc, Y. A. Kourmpetis, A. D. van Dijk, C. J.  
508 ter Braak, Y. Zhou, Q. Gong, X. Dong, W. Tian, M. Falda, P. Fontana, E. Lavezzo, B. Di Camillo,  
509 S. Toppo, L. Lan, N. Djuric, Y. Guo, S. Vucetic, A. Bairoch, M. Linial, P. C. Babbitt, S. E. Brenner,  
510 C. Orengo, B. Rost, S. D. Mooney, and I. Friedberg. A large-scale evaluation of computational protein  
511 function prediction. *Nat Methods*, 10(3):221–227, 2013.
- 512 [26] Y. Jiang, T. R. Oron, W. T. Clark, A. R. Bankapur, D. D’Andrea, R. Lepore, C. S. Funk, I. Kahanda,  
513 K. M. Verspoor, A. Ben-Hur, C. E. Koo da, D. Penfold-Brown, D. Shasha, N. Youngs, R. Bonneau,  
514 A. Lin, S. M. Sahraeian, P. L. Martelli, G. Profiti, R. Casadio, R. Cao, Z. Zhong, J. Cheng, A. Altenhoff,  
515 N. Skunca, C. Dessimoz, T. Dogan, K. Hakala, S. Kaewphan, F. Mehryary, T. Salakoski, F. Ginter,  
516 H. Fang, B. Smithers, M. Oates, J. Gough, P. Toronen, P. Koskinen, L. Holm, C. T. Chen, W. L.  
517 Hsu, K. Bryson, D. Cozzetto, F. Minneci, D. T. Jones, S. Chapman, D. Bkc, I. K. Khan, D. Kihara,

- 518 D. Ofer, N. Rappoport, A. Stern, E. Cibrian-Uhalte, P. Denny, R. E. Foulger, R. Hieta, D. Legge,  
519 R. C. Lovering, M. Magrane, A. N. Melidoni, P. Mutowo-Meullenet, K. Pichler, A. Shypitsyna, B. Li,  
520 P. Zakeri, S. ElShal, L. C. Tranchevent, S. Das, N. L. Dawson, D. Lee, J. G. Lees, I. Sillitoe, P. Bhat,  
521 T. Nepusz, A. E. Romero, R. Sasidharan, H. Yang, A. Paccanaro, J. Gillis, A. E. Sedeno-Cortes,  
522 P. Pavlidis, S. Feng, J. M. Cejuela, T. Goldberg, T. Hamp, L. Richter, A. Salamov, T. Gabaldon,  
523 M. Marcet-Houben, F. Supek, Q. Gong, W. Ning, Y. Zhou, W. Tian, M. Falda, P. Fontana, E. Lavezzo,  
524 S. Toppo, C. Ferrari, M. Giollo, D. Piovesan, S. C. Tosatto, A. Del Pozo, J. M. Fernandez, P. Maietta,  
525 A. Valencia, M. L. Tress, A. Benso, S. Di Carlo, G. Politano, A. Savino, H. U. Rehman, M. Re,  
526 M. Mesiti, G. Valentini, J. W. Bargsten, A. D. van Dijk, B. Gemovic, S. Glisic, V. Perovic, V. Veljkovic,  
527 N. Veljkovic, E. S. D. C. Almeida, R. Z. Vencio, M. Sharan, J. Vogel, L. Kansakar, S. Zhang, S. Vucetic,  
528 Z. Wang, M. J. Sternberg, M. N. Wass, R. P. Huntley, M. J. Martin, C. O'Donovan, P. N. Robinson,  
529 Y. Moreau, A. Tramontano, P. C. Babbitt, S. E. Brenner, M. Linial, C. A. Orengo, B. Rost, C. S.  
530 Greene, S. D. Mooney, I. Friedberg, and P. Radivojac. An expanded evaluation of protein function  
531 prediction methods shows an improvement in accuracy. *Genome Biol*, 17(1):184, 2016.
- 532 [27] I. Friedberg and P. Radivojac. Community-wide evaluation of computational function prediction. *Meth-*  
533 *ods Mol Biol*, 1446:133–146, 2017.
- 534 [28] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski,  
535 S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese,  
536 J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification  
537 of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25–29, 2000.
- 538 [29] B. Z. Kacsóh, S. Barton, Y. Jiang, N. Zhou, S. D. Mooney, I. Friedberg, P. Radivojac, C. S. Greene,  
539 and G. Bosco. New *Drosophila* long-term memory genes revealed by assessing computational function  
540 prediction methods. *G3*, 9(1):251–267, 2019.
- 541 [30] R. P. Huntley, T. Sawford, P. Mutowo-Meullenet, A. Shypitsyna, C. Bonilla, M. J. Martin, and  
542 C. O'Donovan. The GOA database: gene ontology annotation updates for 2015. *Nucleic Acids Res*,  
543 43(Database issue):D1057–1063, 2015.

- 544 [31] Y. Peng, Y. Jiang, and P. Radivojac. Enumerating consistent sub-graphs of directed acyclic graphs: an  
545 insight into biomedical ontologies. *Bioinformatics*, 34(13):i313–i322, 2018.
- 546 [32] L. Wang, J. Law, S. D. Kale, T. M. Murali, and G. Pandey. Large-scale protein function prediction  
547 using heterogeneous ensembles. *F1000Res*, 7, 2018.
- 548 [33] J. Tan, G. Doing, K. A. Lewis, C. E. Price, K. M. Chen, K. C. Cady, B. Perchuk, M. T. Laub, D. A.  
549 Hogan, and C. S. Greene. Unsupervised extraction of stable expression signatures from public compendia  
550 with an ensemble of neural networks. *Cell Syst*, 5(1):63–71, 2017.
- 551 [34] M. S. Skrzypek, J. Binkley, G. Binkley, S. R. Miyasato, M. Simison, and G. Sherlock. The Candida  
552 Genome Database (CGD): incorporation of Assembly 22, systematic identifiers and visualization of high  
553 throughput sequencing data. *Nucleic Acids Res*, 45(Database issue):D592–D596, 2017.
- 554 [35] S. Goyard, P. Knechtle, M. Chauvel, A. Mallet, M. C. Prevost, C. Proux, J. Y. Coppee, P. Schwarz,  
555 F. Dromer, H. Park, S. G. Filler, G. Janbon, and C. d’Enfert. The Yak1 kinase is involved in the  
556 initiation and maintenance of hyphal growth in *Candida albicans*. *Mol Biol Cell*, 19(5):2251–2266,  
557 2008.
- 558 [36] P. Gutierrez-Escribano, A. Gonzalez-Novo, M. B. Suarez, C. R. Li, Y. Wang, C. R. de Aldana, and  
559 J. Correa-Bordes. Cdk-dependent phosphorylation of Mob2 is essential for hyphal development in  
560 *Candida albicans*. *Mol Biol Cell*, 22(14):2458–2469, 2011.
- 561 [37] T. Lassak, E. Schneider, M. Bussmann, D. Kurtz, J. R. Manak, T. Srikantha, D. R. Soll, and J. F.  
562 Ernst. Target specificity of the *Candida albicans* *efg1* regulator. *Mol Microbiol*, 82(3):602–618, 2011.
- 563 [38] R. Martin, G. P. Moran, I. D. Jacobsen, A. Heyken, J. Domey, D. J. Sullivan, O. Kurzai, and B. Hube.  
564 The *Candida albicans*-specific gene *EED1* encodes a key regulator of hyphal extension. *PLoS One*,  
565 6(4):e18394, 2011.
- 566 [39] M. L. Richard, C. J. Nobile, V. M. Bruno, and A. P. Mitchell. *Candida albicans* biofilm-defective  
567 mutants. *Eukaryot Cell*, 4(8):1493–1502, 2005.

- 568 [40] S. M. Bernardo, Z. Khalique, J. Kot, J. K. Jones, and S. A. Lee. *Candida albicans* VPS1 contributes  
569 to protease secretion, filamentation, and biofilm formation. *Fungal Genet Biol*, 45(6):861–877, 2008.
- 570 [41] S. Yi, N. Sahni, K. J. Daniels, K. L. Lu, G. Huang, T. Srikantha, and D. R. Soll. Self-induction of a/a  
571 or alpha/alpha biofilms in *Candida albicans* is a pheromone-based paracrine system requiring switching.  
572 *Eukaryot Cell*, 10(6):753–760, 2011.
- 573 [42] David C. Hess, Chad L. Myers, Curtis Huttenhower, Matthew A. Hibbs, Alicia P. Hayes, Jadine Paw,  
574 John J. Clore, Rosa M. Mendoza, Bryan San Luis, Corey Nislow, Guri Giaever, Michael Costanzo,  
575 Olga G. Troyanskaya, and Amy A. Caudy. Computationally driven, quantitative experiments discover  
576 genes required for mitochondrial biogenesis. *PLoS Genetics*, 5(3):1–16, 03 2009.
- 577 [43] Matthew A. Hibbs, Chad L. Myers, Curtis Huttenhower, David C. Hess, Kai Li, Amy A. Caudy, and  
578 Olga G. Troyanskaya. Directing experimental biology: A case study in mitochondrial biogenesis. *PLoS*  
579 *Computational Biology*, 5(3):1–12, 03 2009.
- 580 [44] I. Blus-Kadosh, A. Zilka, G. Yerushalmi, and E. Banin. The effect of *pstS* and *phoB* on quorum sensing  
581 and swarming motility in *Pseudomonas aeruginosa*. *PLoS One*, 8(9):e74444, 2013.
- 582 [45] S. L. Kuchma, K. M. Brothers, J. H. Merritt, N. T. Liberati, F. M. Ausubel, and G. A. O’Toole. BifA,  
583 a cyclic-Di-GMP phosphodiesterase, inversely regulates biofilm formation and swarming motility by  
584 *Pseudomonas aeruginosa* PA14. *J Bacteriol*, 189(22):8165–8178, 2007.
- 585 [46] G. L. Winsor, E. J. Griffiths, R. Lo, B. K. Dhillon, J. A. Shay, and F. S. Brinkman. Enhanced annotations  
586 and features for comparing thousands of *Pseudomonas* genomes in the *Pseudomonas* genome database.  
587 *Nucleic Acids Res*, 44(D1):D646–653, 2016.
- 588 [47] L. Friedman and R. Kolter. Genes involved in matrix formation in *Pseudomonas aeruginosa* PA14  
589 biofilms. *Mol Microbiol*, 51(3):675–690, 2004.
- 590 [48] L. Friedman and R. Kolter. Two genetic loci produce distinct carbohydrate-rich structural components  
591 of the *Pseudomonas aeruginosa* biofilm matrix. *J Bacteriol*, 186(14):4457–4465, 2004.

- 592 [49] K. D. Jackson, M. Starkey, S. Kremer, M. R. Parsek, and D. J. Wozniak. Identification of *psl*, a locus  
593 encoding a potential exopolysaccharide that is essential for *Pseudomonas aeruginosa* PAO1 biofilm  
594 formation. *J Bacteriol*, 186(14):4466–4475, 2004.
- 595 [50] Synapse. <https://www.synapse.org/>.
- 596 [51] W. T. Clark and P. Radivojac. Information-theoretic evaluation of predicted ontological annotations.  
597 *Bioinformatics*, 29(13):i53–i61, 2013.
- 598 [52] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped  
599 BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*,  
600 25(17):3389–3402, 1997.
- 601 [53] The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res*,  
602 45(D1):D158–D169, 2017.
- 603 [54] N. T. Liberati, J. M. Urbach, S. Miyata, D. G. Lee, E. Drenkard, G. Wu, J. Villanueva, T. Wei, and  
604 F. M. Ausubel. An ordered, nonredundant library of *Pseudomonas aeruginosa* strain PA14 transposon  
605 insertion mutants. *Proc Natl Acad Sci U S A*, 103(8):2833–2838, 2006.
- 606 [55] S. M. Noble, S. French, L. A. Kohn, V. Chen, and A. D. Johnson. Systematic screens of a *Candida*  
607 *albicans* homozygous deletion library decouple morphogenetic switching and pathogenicity. *Nat Genet*,  
608 42(7):590–598, 2010.
- 609 [56] T. Roemer, B. Jiang, J. Davison, T. Ketela, K. Veillette, A. Breton, F. Tandia, A. Linteau, S. Sillaots,  
610 C. Marta, N. Martel, S. Veronneau, S. Lemieux, S. Kauffman, J. Becker, R. Storms, C. Boone, and  
611 H. Bussey. Large-scale essential gene identification in *Candida albicans* and applications to antifungal  
612 drug discovery. *Mol Microbiol*, 50(1):167–181, 2003.
- 613 [57] H. Liu, J. Kohler, and G. R. Fink. Suppression of hyphal formation in *Candida albicans* by mutation  
614 of a STE12 homolog. *Science*, 266(5191):1723–1726, 1994.

## 615 List of Figures

616	1	A comparison in $F_{\max}$ between the top-five CAFA2 models against the top-five CAFA3 models. Colored boxes encode the results such that (1) the colors indicate margins of a CAFA3 method over a CAFA2 method in $F_{\max}$ and (2) the numbers in the box indicate the percentage of wins. A: CAFA2 top-five models (rows, from top to bottom) against CAFA3 top-five models (columns, from left to right). B: Comparison of performance ( $F_{\max}$ ) of Naïve baselines trained respectively on SwissProt2014 and SwissProt2017. C: Comparison of performance ( $F_{\max}$ ) of BLAST baselines trained on SwissProt2014 and SwissProt2017. Statistical significance was assessed using 10,000 bootstrap samples of benchmark proteins. . . . .	33
624	2	Performance evaluation based on $F_{\max}$ for top CAFA1, CAFA2 and CAFA3 methods. The top 12 methods are shown in this barplot ranked in descending order from left to right. The baseline methods are appended to the right, they were trained on training data from 2017, 2014 and 2011 respectively. Coverage of the methods were shown as text inside the bars. Coverage is defined as percentage of proteins in the benchmark that are predicted by the methods. Color scheme: CAFA2: ivory; CAFA3: green; Naïve: red; BLAST: blue. Note that in MFO and BPO, CAFA1 methods were ranked but not displayed. CAFA1 challenge did not collect predictions for CCO. . . . .	34
632	3	Performance evaluation based on $F_{\max}$ for the top-performing methods in three ontologies. The 95% confidence interval was estimated using 10,000× bootstrap on the benchmark set. Coverage of the methods were shown as text inside the bars. Coverage is defined as percentage of proteins in the benchmark that are predicted by the methods. . . . .	35
636	4	Precision Recall curves for the top-performing methods . . . . .	36
637	5	Evaluation based on the $S_{\min}$ for the top-performing methods . . . . .	37
638	6	Evaluation based on the $F_{\max}$ for the top-performing methods in eukaryotic and prokaryotic species . . . . .	38
640	7	Similarity networks of top 10 methods from CAFA1, CAFA2 and CAFA3. The team names are displayed together with which CAFA challenge they come from in parenthesis. Similarity is calculated as the reciprocal of the Euclidean distance of the prediction scores from each pair of methods. A 0.07 cutoff was applied to the Euclidean distances, i.e. an edge exists if the Euclidean distance is lower than the cutoff. Edge width is directly proportional to similarity, except at the three edges between the three Naïve methods, where the similarity is much larger than the rest. Vertex size is directly proportional to number of edges, or degree of a vertex. Singletons, or vertices without any edges are framed with black circles. The nodes are ranked counter-clockwise, starting after 'BLAST1', by $F_{\max}$ performance in the intersection set of benchmarks in Section 2.1. Color scheme: CAFA1: orange; CAFA2: ivory; CAFA3: green; Naïve: red; BLAST: blue. . . . .	39
651	8	Keyword analysis of all CAFA3 participating methods. Both relative frequency of the keywords and weighted frequency are provided. The weighted frequencies accounts for the performance of the the particular model using the given keyword. If that model performs well (with high $F_{\max}$ ) then it gives more weight to the calculation of the total weighted average of that keyword. . . . .	40
656	9	AUROC of top 5 teams in CAFA- $\pi$ . The best performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. Four baseline models all based on BLAST were computed for <i>Candida</i> , while six baseline models were computed for <i>Pseudomonas</i> , including two based on Expression profiles. . . . .	41

660	10	AUROC of top 5 teams in CAFA- $\pi$ . The best performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. . . . .	42
661			
662	11	AUROC of top five teams in CAFA3. The best performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. . . . .	43
663			
664	12	CAFA participation has been growing. Each Principle Investigator is allowed to head multiple teams, but each member can only belong to one team. Each team can submit up to three models. . . . .	44
665			
666	13	CAFA3 timeline . . . . .	45
667	14	Number of proteins in each benchmark species and ontology. . . . .	46
668	15	Experimental procedure of determining genes associated with the functions biofilm formation and motility in <i>P. aeruginosa</i> . . . . .	47
669			
670	16	Experimental procedure of determining genes associated with the functions biofilm formation in <i>C. albicans</i> . . . . .	48
671			



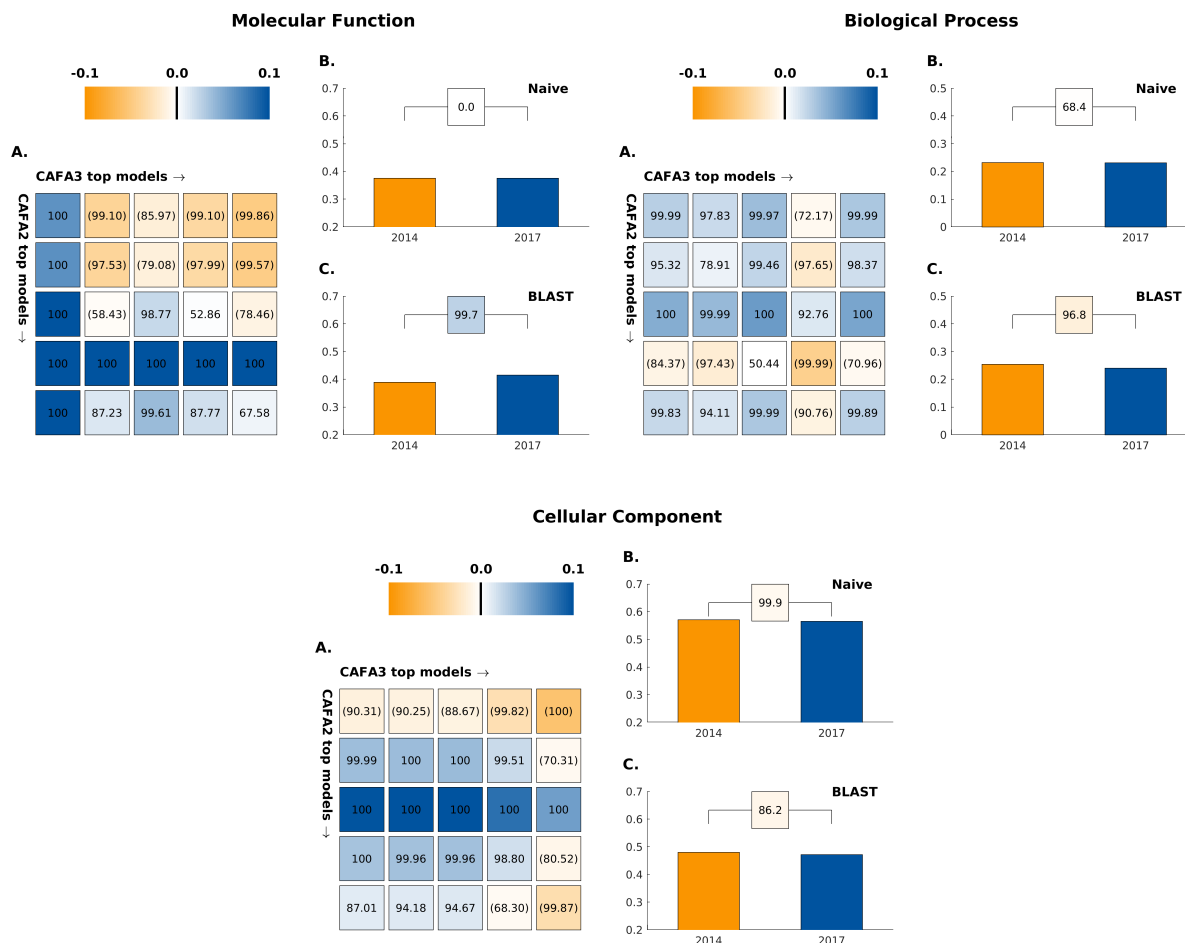


Figure 1: A comparison in  $F_{\max}$  between the top-five CAFA2 models against the top-five CAFA3 models. Colored boxes encode the results such that (1) the colors indicate margins of a CAFA3 method over a CAFA2 method in  $F_{\max}$  and (2) the numbers in the box indicate the percentage of wins. A: CAFA2 top-five models (rows, from top to bottom) against CAFA3 top-five models (columns, from left to right). B: Comparison of performance ( $F_{\max}$ ) of Naïve baselines trained respectively on SwissProt2014 and SwissProt2017. C: Comparison of performance ( $F_{\max}$ ) of BLAST baselines trained on SwissProt2014 and SwissProt2017. Statistical significance was assessed using 10,000 bootstrap samples of benchmark proteins.

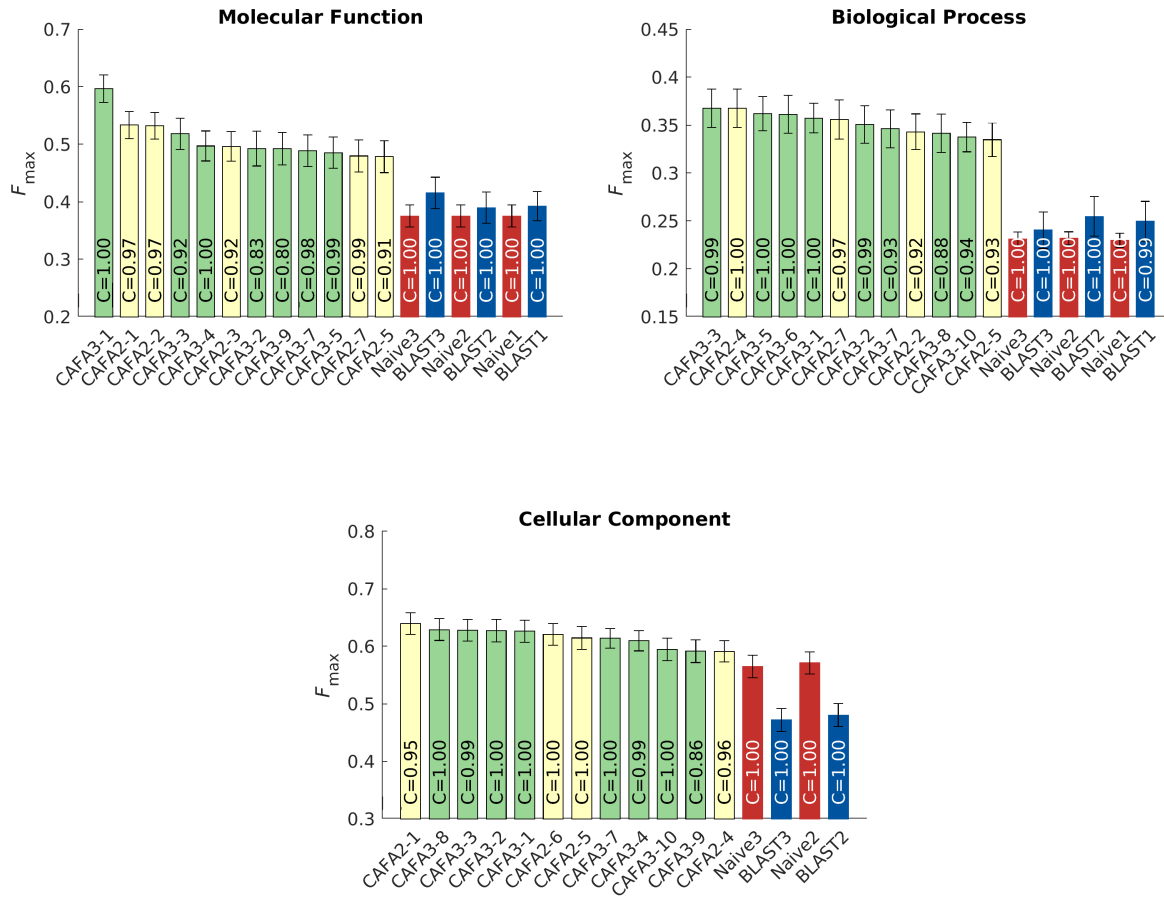


Figure 2: Performance evaluation based on  $F_{\max}$  for top CAFA1, CAFA2 and CAFA3 methods. The top 12 methods are shown in this barplot ranked in descending order from left to right. The baseline methods are appended to the right, they were trained on training data from 2017, 2014 and 2011 respectively. Coverage of the methods were shown as text inside the bars. Coverage is defined as percentage of proteins in the benchmark that are predicted by the methods. Color scheme: CAFA2: ivory; CAFA3: green; Naive: red; BLAST: blue. Note that in MFO and BPO, CAFA1 methods were ranked but not displayed. CAFA1 challenge did not collect predictions for CCO.

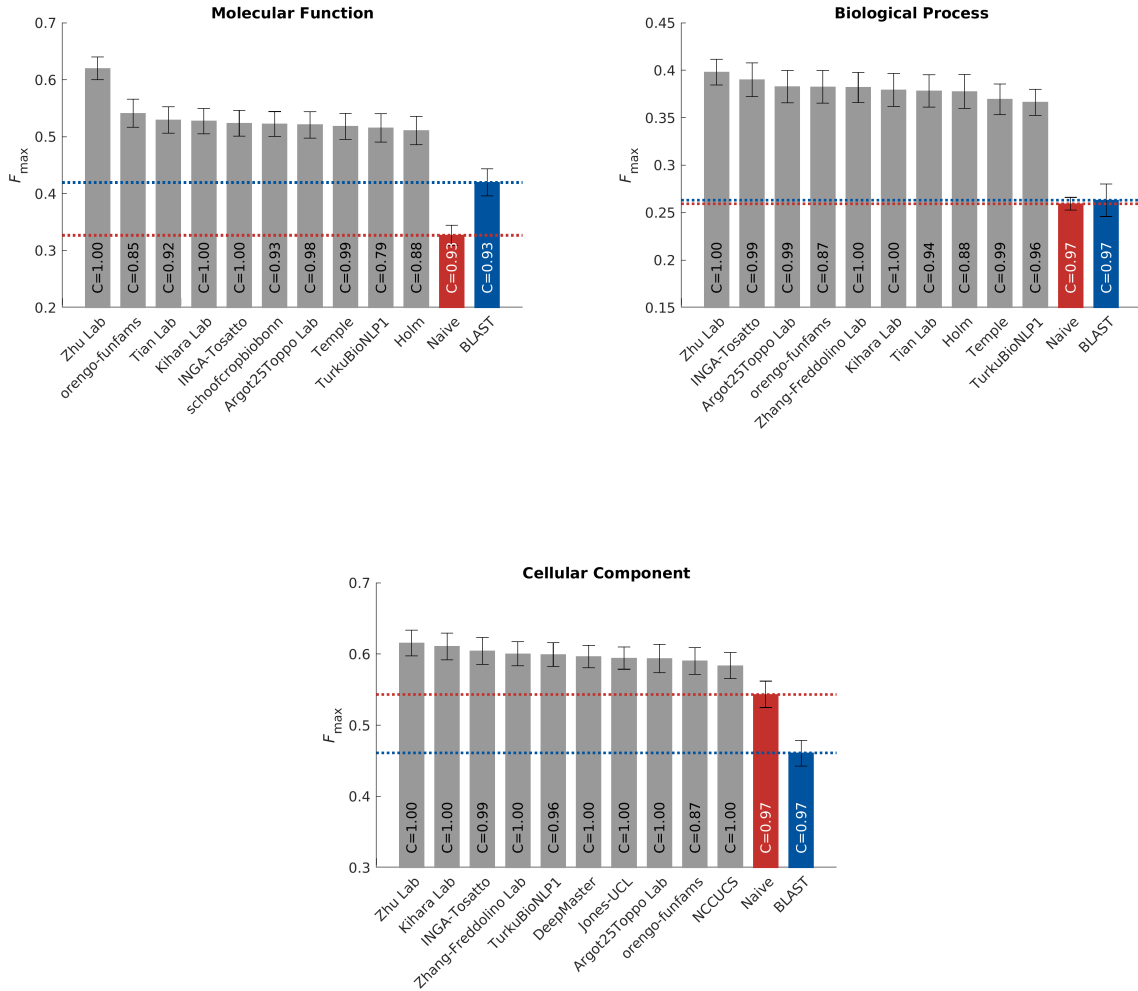


Figure 3: Performance evaluation based on  $F_{max}$  for the top-performing methods in three ontologies. The 95% confidence interval was estimated using  $10,000 \times$  bootstrap on the benchmark set. Coverage of the methods were shown as text inside the bars. Coverage is defined as percentage of proteins in the benchmark that are predicted by the methods.

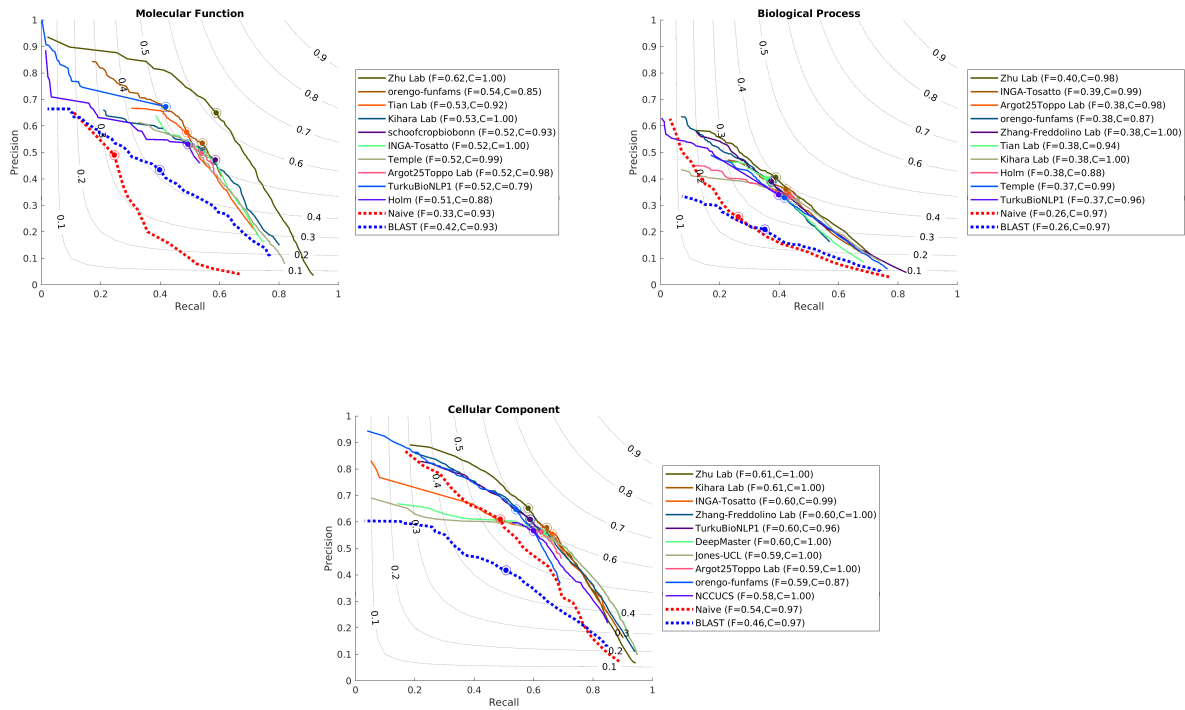


Figure 4: Precision Recall curves for the top-performing methods

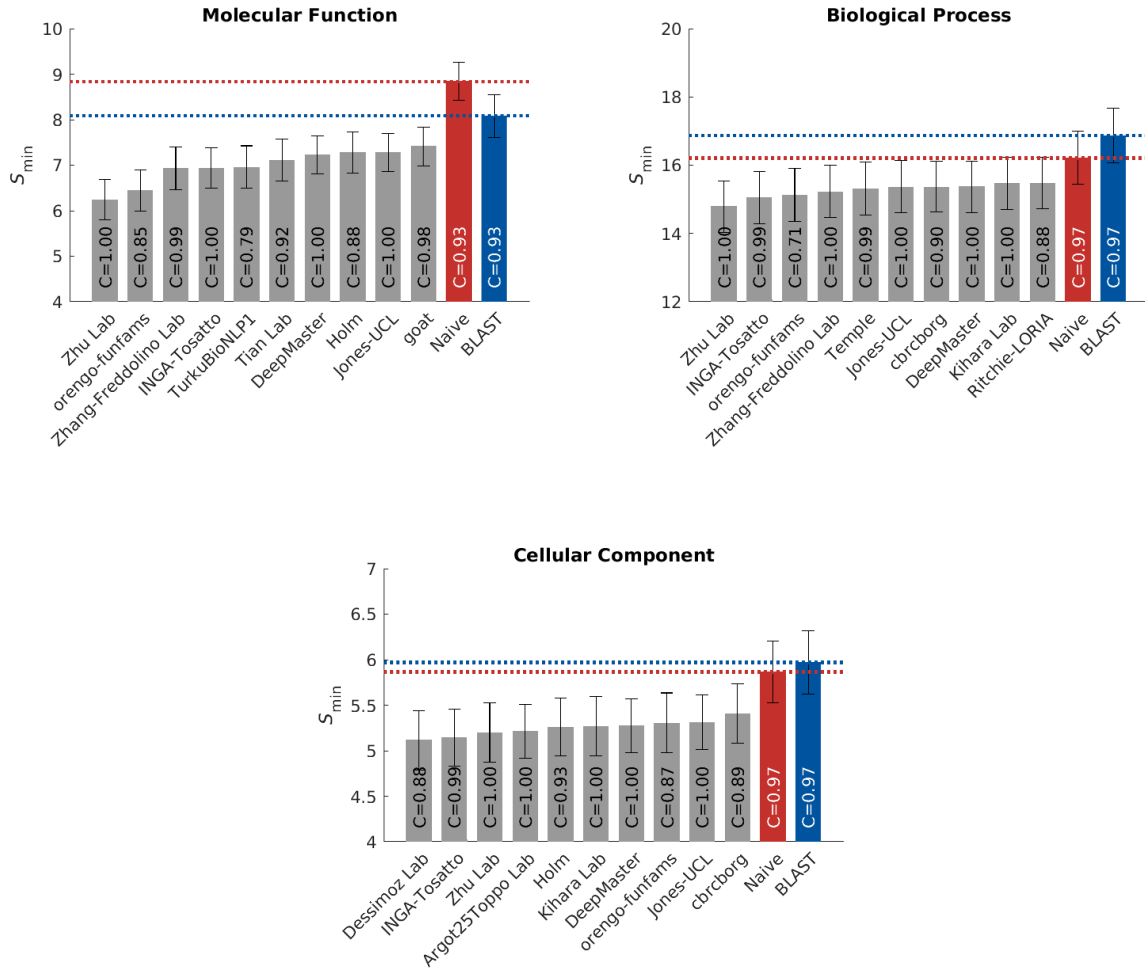
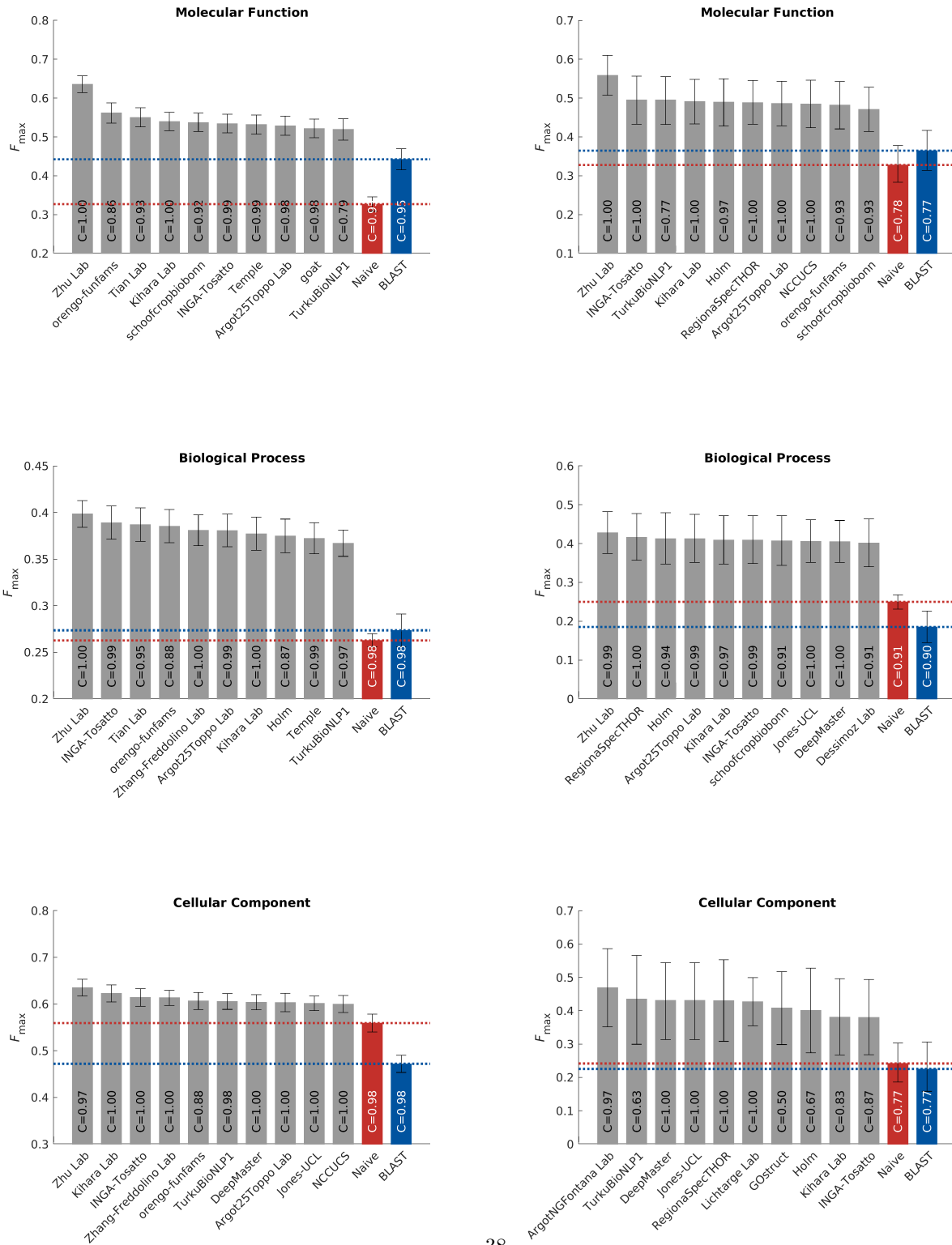


Figure 5: Evaluation based on the  $S_{\min}$  for the top-performing methods



(a) Eukarya

(b) Prokarya

Figure 6: Evaluation based on the  $F_{max}$  for the top-performing methods in eukaryotic and prokaryotic species

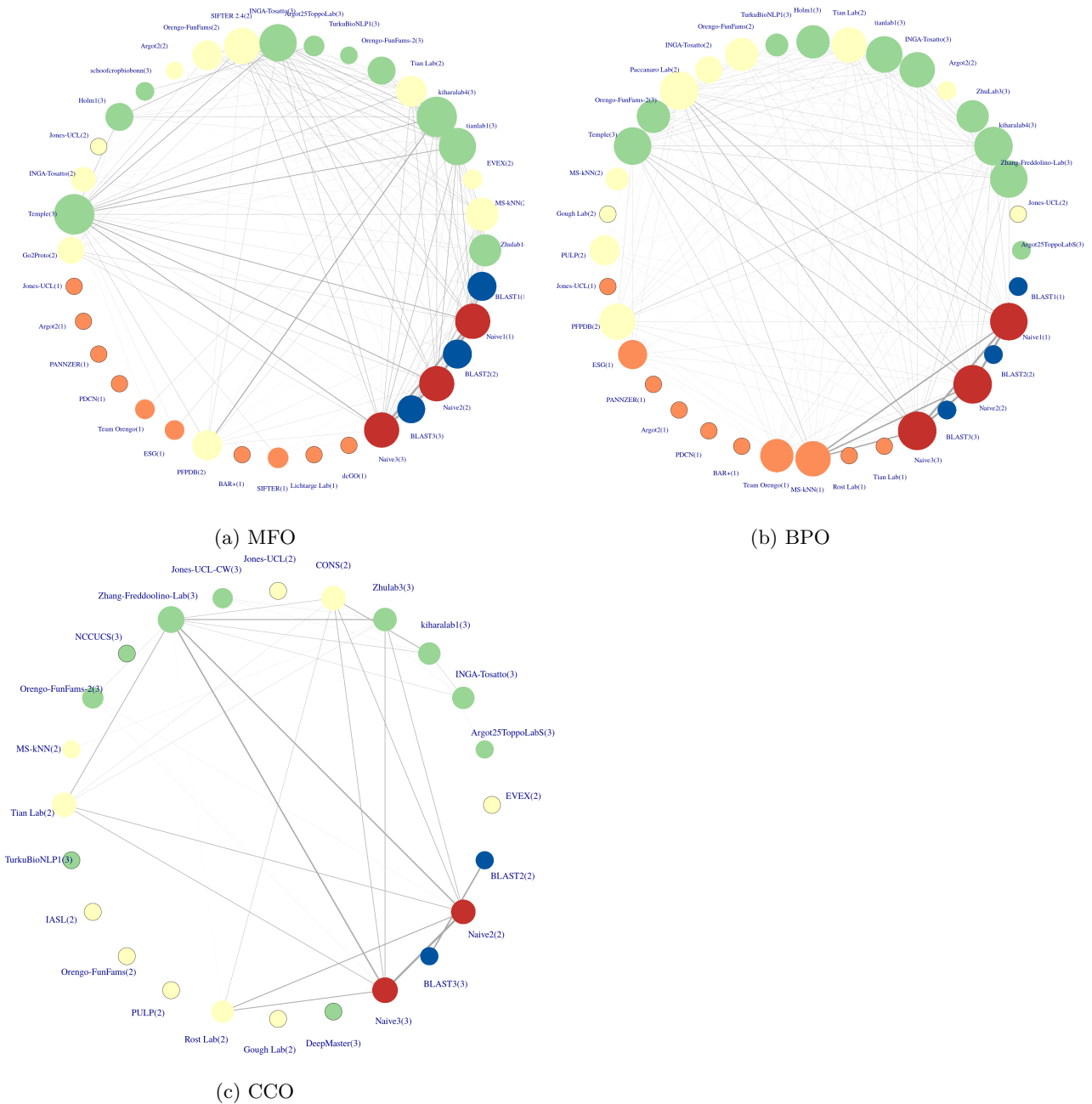
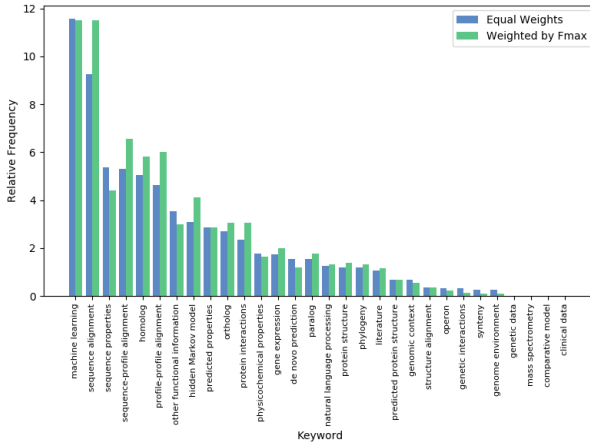
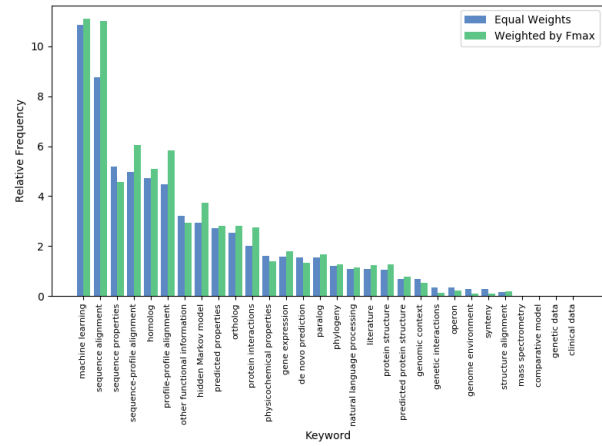


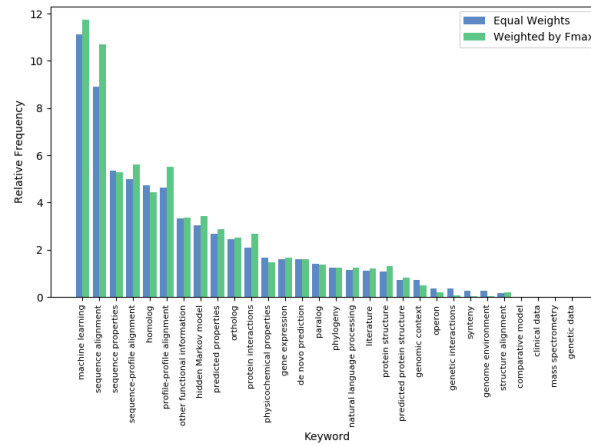
Figure 7: Similarity networks of top 10 methods from CAFA1, CAFA2 and CAFA3. The team names are displayed together with which CAFA challenge they come from in parenthesis. Similarity is calculated as the reciprocal of the Euclidean distance of the prediction scores from each pair of methods. A 0.07 cutoff was applied to the Euclidean distances, i.e. an edge exists if the Euclidean distance is lower than the cutoff. Edge width is directly proportional to similarity, except at the three edges between the three Naïve methods, where the similarity is much larger than the rest. Vertex size is directly proportional to number of edges, or degree of a vertex. Singletons, or vertices without any edges are framed with black circles. The nodes are ranked counter-clockwise, starting after 'BLAST1', by  $F_{\max}$  performance in the intersection set of benchmarks in Section 2.1. Color scheme: CAFA1: orange; CAFA2: ivory; CAFA3: green; Naïve: red; BLAST: blue.



(a) Molecular Function



(b) Biological Process



(c) Cellular Component

Figure 8: Keyword analysis of all CAFA3 participating methods. Both relative frequency of the keywords and weighted frequency are provided. The weighted frequencies accounts for the performance of the the particular model using the given keyword. If that model performs well (with high  $F_{max}$ ) then it gives more weight to the calculation of the total weighted average of that keyword.



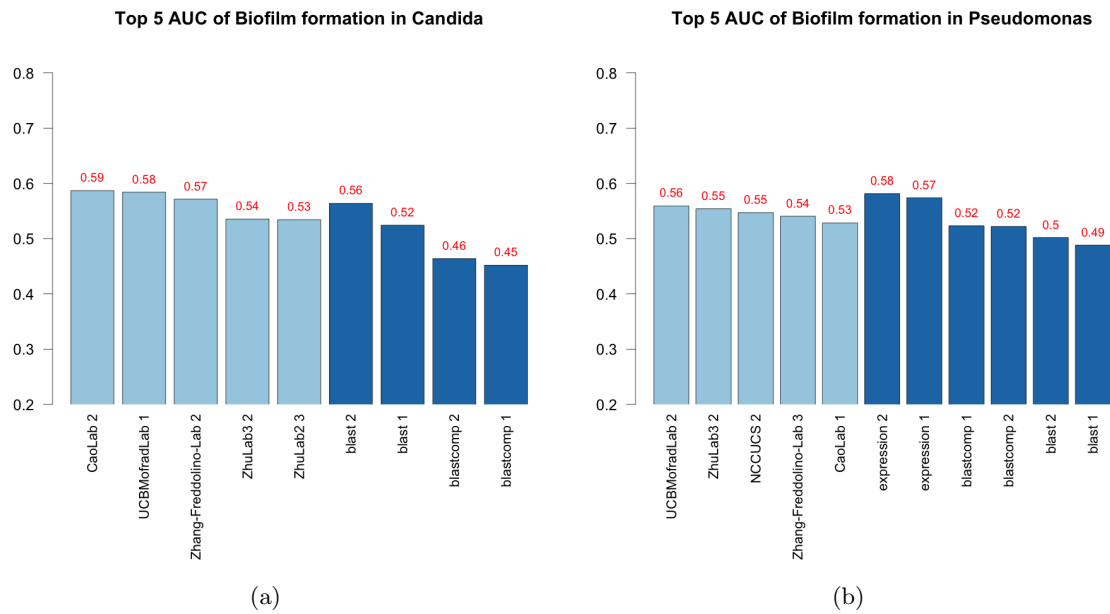


Figure 9: AUROC of top 5 teams in CAFA- $\pi$ . The best performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1. Four baseline models all based on BLAST were computed for *Candida*, while six baseline models were computed for *Pseudomonas*, including two based on Expression profiles.

### Top 5 AUC of Motility in Pseudomonas

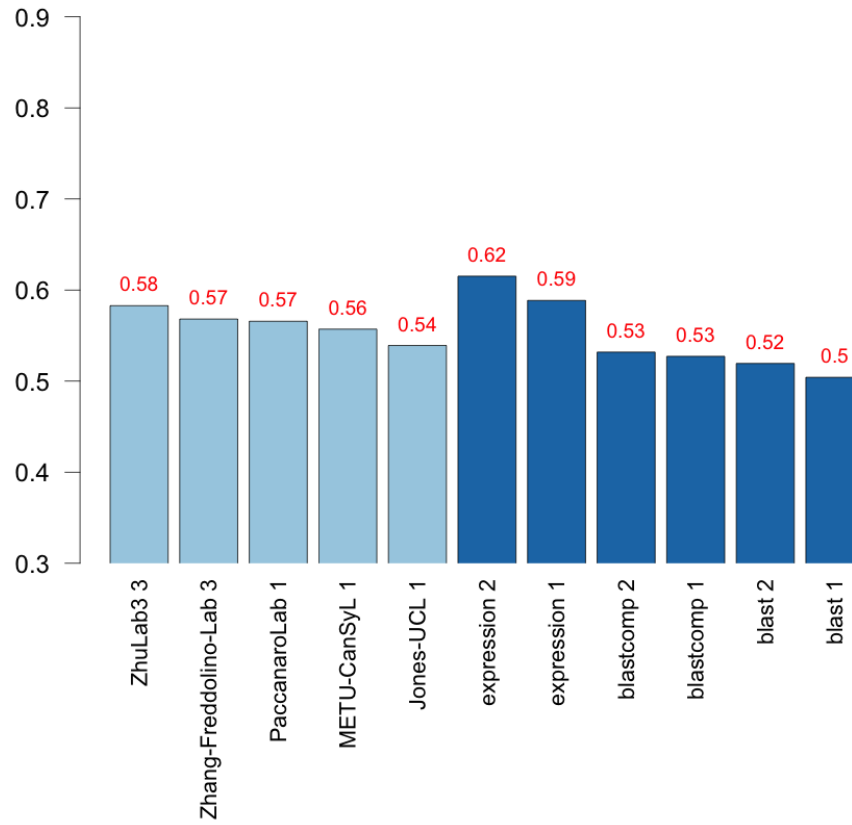


Figure 10: AUROC of top 5 teams in CAFA- $\pi$ . The best performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1.

### Top 5 AUC of long-term memory in Drosophila

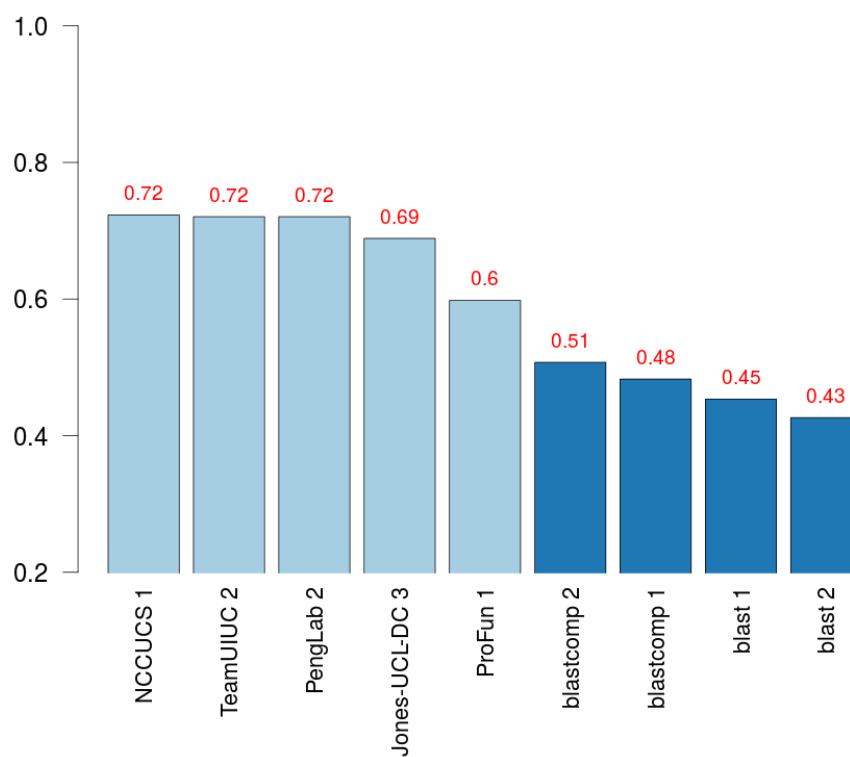


Figure 11: AUROC of top five teams in CAFA3. The best performing model from each team is picked for the top five teams, regardless of whether that model is submitted as model 1.

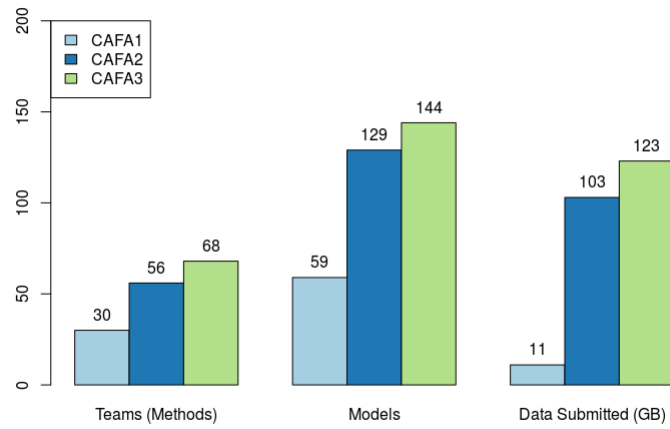


Figure 12: CAFA participation has been growing. Each Principle Investigator is allowed to head multiple teams, but each member can only belong to one team. Each team can submit up to three models.

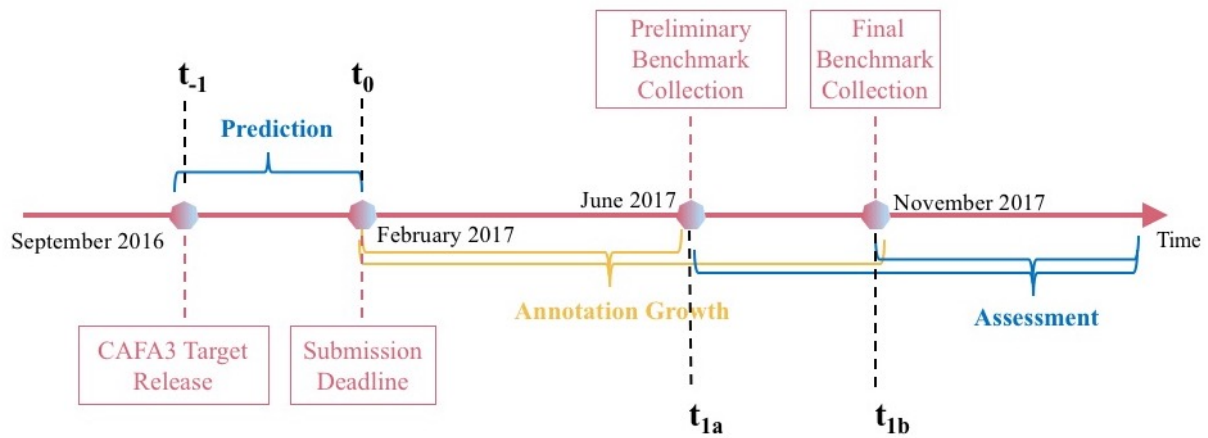


Figure 13: CAFA3 timeline

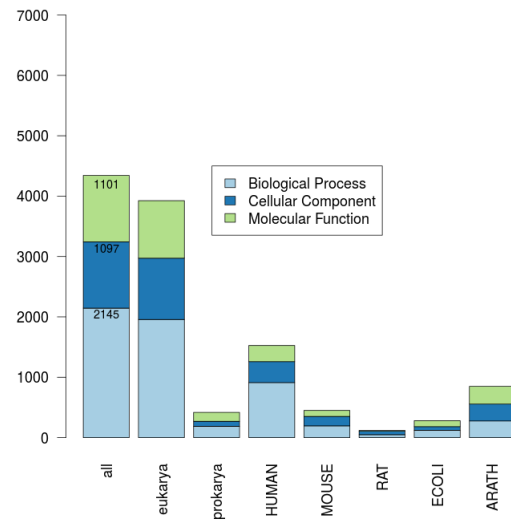
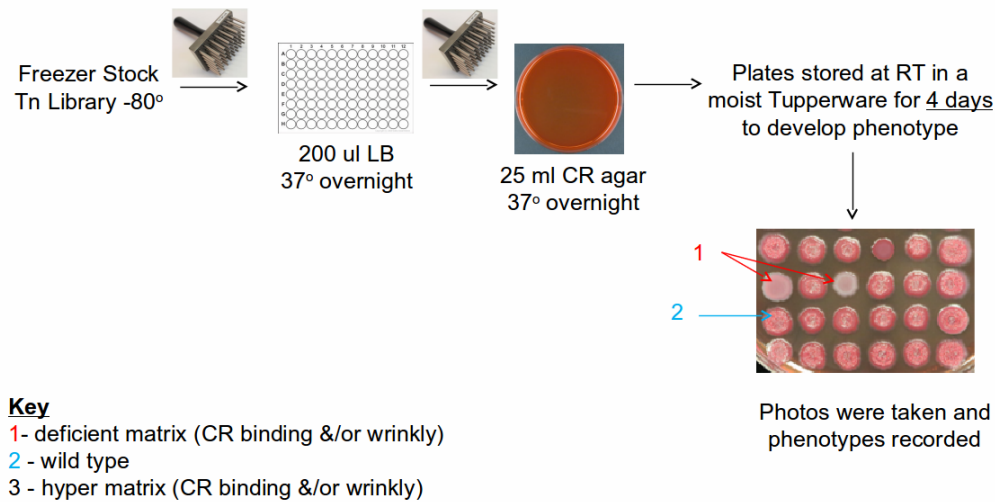
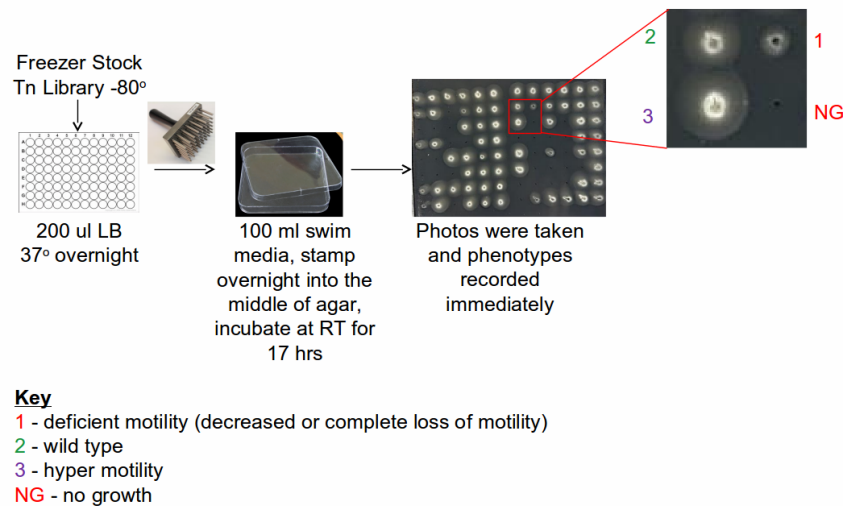


Figure 14: Number of proteins in each benchmark species and ontology.

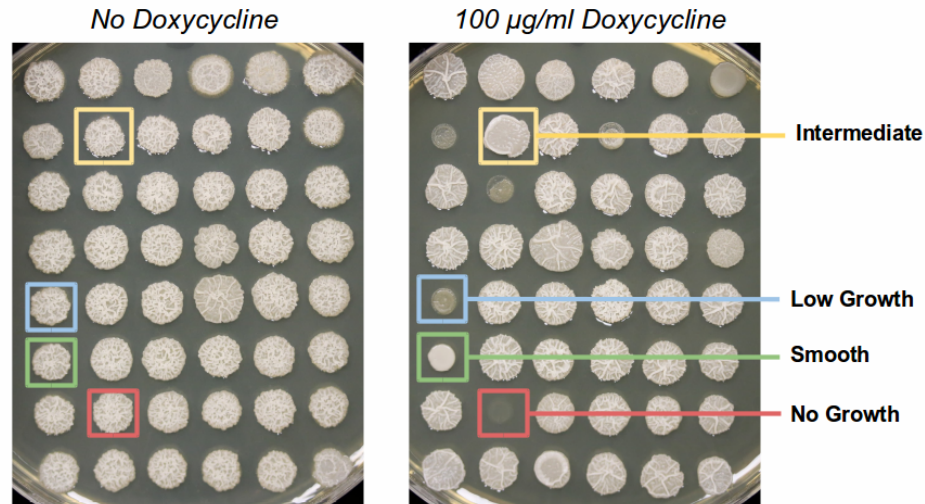


(a) Biofilm screen

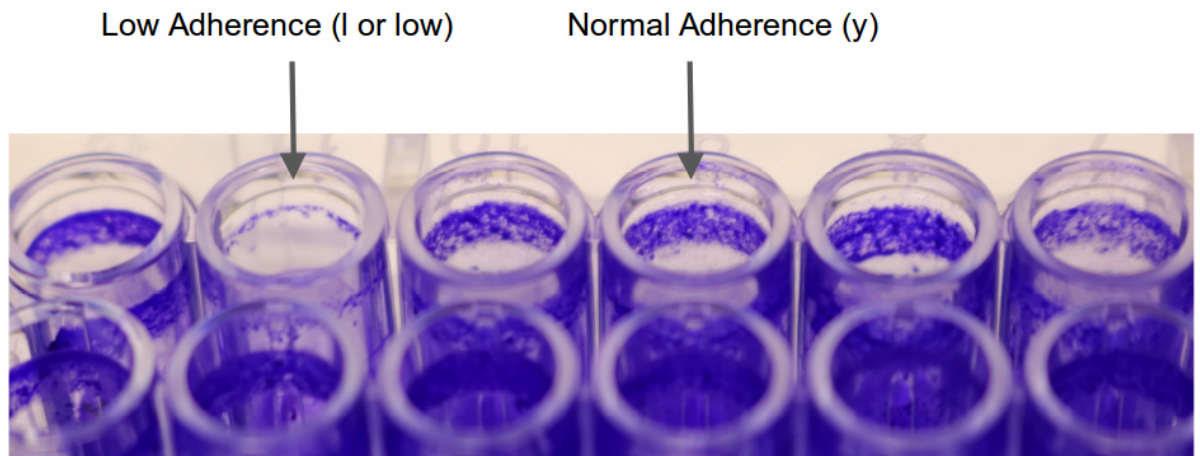


(b) Motility screen

Figure 15: Experimental procedure of determining genes associated with the functions biofilm formation and motility in *P. aeruginosa*



(a) wrinkle phenotype



(b) adherence phenotype

Figure 16: Experimental procedure of determining genes associated with the functions biofilm formation in *C. albicans*