a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC Bioinformatics*, 19(1):334, Sep 2018.
23. Maxat Kulmanov, Mohammed Asif Khan, and Robert Hoehndorf. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. *Bioinformatics*, 34(4):660–668, 10 2017. ISSN 1367-4803. doi: 10.1093/bioinformatics/btx624.
24. R. Cao, C. Freitas, L. Chan, M. Sun, H. Jiang, and Z. Chen. ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. *Molecules*, 22(10), Oct 2017.
25. J. J. Almagro Armenteros, C. K. S?nderby, S. K. S?nderby, H. Nielsen, and O. Winther. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics*, 33(21):3387–3395, Nov 2017.
26. Ariel S Schwartz, Gregory J Hannum, Zach R Dwiel, Michael E Smoot, Ana R Grant, Jason M Knight, Scott A Becker, Jonathan R Eads, Matthew C LaFave, Harini Eavani, Yinyin Liu, Arjun K Bansal, and Toby H Richardson. Deep semantic protein representation for annotation, discovery, and engineering. *bioRxiv*, 2018. doi: 10.1101/365965.
27. A. Sureyya Rifaioglu, T. Do?an, M. Jesus Martin, R. Cetin-Atalay, and V. Atalay. DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks. *Sci Rep*, 9(1):7344, May 2019.
28. Y. Li, S. Wang, R. Umarov, B. Xie, M. Fan, L. Li, and X. Gao. DEEPre: sequence-based enzyme EC number prediction by deep learning. *Bioinformatics*, 34(5):760–769, 03 2018.
29. J. Hou, B. Adhikari, and J. Cheng. DeepSF: deep convolutional neural network for mapping protein sequences to folds. *Bioinformatics*, 34(8):1295–1303, 04 2018.
30. Maria Littmann, Michael Heinzinger, Christian Dallago, Tobias Olenyi, and Burkhard Rost. Embeddings from deep learning transfer go annotations beyond homology. *Scientific reports*, 11(1):1–14, 2021.
31. Mohammed AlQuraishi. End-to-end differentiable learning of protein structure. *Cell systems*, 8(4):292–301, 2019.
32. Andrew W Senior, Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Zidek, Alexander WR Nelson, Alex Bridgland, et al. Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792): 706–710, 2020.
33. Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020. ISSN 0027-8424. doi: 10.1073/pnas.1914677117.
34. Jianyi Yang, Ivan Anishchenko, Hahnbeom Park, Zhenling Peng, Sergey Ovchinnikov, and David Baker. Improved protein structure prediction using predicted interresidue orientations. *Proceedings of the National Academy of Sciences*, 117(3):1496–1503, 2020.
35. Yilun Du, Joshua Meier, Jerry Ma, Rob Fergus, and Alexander Rives. Energy-based models for atomic-resolution protein conformations. In *International Conference on Learning Representations*, 2019.
36. Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. 139:8844–8856, 18–24 Jul 2021.
37. Adam J Riesselman, John B Ingraham, and Debora S Marks. Deep generative models of genetic variation capture the effects of mutations. *Nature methods*, 15(10):816–822, 2018.
38. Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15), 2021.
39. Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Xi Chen, John Canny, Pieter Abbeel, and Yun S Song. Evaluating protein transfer learning with tape. *Advances in Neural Information Processing Systems*, 32:9689, 2019.
40. J. Frazer, P. Notin, M. Dias, A. Gomez, J. K. Min, K. Brock, Y. Gal, and D. S. Marks. Disease variant prediction with deep generative models of evolutionary data. *Nature*, 599(7883): 91–95, 11 2021.
41. Kevin K Yang, Zachary Wu, and Frances H Arnold. Machine-learning-guided directed evolution for protein engineering. *Nature methods*, 16(8):687–694, 2019.
42. Stanislav Mazurenko, Zbynek Prokop, and Jiri Damborsky. Machine learning in enzyme engineering. *ACS Catalysis*, 10(2):1210–1223, 2019.
43. S. Biswas, G. Khimulya, E. C. Alley, K. M. Esvelt, and G. M. Church. Low-N protein engineering with data-efficient deep learning. *Nat Methods*, 18(4):389–396, 04 2021.
44. Ali Madani, Bryan McCann, Nikhil Naik, Nitish Shirish Keskar, Namrata Anand, Raphael R Eguchi, Po-Ssu Huang, and Richard Socher. Progen: Language modeling for protein generation. *arXiv preprint arXiv:2004.03497*, 2020.
45. Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, et al. Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing. *arXiv preprint arXiv:2007.06225*, 2020.
46. I. Anishchenko, S. J. Pellock, T. M. Chidyausiku, T. A. Ramelot, S. Ovchinnikov, J. Hao, K. Bafna, C. Norn, A. Kang, A. K. Bera, F. DiMaio, L. Carter, C. M. Chow, G. T. Montelione, and D. Baker. De novo protein design by deep network hallucination. *Nature*, 600(7889): 547–552, 12 2021.
47. Drew H Bryant, Ali Bashir, Sam Sinai, Nina K Jain, Pierce J Ogden, Patrick F Riley, George M Church, Lucy J Colwell, and Eric D Kelsic. Deep diversification of an aav capsid protein by machine learning. *Nature Biotechnology*, pages 1–6, 2021.
48. N. Brandes, D. Ofer, Y. Peleg, N. Rappoport, and M. Linial. ProteinBERT: A universal deep-learning model of protein sequence and function. *Bioinformatics*, Jan 2022.
49. M. L. Bileschi, D. Belanger, D. H. Bryant, T. Sanderson, B. Carter, D. Sculley, A. Bateman, M. A. DePristo, and L. J. Colwell. Using deep learning to annotate the protein universe. *Nat Biotechnol*, 40(6):932–937, 06 2022.
50. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.
51. Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015.
52. David Dohan, Andreea Gane, Maxwell Bileschi, David Belanger, and Lucy Colwell. Improving protein function annotation via unsupervised pre-training: Robustness, efficiency, and insights. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2021.
53. Predrag Radivojac, Wyatt T Clark, Tal Ronnen Oron, Alexandra M Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, Christopher Funk, Karin Verspoor, Asa Ben-Hur, et al. A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227, 2013.
54. Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *The International Conference on Learning Representations*, 2015.
55. Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. In *Advances in neural information processing systems*, pages 1223–1231, 2012.
56. Christopher J. Shallue, Jaehoon Lee, Joseph M. Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the effects of data parallelism on neural network training. *CoRR*, abs/1811.03600, 2018.
57. Julian Gough, Arun Prasad Pandurangan, Ben Smithers, Jonathan Stahlhacke, and Matt E Oates. The SUPERFAMILY 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Research*, 47(D1):D490–D494, 11 2018. ISSN 0305-1048. doi: 10.1093/nar/gky1130.
58. T. K. Attwood, P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri. PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res.*, 31(1):400–402, Jan 2003.
59. D. H. Haft, J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu, and E. Beck. TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Res.*, 41(Database issue):D387–395, Jan 2013.
60. H. Mi, S. Poudel, A. Muruganujan, J. T. Casagrande, and P. D. Thomas. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.*, 44 (D1):D336–342, Jan 2016.
61. S. Hunter, R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu, and C. Yeats. InterPro: the integrative protein signature database. *Nucleic Acids Res.*, 37(Database issue): D211–215, Jan 2009.
62. Baris E Suzek, Yuqi Wang, Hongzhan Huang, Peter B McGarvey, Cathy H Wu, and UniProt Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.
63. Naihui Zhou, Yuxiang Jiang, Timothy R Bergquist, Alexandra J Lee, Balint Z Kacsoh, Alex W Crocker, Kimberley A Lewis, George Georghiou, Huy N Nguyen, Md Nafiz Hamid, et al. The cafa challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. *Genome biology*, 20(1):1–23, 2019.
64. J. Gillis and P. Pavlidis. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). *BMC Bioinformatics*, 14 Suppl 3:S15, 2013.
65. Alex Warwick Vesztrocy and Christophe Dessimoz. Benchmarking gene ontology function predictions using negative annotations. *Bioinformatics*, 36(Supplement_1):i210–i218, 2020.
66. Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)*, pages 265–283, 2016.
67. Sean R Eddy. Accelerated profile hmm searches. *PLoS computational biology*, 7(10): e1002195, 2011.
68. Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. *CoRR*, abs/1512.04150, 2015.
69. Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.
70. D. Y. Chao, Y. Chen, J. Chen, S. Shi, Z. Chen, C. Wang, J. M. Danku, F. J. Zhao, and D. E. Salt. Genome-wide association mapping identifies a new arsenate reductase enzyme critical for limiting arsenic accumulation in plants. *PLoS Biol*, 12(12):e1002009, Dec 2014.
71. A. Bartels, H. P. Mock, and J. Papenbrock. Differential expression of Arabidopsis sulfurtransferases under various growth conditions. *Plant Physiol Biochem*, 45(3-4):178–187, 2007.
72. Ethan C Alley, Grigory Khimulya, Surojit Biswas, Mohammed AlQuraishi, and George M Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12):1315–1322, 2019.
73. Daniel Smilkov, Nikhil Thorat, Yannick Assogba, Ann Yuan, Nick Kreeger, Ping Yu, Kangyi Zhang, Shanqing Cai, Eric Nielsen, David Soergel, Stan Bileschi, Michael Terry, Charles Nicholson, Sandeep N. Gupta, Sarah Sirajuddin, D. Sculley, Rajat Monga, Greg Corrado, Fernanda B. Viegas, and Martin Wattenberg. Tensorflow.js: Machine learning for the web and beyond, 2019.
74. Philip Jones, David Binns, Hsin-Yu Chang, Matthew Fraser, Weizhong Li, Craig McAnulla, Hamish McWilliam, John Maslen, Alex Mitchell, Gift Nuka, Sebastien Pesseat, Antony F. Quinn, Amaia Sangrador-Vegas, Maxim Scheremetjew, Siew-Yit Yong, Rodrigo Lopez, and Sarah Hunter. InterProScan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240, 01 2014. ISSN 1367-4803. doi: 10.1093/bioinformatics/btu031.
75. Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.
76. Gene Ontology Consortium. The gene ontology resource: 20 years and still going strong. *Nucleic acids research*, 47(D1):D330–D338, 2019.
77. Seth Carbon, Amelia Ireland, Christopher J Mungall, ShengQiang Shu, Brad Marshall, Suzanna Lewis, AmiGO Hub, and Web Presence Working Group. Amigo: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289, 2009.
78. Yuxiang Jiang, Wyatt T Clark, Iddo Friedberg, and Predrag Radivojac. The impact of in-

complete knowledge on the evaluation of protein function prediction: a structured-output learning perspective. *Bioinformatics*, 30(17):i609–i616, 2014.
79. L. Li, C. J. Stoeckert, and D. S. Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, 13(9):2178–2189, Sep 2003.
80. Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. *CoRR*, abs/1706.04599, 2017.
81. Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016.
82. Alex Mitchell, Hsin-Yu Chang, Louise Daugherty, Matthew Fraser, Sarah Hunter, Rodrigo Lopez, Craig McAnulla, Conor McMenamin, Gift Nuka, Sebastien Pesseat, et al. The interpro protein families database: the classification resource after 15 years. *Nucleic acids research*, 43(D1):D213–D221, 2015.

Supplementary Note 1: Figure supplements

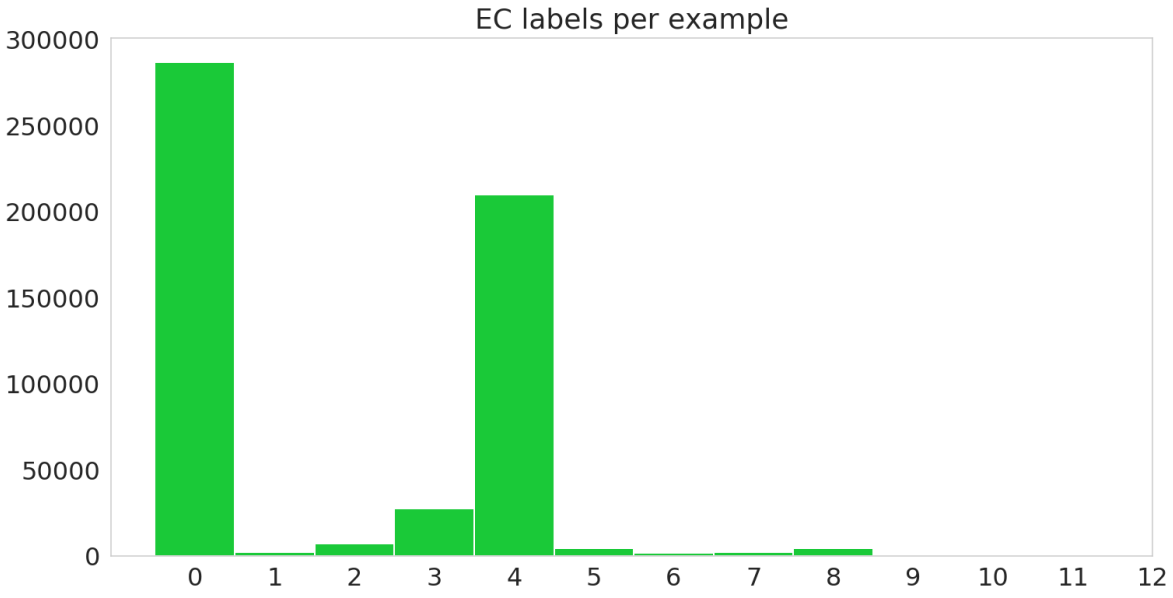


Fig. S1. [Figure 3 - Figure supplement 1] Histogram of number of labels per sequence, including hierarchical labels, on the random dataset.

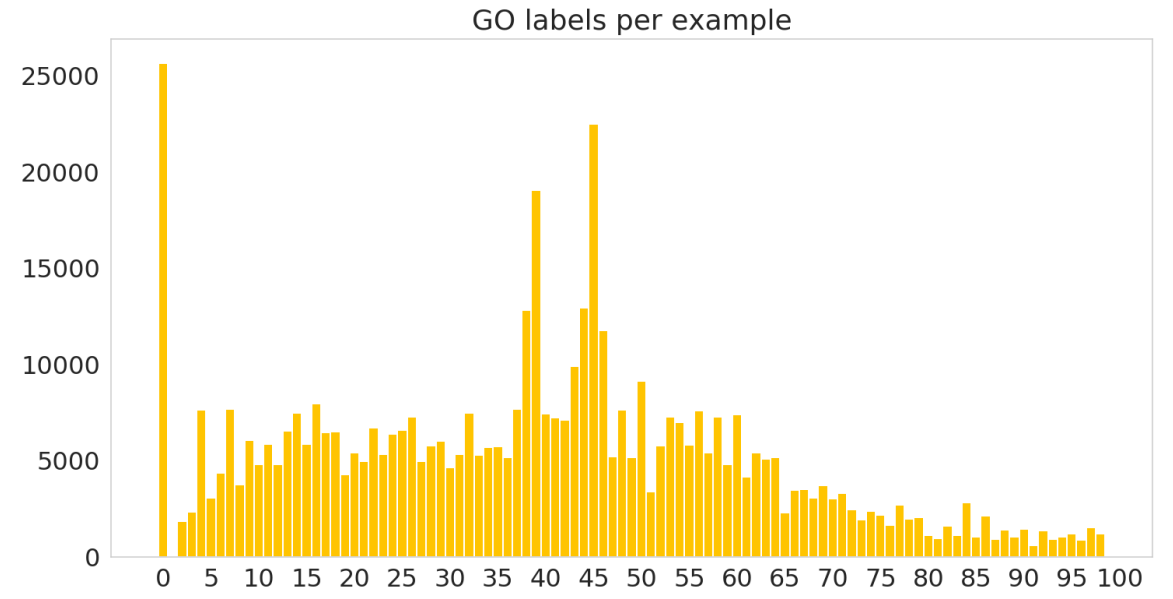


Fig. S2. [Figure 3 - figure supplement 2] Histogram of number of labels per sequence, including hierarchical labels, on the random dataset.

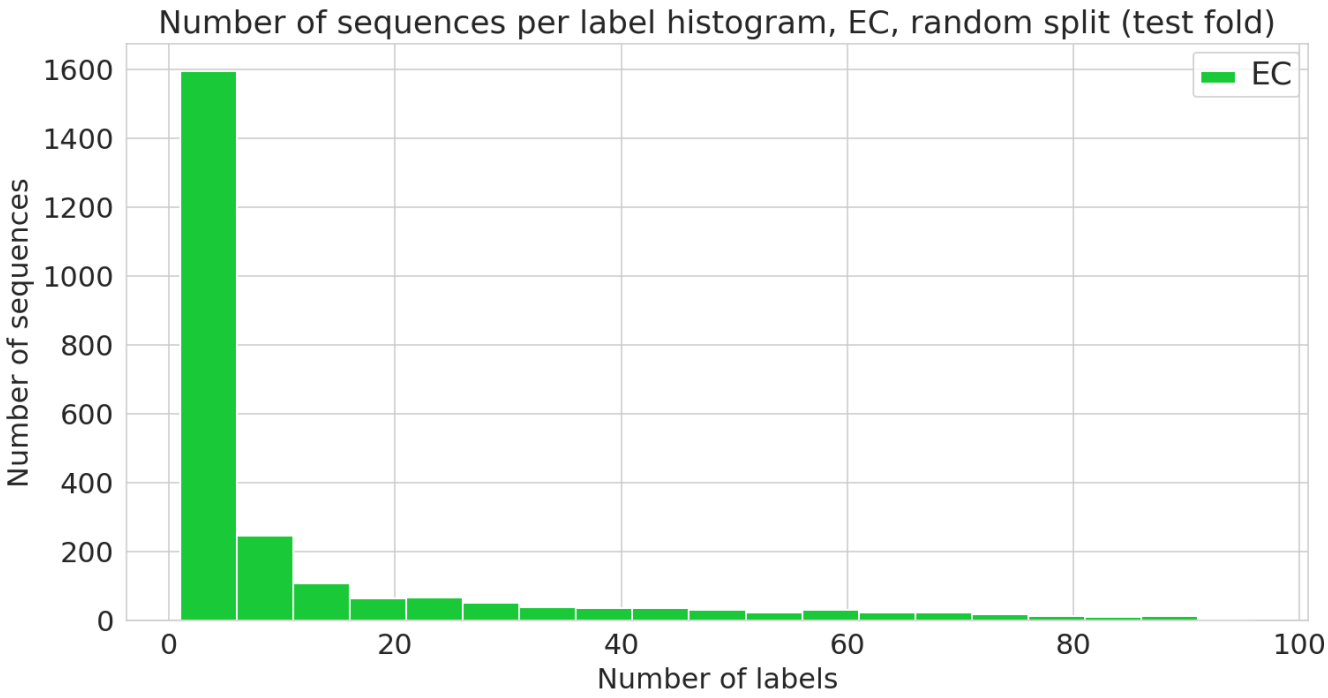


Fig. S3. [Figure 3 - figure supplement 3] Number of sequences annotated with a given functional label (EC class) in the random dataset.

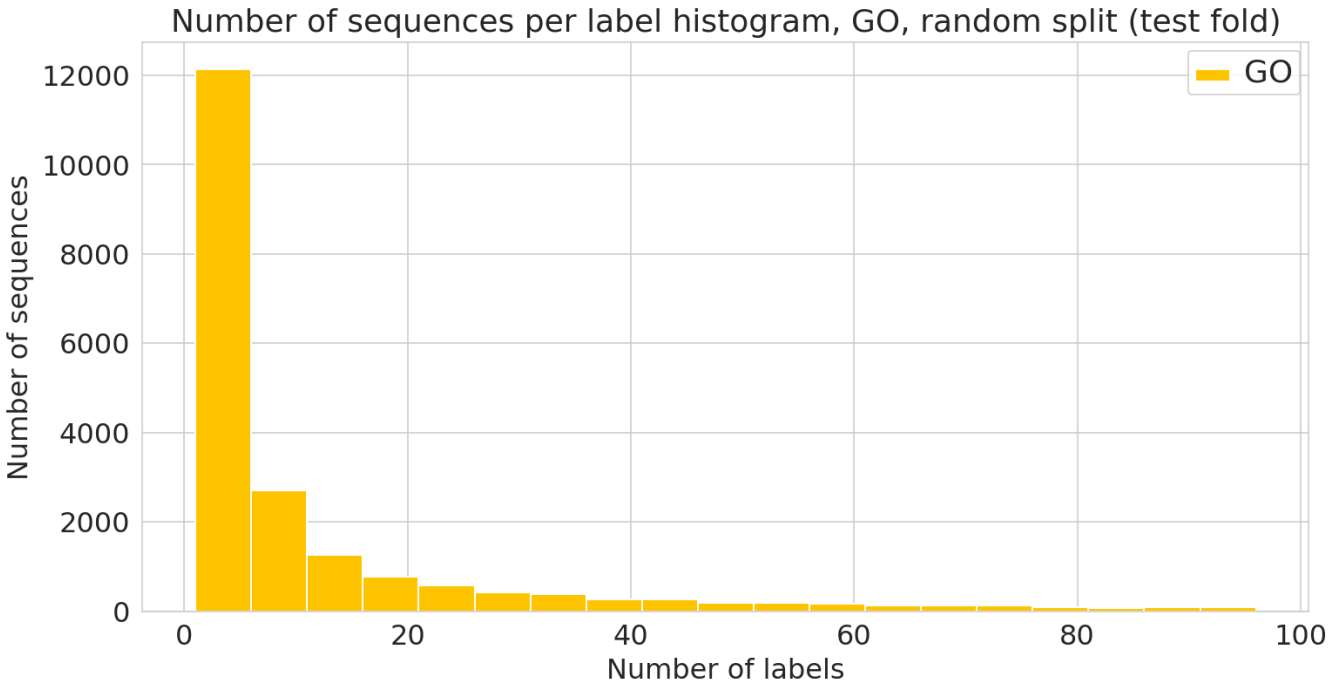


Fig. S4. [Figure 3 - figure supplement 4] Number of sequences annotated with a given functional label (GO label) in the random dataset.

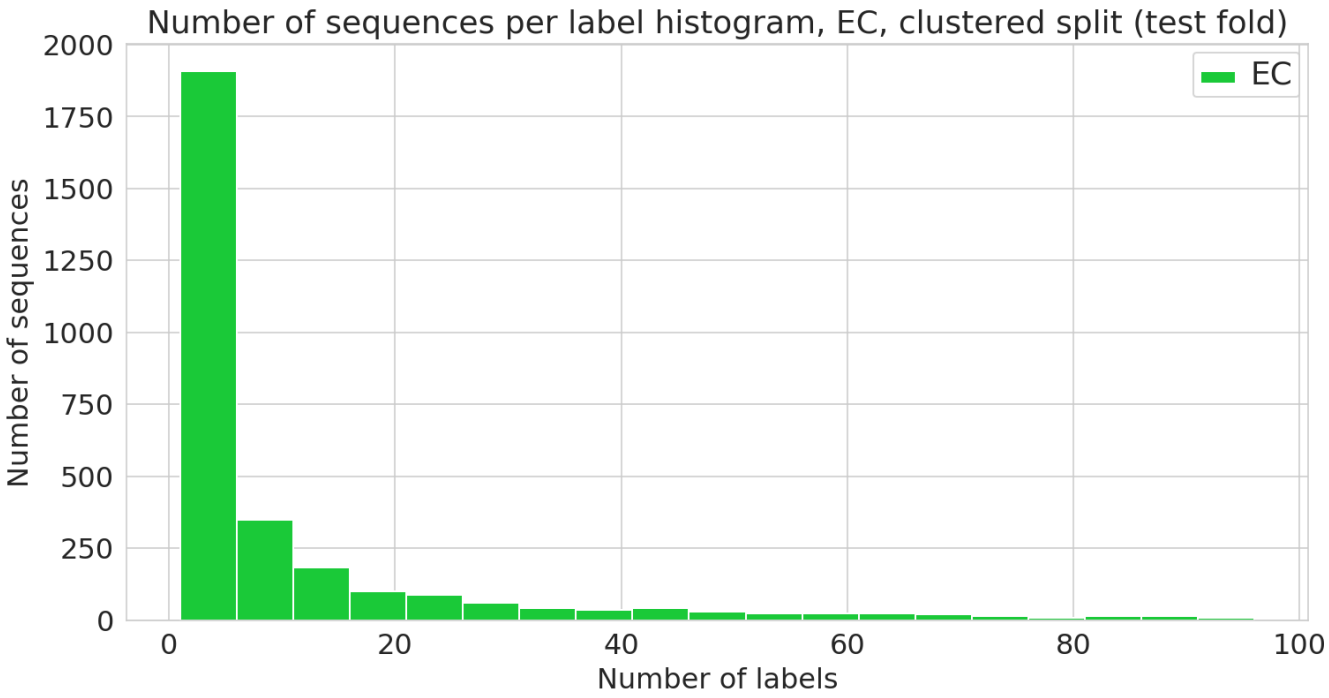


Fig. S5. [Figure 3 - figure supplement 5] Number of sequences annotated with a given functional label (EC class) in the clustered dataset.

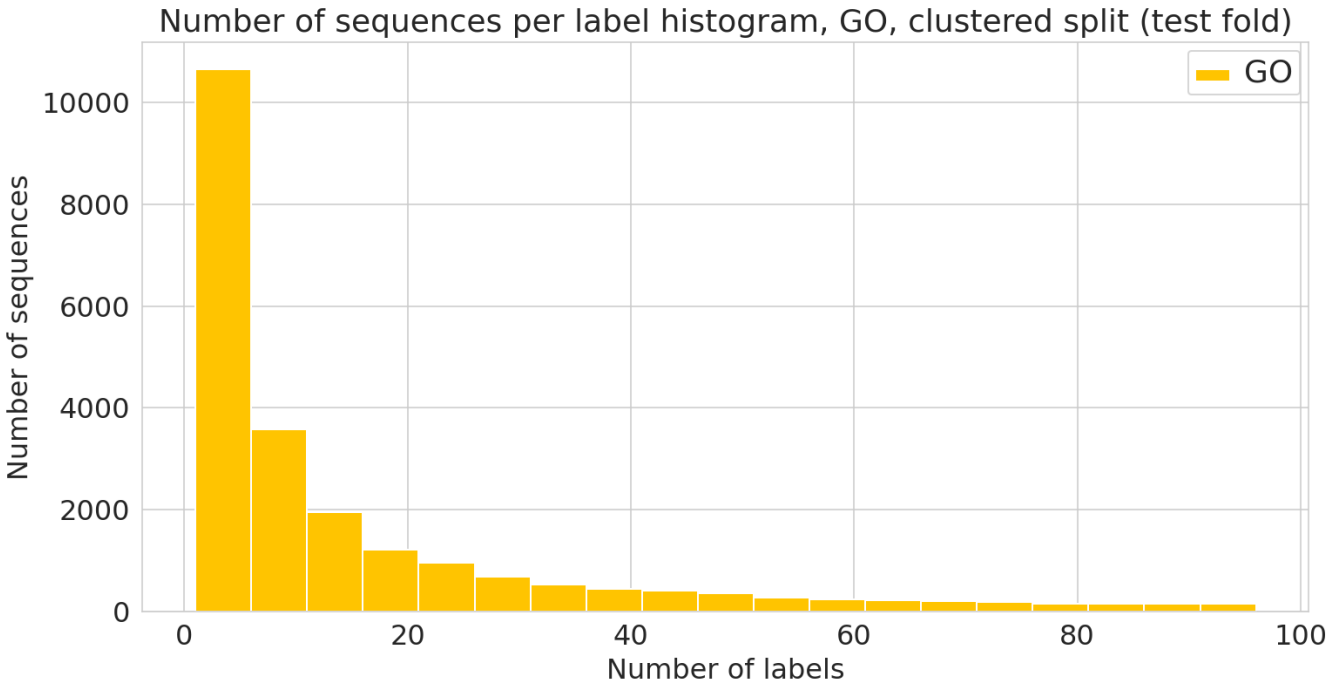
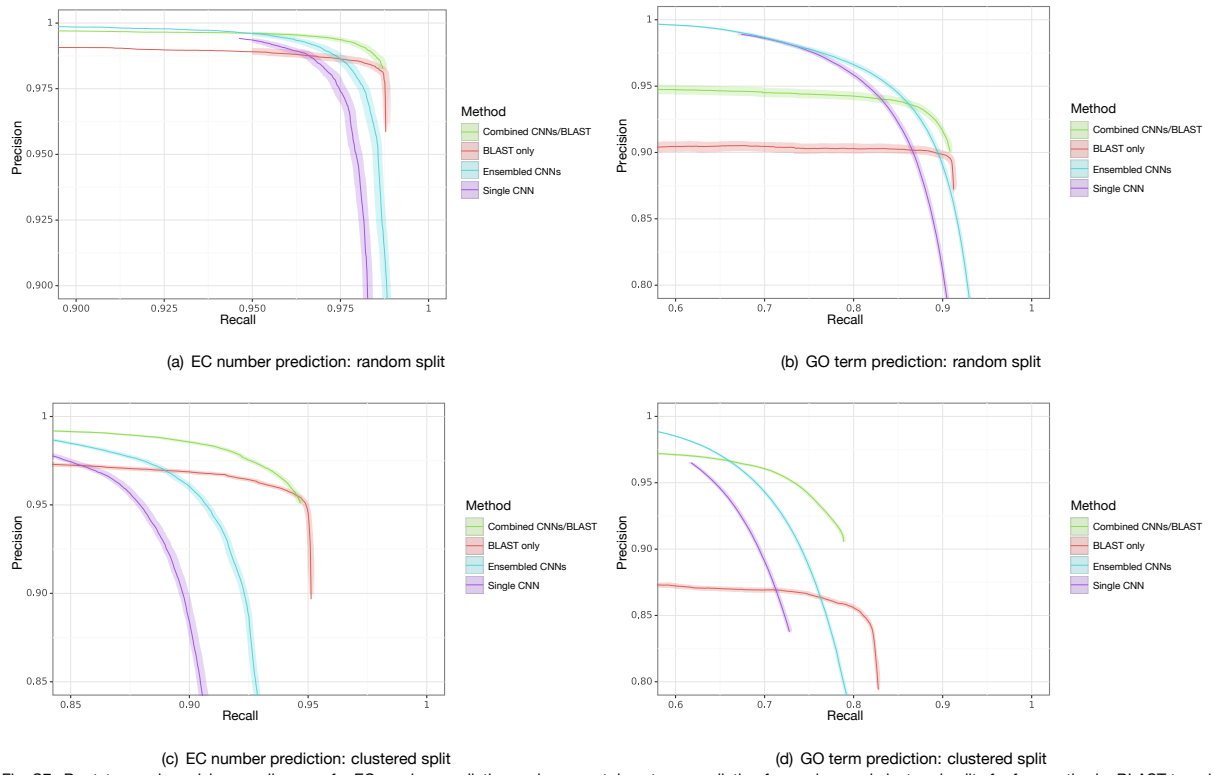


Fig. S6. [Figure 3 - figure supplement 6] Number of sequences annotated with a given functional label. (GO label) in the clustered dataset.



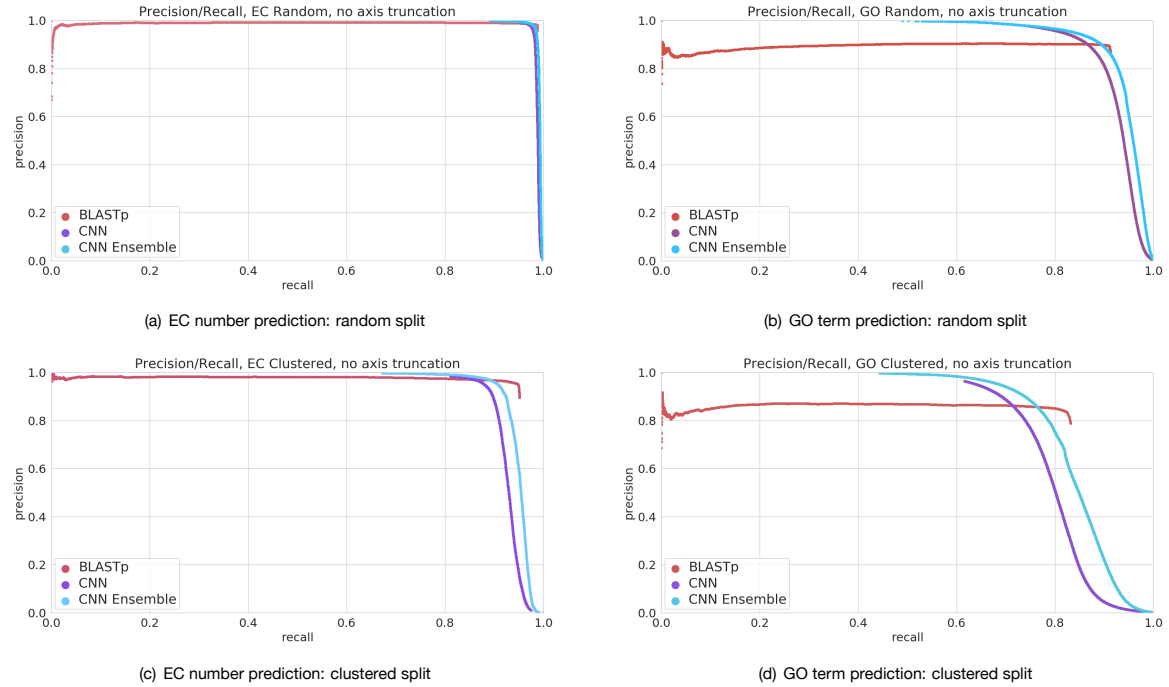
(a) EC number prediction: random split

(b) GO term prediction: random split

(c) EC number prediction: clustered split

(d) GO term prediction: clustered split

Fig. S7. [Figure 3 - figure supplement 7] Bootstrapped precision-recall curves for EC number prediction and gene ontology term prediction for random and clustered splits for four methods: BLAST top pick, single ProteInfer CNN, ensembled ProteInfer CNNs, and ensembled ProteInfer CNNs scaled by BLAST score.



(a) EC number prediction: random split

(b) GO term prediction: random split

(c) EC number prediction: clustered split

(d) GO term prediction: clustered split

Fig. S8. [Figure 3 - Figure supplement 8] Full precision-recall curves for EC number prediction and gene ontology term prediction for random and clustered splits for four methods: BLAST top pick, single ProteInfer CNN, ensembled ProteInfer CNNs

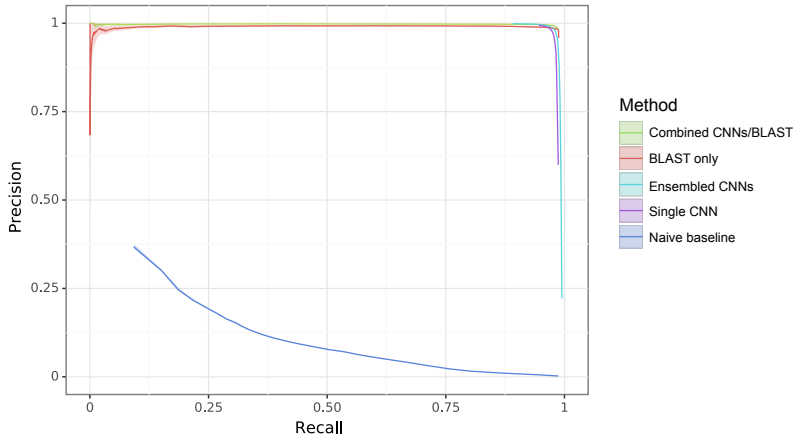


Fig. S9. [Figure 3 - Figure supplement 9] EC random task with different methods compared against a naive baseline where the predictor is simply the frequency in the training set.

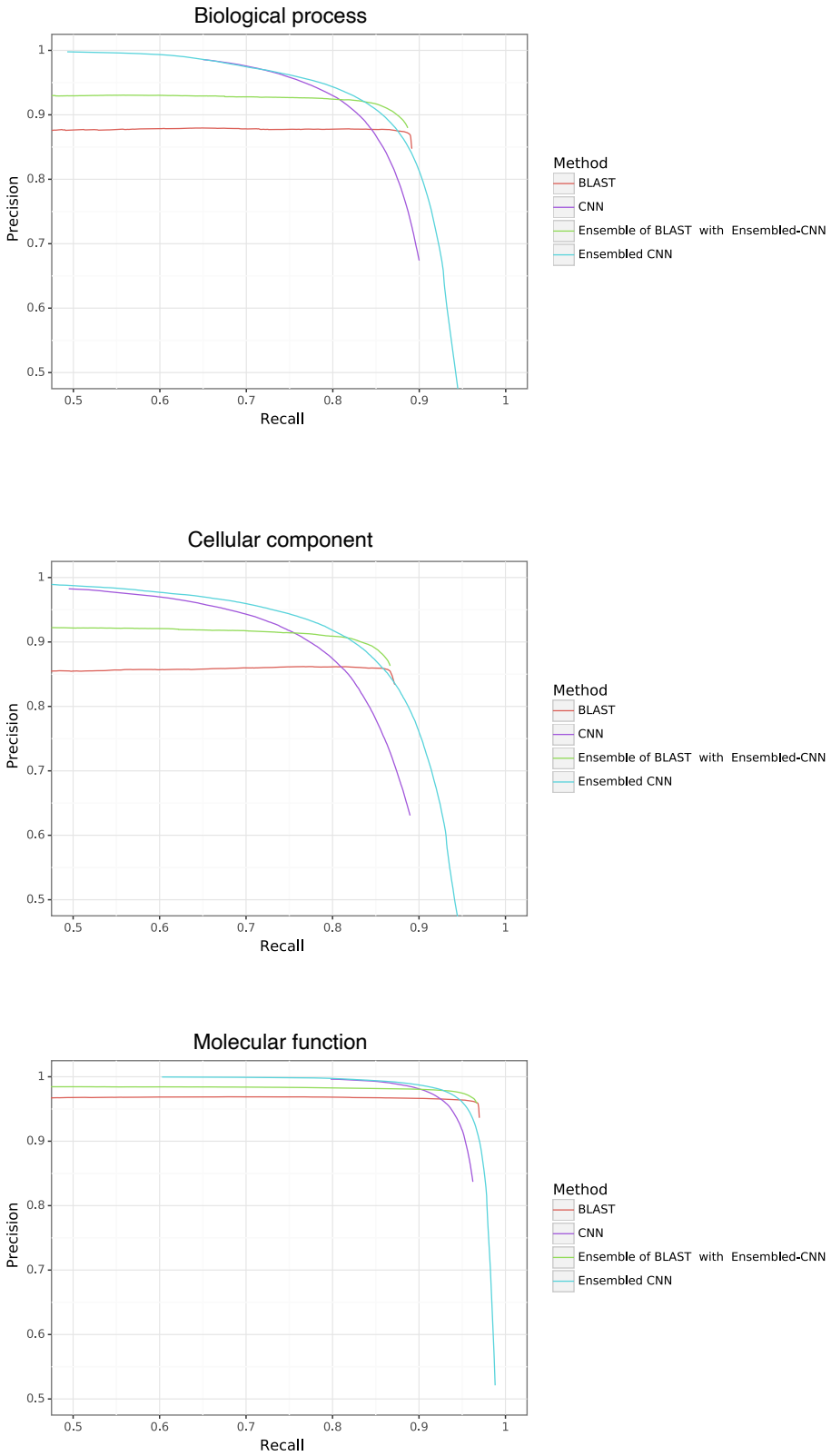


Fig. S10. [Figure 3 - Figure supplement 10] GO performance stratified by method and ontology type.

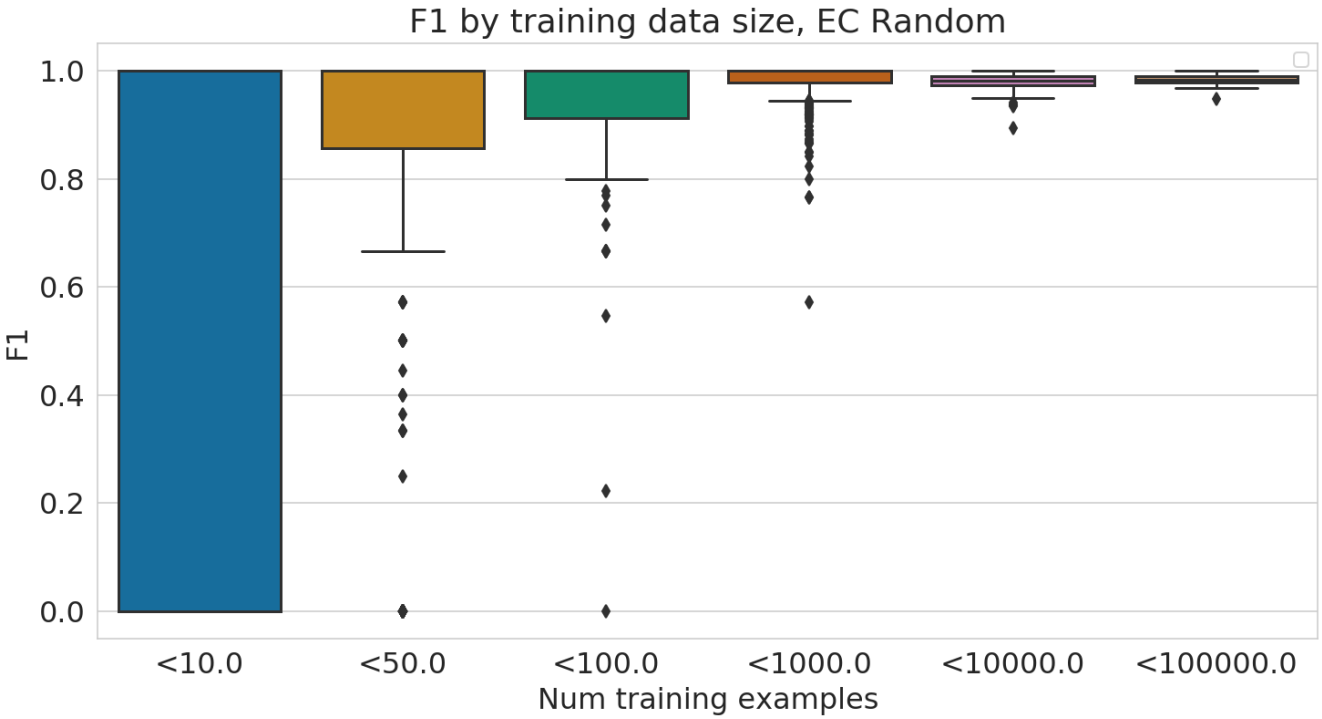


Fig. S11. [Figure 3 - Figure supplement 11] Performance of EC model stratified by number of training examples available for each test example.

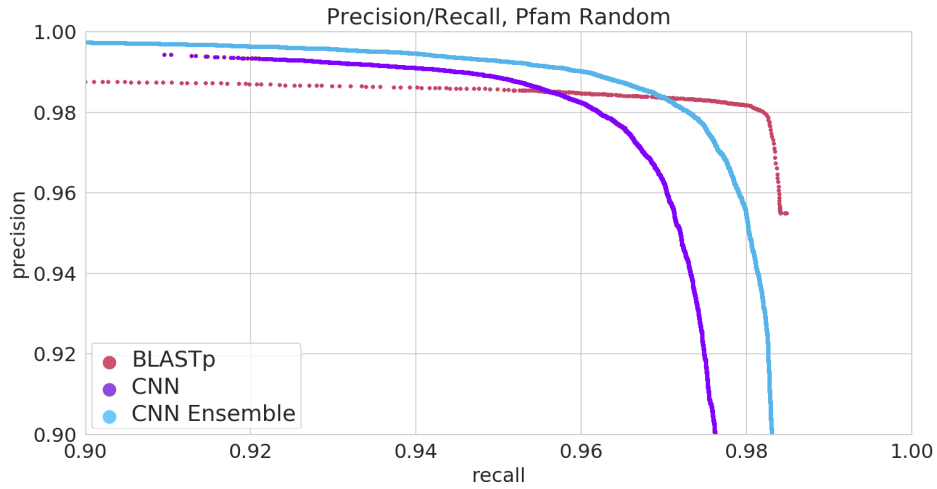


Fig. S12. [Figure 3 - Figure supplement 12] The ProteInfer algorithm is set up to allow any desired training vocabulary to be used. We demonstrated this by additionally training a model for predicting Pfam families from full-length protein sequences, which is available through our CLI-tool, and performs as shown here.