HiFi-NN annotates the microbial dark matter with Enzyme Commission numbers

Gavin Ayres
Basecamp Research Ltd.
London, United Kingdom
gavin@basecamp-research.com

Geraldene Munsamy
Basecamp Research Ltd.
London, United Kingdom

Michael Heinzinger
Department of Informatics, Bioinformatics & Computational Biology
Technical University of Munich
Munich, Germany

Noelia Ferruz
Molecular Biology Institute of Barcelona
Barcelona, Spain

Kevin Yang
Microsoft Research New England
Cambridge, MA

Philipp Lorenz
Basecamp Reseearch Ltd.
London, United Kingdom
phil@basecamp-research.com

Abstract

The accurate computational annotation of protein sequences with enzymatic function, especially those that are part of the functional and taxonomic dark matter, remains a fundamental challenge in bioinformatics. Here, we present HiFi-NN, (Hierarchically-Finetuned Nearest Neighbor search) which annotates protein sequences to the 4th level of EC (enzyme commission) number with greater precision and recall than all existing deep learning methods. HiFi-NN is a hierarchically-finetuned deep learning method based on a combination of semi-supervised representation learning and a nearest neighbours classifier. Furthermore, we show that this method can correctly identify the EC number of a given sequence to identities below 40%, where the current state of the art annotation tool, BLASTp, cannot. We proceed to improve the representations learned by increasing the diversity of the training set, not just in sequence space but also in terms of the environment the sequences have been sampled from. Finally, we use HiFi-NN to annotate a portion of microbial dark matter sequences in the MGnify database.

1 Introduction

Enzymes are efficient catalysts capable of accelerating chemical reactions by several orders of magnitude [1]. They play a crucial role in a myriad of processes within living organisms, encompassing functions from respiration and digestion to facilitating muscle and nerve activity. Sequence databases are experiencing unprecedented growth, providing an increasing number of enzymatic sequences that span a wide range of microbial genomes [11] [12]. While these developments have led to impressive success in training unsupervised models [19][24], a substantial portion of this sequence space remains

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

functionally and taxonomically unannotated and it has been termed the "microbial dark matter" (MDM) [13] [14]. At least one-third of microbial proteins cannot be annotated by aligning them with functionally characterized sequences, and recent studies on the entire AlphaFold database provide evidence that up to 34% of the protein space qualifies as dark matter [25]. Enzymes are exceptionally attractive in biotechnology, catalyzing a wide array of chemical reactions under mild, non-toxic conditions [1]. Given the vast potential within the MDM, it is imperative that we develop innovative methodologies for more accurate and cost-efficient enzyme sequence annotation.

The catalytic function of enzymes is commonly annotated with Enzyme Commission (EC) numbers, categorized mainly into oxidoreductases, transferases, hydrolases, lyases, isomerases, ligases, and transferases [2] [3]. Enzyme commission numbers are an effective proving ground for functional annotation methods. This is because the EC number describes a reaction catalysed by a protein, through convergent evolution different folds can catalyse the same reaction [1] [2] [3] and so annotation methods which rely solely on sequence homology may fail to generalise to novel folds or sequence motifs. Additionally, a given protein may have multiple EC numbers. A total of 8,243 EC numbers have been identified in the BRENDA [3] database. The numbering system is structured in a hierarchical manner, with 7 top level categories. Each level of the hierarchy denotes a more specific type of reaction than the previous. For example, all carbonic anhydrases (EC:4.2.1.1) are hydro-lyases (EC:4.2.1) and that all hydro-lyases are lyases (EC:4).

Several computational methods have been developed to annotate EC numbers from amino-acid sequence alone, such as the sequence-homology-based BLASTp [4], or methods based on curation of protein families or sequence profiles [5], [6], [7], as well as deep-learning methods that were developed more recently: These include DEEPre [8], DeepEC [9], and CLEAN [10]. The latter is considered the current stat-of-the art deep-learning method for predicting EC numbers from sequence. Despite the aforementioned advances in deep-learning based enzyme functional annotation, and language models being built to understand the language of life for the bacterial and archaeal kingdoms, comprehensively annotating the microbial dark matter remains a challenge [15] [16]. To address this challenge, we present HiFi-NN (Hierarchically-Finetuned Nearest Neighbor search). HiFi-NN is based on contrastive learning, which has been applied to various protein-sequence related tasks [10], [17], which we have optimised for annotation of the MDM.

HiFi-NN serves as a method by which a query amino acid sequence can be compared to a set of protein sequence embeddings to find those most similar to each query. To this end, we provide a model that has been trained using contrastive learning to map ESM-2 embeddings [19] to a new feature space where distances between the embeddings of sequences correspond to the similarities of their respective EC numbers. The contributions of our manuscript are three-fold: (1) We develop a method that can correctly identify the EC number of sequences below the twilight zone by incorporating the inherent hierarchy in EC numbers in our contrastive loss, surpassing current methods. (2) We show that the model can improve by increasing the sequence and environmental diversity of the training set from our proprietary database. (3) We annotate a subset of the dark space in the MGnify database [12].

2 Related work and Methods

Related work and methods are outlined in the supplementary information sections A and B, respectively.

3 Results

We train HiFi-NN using contrastive learning, where the training objective is to learn an embedding space of vectors where the Euclidean distance among data points represents the similarities among their functionalities (EC classes) (Fig. 1). In particular, for each sequence in the training set (anchor), we select a positive and a negative sequence, which belong to the same and different EC classes, respectively (Fig. 1a). HiFi-NN is then trained with a triplet loss where distances between positive and anchor sequences are minimized, and between anchor and negative sequences are maximized (Fig. 1a).

We leverage the inherent hierarchy of the EC annotation system as a natural augmentation by sampling positive and negative examples for a given anchor across each of the four levels. The loss function is

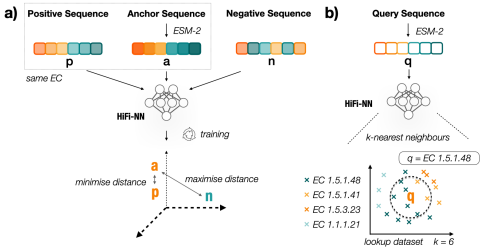

Figure 1: The contrastive learning protocol followed by HiFi-NN during (a) training and (b) inference.

a weighted sum of triplet losses applied to each level of the hierarchy. The selection of positives and negatives for a given anchor is outlined in Table 1. After training, the model assigns the EC label to a query sequence by applying k-nearest neighbor to a pre-embedded lookup table (Fig. 1b).

Table 1: Positive and negative labels for a given anchor

Anchor: 1.1.1.1		
Level	Positive	Negative
1	1.2.4.6	7.1.4.21
2	1.1.3.5	1.3.2.7
3	1.1.1.2	1.1.6.11
4	1.1.1.1	1.1.1.2

We aimed to assess the performance of HiFi-NN in diverse scenarios, and to do so, we assembled different datasets, each with a specific purpose. First, we wanted to compare how our model performs at varying clustering identities to the current gold standard protein annotation tool, BLASTp, which requires several validation sets representing each identity threshold. Second, we prepared a dataset to calibrate our choice of k nearest neighbours and distance thresholds at which our model should refuse to annotate, outlined in supplementary information D and E. Lastly, we further curated a dataset to annotate the MDM. For this we use Swissprot clustered to 30%, with 100 sequences removed as a validation set, and supplemented with sequences from our in-house knowledge graph. In each instance we train a separate model, one for each lookup set used. This ensures that at training time the model does not have access to sequences with the sequence similarity to the training set which we aim to evaluate performance on.

3.1 Performance relative to BLASTp

For the first purpose, we clustered the Swissprot database to sequence identities ranging from 10% to 90% in increments of 10. Our results showed that HiFi-NN outperforms Blast at almost all identity ranges in recall, precision and F1 score (Figure 2a-c), particularly excelling at the low identity range (10-50%).

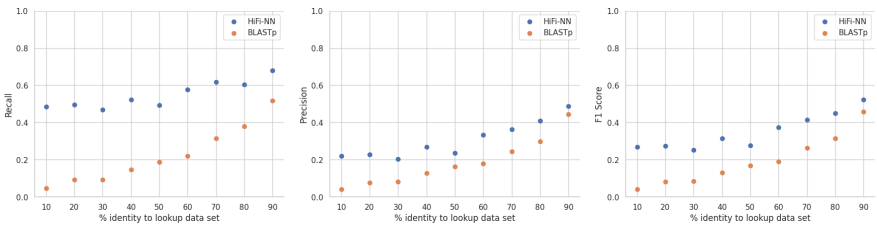

Figure 2: HiFi-NN can annotate proteins to low sequence identity better than homology based approaches.

We expand on the performance comparison between HiFi-NN and BLASTp as well as the make-up of the clustered sequence sets in supplementary information section C.

3.2 Annotating benchmarking datasets outside public databases

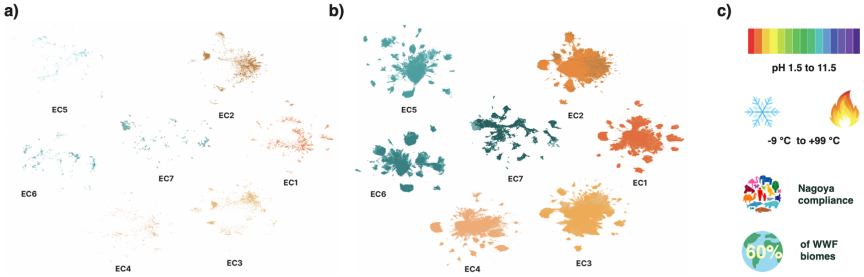

Figure 3: Training sets: Swissprot and sequence supplementation from an in-house metagenomic knowledge graph. a) UMAP of Swissprot sequences with an EC number used for training HiFi-NN. b) UMAP with sequences from a) overlayed with 3 million sequences from an in-house metagenomic knowledge graph. c) Key features of the knowledge graph ensuring diverse sampling origin and Nagoya compliance [18], from which the subset of 3 million sequences shown in b) were derived.

Seeing that HiFi-NN ourperforms BLASTp especially at the low sequence identity range (compared to the lookup data sets), we wanted to test the performance on benchmarking datasets comprised of novel enzyme sequences outside public databases. To that end, we hypothesised that our model would further benefit from supplementing the training dataset with sequences from diverse and under-studied environments (Fig 3a and 3b). For this, we use a proprietary, Nagoya-compliant [18] metagenomic knowledge graph, covering broad pH, temperature, and biome ranges (Fig 3c). A curated subset of 3 million sequences (Fig 3b) from this knowledge graph was added to the training set and a new model was trained on the bigger sequence set. We further expand on the selection and constitution of the training sets used for HiFi-NN models in supplementary information section B.

We compare our method to other state of the art deep learning protein function annotation tools as well as the current most widely used annotation tool, BLASTp, on the Price data set [35]. This data set was introduced for benchmarking the performance of EC number annotation by [28]. It is composed of 149 sequences covering 56 EC numbers. As shown in table 2, HiFi-NN (trained on Swissprot only) outperforms BLASTp and deep-learning annotation methods in recall and F1-score for the task of annotating the Price enzyme dataset [35]. When supplemented with 3 million curated sequences from our in-house database, HiFi-NN outperforms all the aforementioned methods in recall, precision, and F1-score, including HiFi-NN trained on Swissprot only. Crucially, the sequence supplementation to the training set also increases the confidence score of correct annotations (supplementary information section F).

3.3 Using HiFi-NN for annotating the microbial dark matter

With HiFi-NN performing particularly well on annotating the low sequence identify range and microbial benchmarks, we utilise the best performing model on the Price-149 data set to annotate microbial dark matter. For this, we curated a representative set of 2 million amino acid sequences from the microbial database MGnify [12] (supplementary information section G). When we annotate this subset with BLASTp (same parameters as section 3.1), and HiFi-NN, with a more conservative confidence cutoff than in section 3.2 to optimise for higher precision (supplementary information section G). As a result, BLASTp annotates 548,587 sequences and HiFi-NN annotates 1,673,827 sequences.

With no ground truth for these sequences we seek to validate our annotations using the predicted structures for these MGnify sequences available from the ESM Metagenomic Atlas [19]. The

Method	Recall	Precision	F1-score
ECPred	0.0197	0.0197	0.0197
DEEPre	0.0403	0.0415	0.0386
DeepEC	0.0724	0.1184	0.0846
ProteInfer	0.1382	0.2434	0.1662
ProteinVec	0.2961	0.4901	0.3378
BLASTp	0.3750	0.5083	0.3852
DeepECtransformer	0.3026	0.5263	0.3511
CLEAN	0.4671	0.5844	0.4947
HiFi-NN (Swissprot)	0.5724	0.5505	0.5304
HiFi-NN (Swissprot + 3 million curated sequences)	**0.5921**	**0.6657**	**0.6015**

Table 2: Recall, precision and, F1 scores on a data set of enzymes from [35] (referred to as the Price-149 dataset in [10], [49]. Each reported result using HiFi-NN is using the $k = 20$ nearest neighbours. The trade-off between recall and precision is discussed in the supplementary information. As in [10] we report the weighted average of each metric to account for class imbalance. The scores from each competing method are as reported in [10].

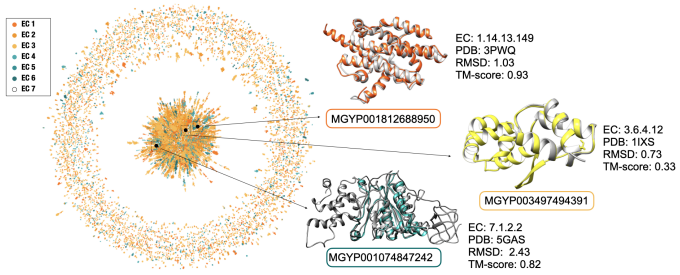

Figure 4: Sequence similarity graph of the MGnify sequences annotated by HiFi-NN but not by BLASTp. We highlight some representative examples which have matches in the PDB.

hypothesis we seek to test is that for at least some of our annotations, are there proteins in the Protein Data Bank (PDB) [44] which have high structural similarity to the predicted structures of the subset of MGnify we chose. Furthermore, we are concerned with which of these have high structural similarity to PDB entries annotated with the same EC number as predicted by HiFi-NN. Of the 2 million MGnify sequences we select 116,385 and their corresponding structures - these sequences both have high confidence HiFi-NN annotations (above 0.7), and no BLASTp annotation (supplementary information section G). We then use Foldseek [46] to calculate the pairwise TM Score [47] between this set of structures and the structures available in the PDB.

Attesting to the 'dark matter' nature of these sequences, only 1,652 of these structures have a minimum TM score of 0.5 to their closest match in the PDB (0.5 being the recommended cutoff suggested in [47]). We then consider an agreement between the HiFi-NN annotations and the structural similarity if, for a given structure from MGnify, of the hits in the PDB with a TM score greater than 0.5 at least one has the same EC as that annotated by HiFi-NN (supplementary information section G). We show examples of structural superimpositions alongside a sequence similarity network of dark matter annotations in (Fig. 4). There are several limitations to validating these annotations *in silico*, for example due to the fact that not all EC numbers are represented in the PDB. Ultimately, these annotations will have to be validated experimentally.

4 Conclusion and Discussion

Here we develop and benchmark the performance of a new deep learning model for enzyme annotation, HiFi-NN (Hierarchically-Finetuned Nearest Neighbor search). Our model is a contrastive deep learning-based method that differs from previous models in 2 key aspects: (1) we use the inherent

hierarchy of the EC annotation system as a natural augmentation method and (2) supplement the training set with microbial sequences sampled from diverse environments. It outperforms the current state-of-the-art method and tool of choice in bioinformatics, BLASTp, as well as all other deep learning tools, when assessed with a microbial enzyme benchmarking dataset [35]. The fact that our model outperforms other models on the task of microbial enzyme annotation, combined with the information that it performs particularly well at the low sequence identity range (compared to the look-up dataset), led us to conclude that HiFi-NN is well-suited for annotating the microbial dark matter [13], [14]. To that end, we have curated a representative subset of the MGnify database [12] and annotated sequences with HiFi-NN. Our model annotates a significant portion of this set.

The annotation outcomes of every bioinformatic tool or deep learning model are, in part, affected by the choice of parameters. The selection of parameters for a given annotation task have to be carefully chosen each time. We are not necessarily able to directly compare the confidence score we established for HiFi-NN (discussed in supplementary section F) and the E-value thresholds used for BLAST [4], or confidence scores used in other methods, such as CLEAN [10]. To that effect, we believe each tool used for annotation has their place and different advantages and disadvantages. We therefore do not claim there is any tool that could overall be considered the "best" method. However, here we have shown evidence of HiFi-NN performing particularly well on microbial, low-identity sequences (compared to the look-up set), which is why we propose its usage for the task of annotating microbial dark matter sequences. Following on from these findings, we are expanding on this work by a) training larger models that are supplemented with tens of millions and hundreds of millions of diverse sequences from our knowledge graph, respectively, b) annotating the entirety of MGnify [12], and c) generating comprehensive wet lab validations for HiFi-NN annotations.

References

[1] Robinson, Peter K. (2015) Enzymes: principles and biotechnological applications. *Essays Biochem.* **59**: 1-41.

[2] Webb, E. C. (1992) Enzyme nomenclature 1992: recommendations of the Nomenclature Committee of the International Union of Biochemistry and molecular biology on the nomenclature and classification of enzymes. *Academic Press.*

[3] Chang, Antje Jeske, Lisa Hofman, Julia Koblitz, Julia Schomburg, Ida Neumann-Schaal, Meina Jahn, Dieter Schomburg, Dietmar (2021) BRENDA, the ELIXIR core data resource in 2021: new developments and updates. *Nucleic Acids Research* **49**(D1): 498-508.

[4] Altschul, Stephen F. Gish, Warren Miller, Webb (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**(3): 403-410

[5] Claudel-Renard, Clotilde Chevalet, Claude Faraut, Thomas Kahn, Daniel (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res.* **31**(22): 6633-6639.

[6] Shen, Hong-Bin Chou Kou-Chen (2007) EzyPred: a top-down approach for predictng enzyme functional classes and subclasses. *Biochem Biophys Res commun.* **364**(1): 53-59.

[7] Yu, Chenggang Zavaljevski, Nela Desai, Valmik Reifman, Jaques (2008) Genome-wide enzyme annotation with precision control: Catalytic families (CatFam) databases. *Proteins* **74**(2): 449-460.

[8] Li, Yu Umarov, Ramzan Xie, Bingqing Fan, Ming Li, Lihua Gao, Xin (2018) DEEPre: sequence-based enyzme EC number prediction by deep learning. *Bioinformatics* **34**(5): 760-769.

[9] Ryu, Jae Y. Kim, Hyun U. Lee, Sang Y. (2019) Deep Learning enables high-qauality and high-throughput prediction of enzyme commission numbers. *PNAS* **116**(28): 13996-14001.

[10] Yu, T. Cui, H. Li, J. et al. (2023) Enzyme function prediction using contrastive learning. *Science* **379**(6639): 1358-1363.

[11] Sayers, E. Bolton, E. Brister, J. et al. (2023) Database resources of the National Center for Biotechnology Information in 2023. *NAR* **51**(D1): D29-D38.

[12] Richardson, L. Allen, B. Baldi, G. et al. (2023) MGnify: the microbiome sequence data analysis resource in 2023. *NAR* **51**(D1): D753-D759.

[13] Rinke, C. Schwientek P. Sczyrba A. et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**(7459): 431-437.

[14] Jiao, J. Liu, L. Hua Z. et al. (2021) Microbial dark matter coming to light: challenges and opportunities. *Nat. Science Review* **8**(3).

[15] Vanni, C. Schechter, M. Acinas, S. et al. (2022) Unifying the known and unknown microbial coding sequence space. *eLife* **11**:e67667.

[16] Hoarfrost, A. Aptekmann, A. Faranuk, G. et al. (2022) Deep learning of a bacterial and archaeal univeral language of life enables transfer learning and illuminates microbial dark matter. *Nature Communications* **13**(1): 1-12.

[17] Heinzinger, M. Littman, M. Sillitoe, I. et al. (2022) Contrastive learning on protein embeddings enlightens midnight zone. *NAR Genomics and Bioinformatics* **4**(2).

[18] United Nations Convention on Biological Diversity. (2011) Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization to the Convention on Biological Diversity, 2010. *Montreal, QC: Secretariat of the Convention on Biological Diversity*.

[19] Lin, Z. Akin, H. Rao, R. et al. (2023) Language models of protein sequences at the scale of evolution enable accurate structure prediction. *Science* **379**(6637): 1123-1130.

[20] The UniProt Consortium (2023) UniProt: the Univeral Protein Knowledgebase in 2023. *Nucleic Acids Research* **51**(D1): D523-D531.

[21] AlQuraishi, M. (2019) ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics* **20**:211.

[22] Steinegger, M. Soeding, J. (2017) MMSeqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35**: 1026-1028.

[23] Olson, D. Dinerstein, E. Wikramanayake, E. et al. (2001) Terrestrial ecoregions of the world: a new map of life on Earth. *Bioscience* **51**(11): 933-938.

[24] Elnaggar, A. Heinzinger, M. Dallago, C et al. (2022) ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell.* **44**(10): 7112-7127.

[25] Durairaj, J. Waterhouse, A. M. Mets, T. et al. (2023). Uncovering new families and folds in the natural protein universe. *Nature*. doi: https://doi.org/10.1038/s41586-023-06622-3.

[26] Radford, A. Hallacy, C. Ramesh, A. et al. (2021) Learning Transferable Visual Models From Natural Language Supervision. *bioarxiv* doi: https://doi.org/10.48550/arXiv.2103.00020.

[27] Schroff, F. Kalenichenko, D. Philbin, J. et al. (2015) Facenet: A unified embedding for face recognition and clustering. *Proceedings of IEEE conference on computer vision and pattern recognition* pp.815-823

[28] Sanderson, T. Bileschi, M. Belanger, D. Colwell, J. (2023) ProteInfer, deep neural networks for protein functional inference. *eLife* 12:e80942

[29] Dalkiran, A. Rifaioglu, A. S. Martin, M. J. Cetin-Atalay, R. Atalay, V. Doğan, T. (2018). ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature. *BMC bioinformatics*, **19**(1), 334.

[30] Gligorijevic, V. Renfrew, P. D. Kosciolek, T. Leman, J. K. Cho, K. Vatanen, T. Berenberg, D. Taylor, B. Fisk, I. M. Xavier, R. J. Knight, R. Bonneau, R. (2019) Structure-Based Function Prediction using Graph Convolutional Networks. *bioRxiv*.

[31] Johnson, J. Douze, M. Jégou, H. (2019) Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, **7** (3) pp.535-547

[32] Littmann, M. Heinzinger, M. Dallago, C. et al. (2021) Embeddings from deep learning transfer GO annotations beyond homology. *Sci Rep* 11, 1160.

[33] Loshchilov, I. Hutter, F. (2019) Decoupled Weight Decay Regularization. *arXiv* doi: https://doi.org/10.48550/arXiv.1711.05101.

[34] Zhou, M. Niu, Z. Wang, L. Gao, Z. Zhang, Q. Hua, G. (2019) Ladder Loss for Coherent Visual-Semantic Embedding. *arXiv* doi: https://doi.org/10.48550/arXiv.1911.07528.

[35] Price, M.N. Wetmore, K.M. Waters, R.J. et al. (2018) Mutant phenotypes for thousands of bacterial genes of unknown function. *Nature*, Vol 557(7706), pp. 503-509.

[36] Chopra, S. Hadsell, R. LeCun, Y. (2005) Learning a similarity metric discriminatively, with application to face verification. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, pp. 539-546 vol. 1, doi: 10.1109/CVPR.2005.202.

[37] Zhang, ML. Zhou, ZH (2005) A k-nearest neighbor based algorithm for multi-label classification *2005 IEEE International Conference on Granular Computing, Beijing, China* pp. 718-721 Vol. 2, doi: 10.1109/GRC.2005.1547385.

[38] Frosst, N. Papernot, N. Hinton, G. (2019) Analyzing and Improving Representations with the Soft Nearest Neighbor Loss. *arXiv* doi: https://doi.org/10.48550/arXiv.1902.01889.

[39] Goldberger, J. Hinton, G. E. Roweis, S. Salakhutdinov, R. R. (2004) Neighbourhood Components Analysis. *Advances in Neural Information Processing Systems*, **17**.

[40] Buchfink, B Reuter, K Drost, HG (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods* **18**, 366–368. doi:10.1038/s41592-021-01101-x.

[41] Papadopoulos, H. (2014) A Cross-Conformal Predictor for Multi-label Classification. *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, 241–250.

[42] Cauchois, M. Gupta, S. Duchi, J. C. (2021) Knowing what You Know: valid and validated confidence sets in multiclass and multilabel prediction. *Journal of Machine Learning Research*, **22** (81) pp. 1-42.

[43] Ioffe, S. Szegedy, C. (2015) Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *ICML'15: Proceedings of the 32nd International Conference on International Conference on Machine Learning* **37** pp.448–456.

[44] Berman, H.M. Westbrook, J. Feng, Z. et al. (2000) The Protein Data Bank. *Nucleic Acids Research*, **28**: 235-242.

[45] Shen, W. Le, S. Li, Y. Hu, F. (2016) SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE*, **11**(10): e0163962.

[46] Kempen, M.v. Kim, S.S. Tumescheit C. et al. (2023) Fast and accurate structure search with Foldseek. *Nature Biotechnology*, doi: https://doi.org/10.1038/s41587-023-01773-0.

[47] Zhang, Y. Skolnick, J. (2005) TM-align: A protein structure alignment algorithm based on TM-score. *Nucleic Acids Research*, **33**: 2302-2309.

[48] Aman Memon, S. Aamir Khan, K. Naveed, H. (2020) HECNet: a hierarchical approach to enzyme function classification using a Siamese Triplet Network, *Bioinformatics, Volume 36, Issue 17, pp. 4583–4589.*

[49] Kim, G.B. Kim, J.Y. Lee, J.A. et al. Functional annotation of enzyme-encoding genes using deep learning with transformer layers. Nat Commun 14, 7370 (2023).

Supplementary information

A Related work

A.1 Deep metric learning

Contrastive learning Contrastive learning involves comparing examples to each other and imposing a loss such that similar examples will be close in feature space and dissimilar examples far away. It has proven an effective tool for representation learning and has lead to state of the art performance on several image classification and text classification benchmarks as well as providing a means of aligning the feature spaces of data from multiple modalities [26].

Triplet loss The triplet loss was originally proposed in [27] as a solution to the problem of facial recognition. The aim is to learn a feature space $f_\theta(x)$ in which embeddings of the same face (the anchor) under different poses (positives) are closer to each other than those of different faces (negatives). Formally, given a set of training examples X and a function $f_\theta(x) \in \mathbb{R}^D$ we aim to minimise:

$$L(\theta) = \sum_{i=1}^{N} \left[||f_\theta(x_i^a) - f_\theta(x_i^p)||_2^2 - ||f_\theta(x_i^a) - f_\theta(x_i^n)||_2^2 + \gamma \right]_+, \tag{1}$$

where γ denotes a margin term and the superscripts a, p and, n denote an anchor, positive and negative respectively. Typically f_θ is a Siamese neural network [36] with the same weights used to embed each instance involved in the triplet.

A.2 EC number classification

Preliminary work The task of assigning multiple labels to a test instance can be tackled in many different ways. DeepFRI [30] uses a neural network architecture with a softmax output layer and is trained in a supervised manner. The authors attempt to alleviate the issues that come with an imbalanced data set by using a weighted binary cross entropy loss function. Similarly, ProteInfer [28] is trained in a supervised manner and therefore faces a similar problem. The authors of CLEAN [10] showed that ProteInfer failed to maintain predictive power for underrepresented classes. DeepECtransformer [49] also treats the task of EC annotation as a supervised learning problem, however they include an unsupervised pre-training step in their method. The class imbalance is addressed by the use of a focal loss function. ECPred [29] opts for an alternative method using an ensemble of classifiers, one for each 4th level EC number. However, coverage only extends to 858 EC classes, a problem not faced by the current most widely used annotation tool, BLASTp [4]. HECNet [48]incorporates the inherent hierarchy in the EC labeling system by including a hierarchical triplet loss as an inital training loss upstream of a feed forward neural network with a softmax loss function. Our method is similar in the use of a triplet loss, however we benefit from the use of ESM-2 as the pre-trained sequence encoder. The latest deep learning tool to attempt functional annotation of protein sequences, [10], pairs an optimised feature space with a nearest neighbours classifier. This is a step towards deep learning based functional annotation which can handle class imbalance and the authors showed that for certain data sets it significantly outperforms BLASTp. Building on the work of [10] we use a nearest neighbours classifier on an optimised feature space to annotate protein sequences. Our method differs from [10] in how we represent the classes. Rather than a set of class prototypes (embeddings representing a single EC number) we use each example in the training set as an example of their associated class. This has a few practical benefits; it is straightforward to incorporate new

Machine Learning for Structural Biology Workshop, NeurIPS 2023.

EC numbers into our representation, a practitioner can choose to trade off precision and recall at inference time by varying the choice of k, and we can incorporate label density information into our confidence estimates. The practical drawbacks of our approach, having to store the entirety of the training set, are largely alleviated by the use of approximate nearest neighbour algorithms, e.g. FAISS [31]

k-NN The method we propose is similar in spirit to BLASTp and has been shown to be effective for annotating GO terms [32] and CATH annotations [17]. We transform a 'lookup' dataset to our optimised feature space and then perform a nearest neighbour search against this lookup dataset. The annotations of the k nearest neighbours are then transferred to the query protein. Importantly, we only need to transform the training set once. Then at inference time we embed the query sequences and perform a k-NN lookup against the already embedded training set.

B Methods

B.1 Training set construction

All clustering was performed using the tool MMSeqs2 [22] using iterative profile search with the highest sensitivity setting (7.5) for identities below 50%. At each iteration we removed a set of clusters and added only the representatives of these clusters to the test set. We ensure the same EC labels across training and validation datasets by removing sequences in the validation set for which there is no sequence in the training set which has the same EC number.

B.1.1 Supplementing Swissprot

We hypothesized that for the purpose of annotating functional / microbial dark matter, contrastive deep learning models would benefit from inclusion of highly diverse sequences into the training set - not just in sequence space, but in contextual and environmental diversity, too. Since public sequence databases lack consistent metadata collection or labelling such as biome or temperature information, we supplemented the Swissprot-based training set with a subset of sequences derived from a proprietary metagenomic graph database optimised for diverse representation across these parameters. Sequences derived from this database were collected from 5 continents, spanning 60 percent of WWF biomes [23], a pH range from 1.5 to 11.5, and a 108 °C temperature range. Crucially, all sampling efforts were conducted with biodiversity stakeholder consent and engagement as well as landowner permission, following the access and benefit sharing guidelines & regulation for digital sequence informaiton (DSI) as layed out in the Nagoya Protocol [18].

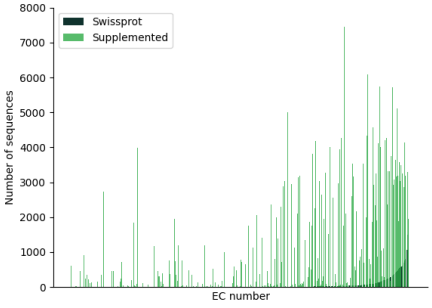

Figure 1: Supplementing Swissprot with 3 million sequences from our in-house knowledge graph. We add sequences to include representation across EC numbers for which Swissprot has few examples.

B.2 Contrastive learning

Hierarchical triplet loss We apply the triplet loss across each level of the hierarchy, a variation of the ladder loss introduced in [34]. We optimise a weighted combination of these loss terms. The

Table 1: Positive and negative labels for a given anchor

Anchor: 1.1.1.1		
Level	Positive	Negative
1	1.2.4.6	7.1.4.21
2	1.1.3.5	1.3.2.7
3	1.1.1.2	1.1.6.11
4	1.1.1.1	1.1.1.2

weighting scheme was chosen to reflect the order of similarity present in the hierarchy, i.e. root nodes are more dissimilar than leaf nodes. The loss is as follows;

$$L = \sum_{i=1}^{|\mathcal{B}|} \sum_{j=1}^{4} w_j \left[||f_\theta(x_i^a) - f_\theta(x_i^{p_j})||_2^2 - ||f_\theta(x_i^a) - f_\theta(x_i^{n_j})||_2^2 + \gamma_j \right]_+, \tag{2}$$

where $w_j \in (1, 0.9, 0.8, 0.7)$ are the weights for each level of the hierarchy, $|\mathcal{B}|$ denotes the size of a batch $\mathcal{B}$ and the choice of margins followed the same scheme as the loss weights, $\gamma_j \in (1, 0.9, 0.8, 0.7)$. Positive examples for protein i at level j of the hierarchy are denoted by $x_i^{p_j}$ and, likewise, negative examples are denoted as $x_i^{n_j}$. Again, f_θ represents a Siamese neural network with which we embed the anchor, positive and negative. The architecture we use is a two-layer multi-layer perceptron (MLP) with a hidden dimension of 1024 and output dimension of 512. We use layer normalisation in the penultimate layer of the model and the GELU activation function.

Triplet formation The authors of [10] use the fourth level of EC number as supervision for constructing sets of positives and negatives given an anchor. However, the inherent hierarchy within the EC numbering system provides a natural augmentation method. Each anchor can be compared to a positive and negative at each of the 4 levels of the hierarchy. Our anchors are the instances in our mini-batch. For each loss term, L_j, we sample uniformly at random the EC number of the positive example, provided it matches the EC number of the anchor up to level j, and the negative EC number, provided it matches up to level $j-1$ and differs at level j. Once we have chosen the EC number of the positive and negative example, we then sample uniformly at random from the set of instances which belong to these labels.

B.3 Training details

Each Swissprot model is trained on an NVIDIA A10 GPU until convergence on a held out validation set. The model trained on Swissprot clustered to 30% identity is trained for 3,180 epochs. Both the models trained on Swissprot clustered to 50% and 90%, with clusters iteratively removed in increments of 10% identity thresholds, are trained for 2000 epochs. The model we train with the addition of 3 million sequences is trained on 8 NVIDIA A100 GPU's for 448 epochs. Notably, contrary to recent models which have used contrastive learning for functional annotation of protein sequences we do not use any hard negative mining during training. We believe this illustrates the effectiveness of the multiple levels of comparison between protein sequences in our loss function. It is likely that improvements to this model could be made by incorporating negative mining whilst maintaining the hierarchical comparisons in the triplets. Each model is trained with a batch size of 8 using the ADAMW optimiser [33] with a learning rate of 0.0003 and default parameters otherwise. The learning rate is decayed according to a cosine annealing schedule to a minimum of 0.000001.

C Performance relative to BLASTp

We compare our method to the current annotation tool widely used for protein function annotation. We run the DIAMOND BLAST [40] version of the tool with default parameters and an e-value cutoff of 1e-3 with the highest sensitivity setting, *–ultra-sensitive*, so we achieve the best possible performance at low percent identities.

C.1 Extending into the midnight zone

To serve as a comparison to BLASTp and to illustrate the utility of protein language model embeddings in moving past homology based annotation at low sequence identities we created a data set following the procedure used for the ProteinNet [21] data set. We cluster Swissprot to sequence identities ranging from 10% to 90% in increments of 10, removing a set of clusters at each iteration and adding only the representatives of these clusters to the validation set. We choose clusters that have a minimum of 5 sequences to avoid validation sets which are composed entirely of protein fragments (typically very short proteins which have no homologs in the rest of the data set) or mis-annotated sequences. The total number of clusters we remove from the training set at each clustering is 300, we then take one sequence from each cluster and add it to the validation set representing the chosen sequence identity. This gives us 9 validation sets with 300 sequences each, each sequence corresponding to an entire cluster. The remaining set of sequences comprises our training set. After clustering we found that there were 830 EC numbers across all our validation sets which did not exist in the training set. It would be impossible for any method to correctly annotate these EC terms and so we remove sequences corresponding to these EC numbers from our validation sets. This ensures we preserve the sequence similarity thresholds we desire. We have a total of 163,632 sequences in our training set. The resulting composition of our validation sets is outlined in table 2 and the performance of each of these sets outlined in figure 2. We use the training set as our lookup set for annotation. We report the sample average of each metric due to the differing test set sizes.

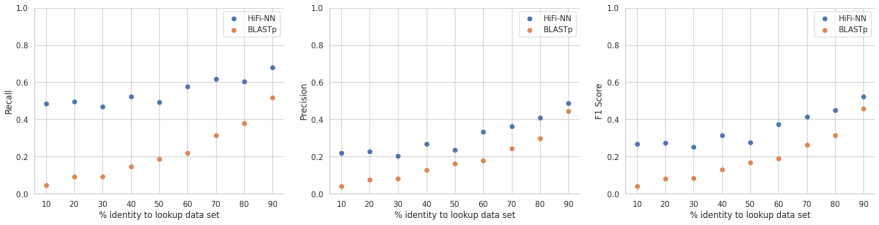

Figure 2: HiFi-NN can annotate proteins to low sequence identity better than homology based approaches.

Table 2: Size of test set for each clustering identity threshold with the number of EC numbers these sets cover.

% Identity Clustering	# Sequences in test set	# EC numbers
10	218	129
20	216	127
30	207	133
40	200	138
50	195	157
60	202	141
70	197	144
80	197	143
90	212	167

C.2 50% sequence identity

The second training set we construct was designed to study the effect of a larger training set and a larger test set. We clustered Swissprot to 50% identity and removed 3000 cluster representatives as the test set. As outlined in the previous paragraph, we then ensure that each EC number which exists in the test set has at least a single representative in the training set. This gives us a final test set size of 1,977 sequences covering 644 EC numbers and a training set size of 166,404 sequences covering 2,808 EC numbers ($\approx 49\%$ of the total number within Swissprot). The performance of HiFi-NN with these data set splits is outlined in figure 3. As before, we report the sample average of each metric.

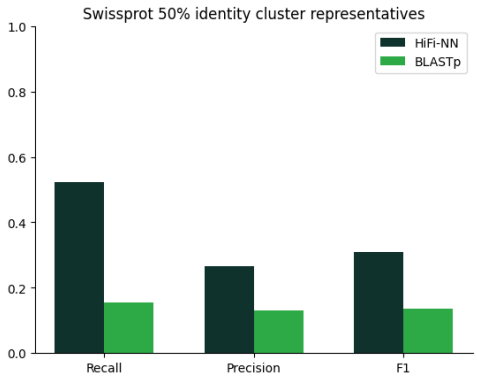

Figure 3: Comparison of HiFi-NN and BLASTp on Swissprot clustered to 50% identity.

C.3 Time based split

To test how HiFi-NN performs in the context of new sequences arriving to a database we decided to construct a test set from a time based split of Swissprot. Specifically, we construct a set of sequences which have been added to Swissprot since the 31st of July, 2023. At the time of writing this is a set of 244 sequences spanning 135 EC numbers. As we can see in figure 4, there is little advantage gained from the addition of a diverse set of metagenomic sequences. The reason for this may be due to the sequence diversity of the test set, our method particularly excelling at low sequence identities. In addition, BLASTp performs quite well on this data set. We use the same benchmarking set up as before where we BLAST against the HiFi-NN training set to ensure a fair comparison.

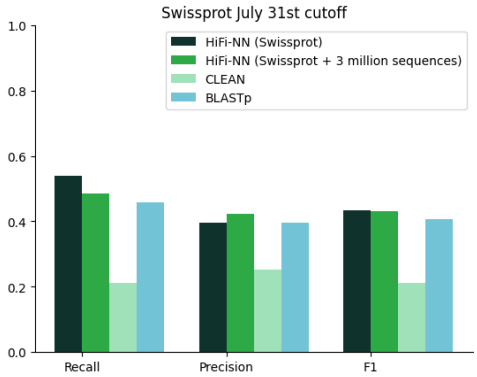

Figure 4: Performance of each method on a time based split of Swissprot. The addition of diverse metagenomic sequences seems to endow no discernible advantage to the model for this data set.

D Choice of k nearest neighbours

The effect of varying the k nearest neighbours on performance on the 50% identity cluster representatives data set is outlined in figure 5. It is worth noting that this is a choice made at *inference*. A practitioner can trade off precision and recall by varying the number of nearest neighbours they wish to retrieve. As a consequence, we provide a confidence score threshold for improving precision and recall for a fixed choice of k. The confidence score threshold takes into account both distances to labels and the density of a label amongst the k nearest neighbours. As such, it becomes more useful as k is increased, for $k = 1$ it will trivially assign a confidence of 1.0 to all annotations. For lower values of k a distance based cutoff is recommended.

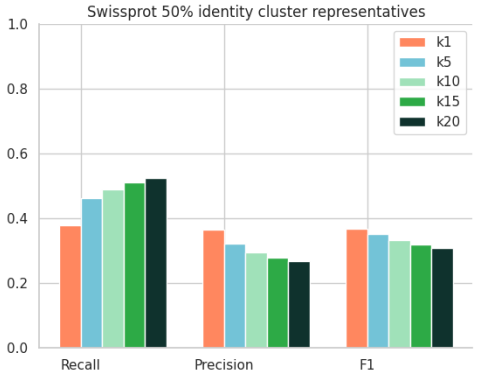

Figure 5: Performance of model as a function of the k nearest neighbours retrieved.

E When not to annotate

To establish a threshold at which we refuse to annotate a protein sequence we use the test set derived from Swissprot clustered to 50% and add sequences from Swissprot which have not been annotated with an EC number and have an annotation score of 5 (a measure of the reliability of the annotations associated with a protein, on a scale of 1-5). The results are illustrated in figure 6. The lookup set used for this study is the training set derived from the 50% clustering. We take the 95th percentile of the distances to the test set which has an EC, a distance of 1066.

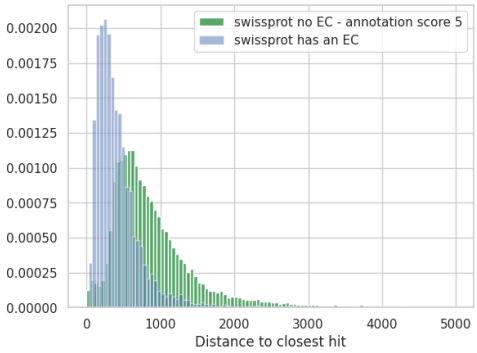

Figure 6: Distance to closest hit in Swissprot clustered to 50% and annotated with an EC number. We see that false positives are almost unavoidable, however the tradeoff between false positives and negatives can be controlled by using a distance threshold on predictions.

F Confidence scores for multi-label k-nearest neighbours

Why we need confidence scores The practical utility of an annotation tool necessitates a reliable confidence score. To this effect there are two broad categories of approaches which we may pursue, confidence scores based on distances from a query to an example, with associated labels, and confidence scores based on the density of labels in the neighbours [37].

Related work The relative distances between queries and neighbours has been optimised as part of the training process and so distance based confidence measures are a natural candidate. Conformal prediction provides a framework for computing such confidence measures through the use of non conformity scores calibrated to a validation set. These non conformity scores have been extended to the case of multi-label classification by considering the non conformity of the entire set of labels

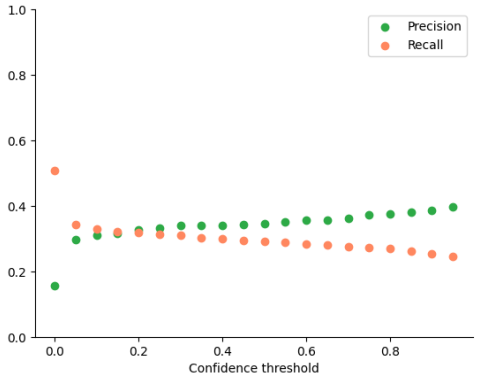

Figure 7: Precision and recall as a function of the confidence threshold.

predicted [41] [42] for a given test instance relative to the power set of the labels. However, in the context of EC number annotation the powerset of all labels is simply too large to have practical application. [10] makes use of statistical properties of the distances between and within EC numbers in order to calculate its confidence measure. This involves fitting a Gaussian Mixture Model to the distribution of within class distances and between class distances.

Our method The approach we opt for uses information from both the density of the labels amongst k nearest neighbours as well as their associated distances. Specifically, we extend probability density estimates for local neighbourhoods [38][39] to the multi-label setting. The setup is as follows; we have a set of real valued vectors, $x_1, ..., x_n \in \mathbb{R}^D$ and a finite set of labels $\mathcal{Y}$. The aim is to learn a classifier which maps from the input space X to the powerset of the labels, assigning scores in proportion to the relevance of a given label to the test data point. For an instance, x, and its associated labelset $Y \subseteq 2^{\mathcal{Y}}$ we will denote the neighbours of x using $N(x)$ and a distance vector for each label l as

$$\vec{y}_{x,t}(l) = \begin{cases} d(t,x), & \text{if } l \in Y \\ 0, & \text{otherwise} \end{cases}$$

where $d(t,x)$ is a distance metric between two vectors $t, x \in \mathbb{R}^D$. We then define the probability of an instance t having label l as a softmax over the distances between t and each instance in the neighbourhood of t with label l,

$$p_{t,l} = \frac{\sum_{x \in N(t)} e^{-\vec{y}_{x,t}(l)/T}}{\sum_{x \in N(t)} e^{-d(t,x)/T}}, \tag{3}$$

where T is a scaling parameter which controls the relative influence of nearby points. We set $T = 100$ for all experiments.

Confidence on the Price data set In figure 7, we see further validation that increasing the sequence diversity of the training set improves the method. The median confidence score for all correct predictions changes from 0.17 to 0.65. The model becomes more confident in its correct predictions, this is likely due to the fact that our confidence measure is based on the Euclidean distance to the nearest neighbour's and the relative density of a label within the k nearest neighbours. Increasing the number of sequences per EC class and the diversity of such sequences should help on both fronts.

G Annotating the microbial dark matter

Selection of sequences to annotate To isolate microbial dark matter sequences, we employed a multi-tiered approach leveraging sequence data from MGnify [12] which was clustered at a 90% identity threshold. This was followed by a further clustering step at 30% identity. Subsequently, we applied a sequence length filter, retaining only those sequences with lengths ranging from 100 to 600 base pairs. To generate a representative subset, we utilized SeqKit [45] to randomly sample 2 million

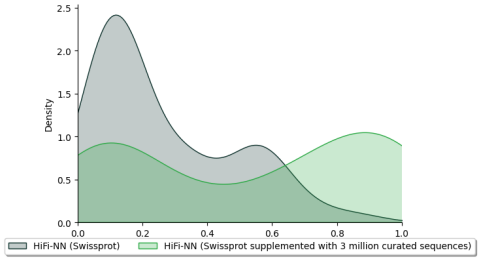

Figure 8: Density of confidence scores for correctly predicted EC numbers in the Price data set.

sequences, using a random seed value of 42 to ensure reproducibility. These selected sequences were then subjected to comprehensive functional annotation via the HiFi-NN pipeline.

Selecting sequences for structural validation We only annotate sequences which have a Euclidean distance to their nearest neighbour of less than 1066 (as discussed in section E) and we use a confidence threshold of 0.1 to avoid spurious annotations. For comparison, CLEAN [10] does not use any threshold and so for the benchmarking performed in section 3.2, we report recall, precision, and F1 score for CLEAN [10], HiFi-NN and others models without thresholds. We however decided to implement a threshold of 0.1 for this task to ensure higher precision and avoid over-annotation (see Figure 6 in this supplementary discussion). The number of sequences which satisfy this criterion is 1,673,827. By comparison, of the two million sequences, BLAST annotates 548,587.

Aware that a portion of these may have been mis-annotated we then curate a subset of these annotations with even higher precision, with which to compare the structures of. To this effect, we take the annotations which have a confidence score greater than 0.7 and only those sequences which have not been annotated by BLASTp. We define an 'agreement' in the following way: If HiFi-NN annotates sequence X with EC 1.1.1.1 and the matching structures in the PDB have associated EC numbers 1.2.3.4 and 1.1.1.1, we say that there is agreement up to the 4th level of EC number between HiFi-NN and the structure search. The resulting number of 'agreements' per level are as follows; 607 have no match between the EC numbers, 1045 agree to the first level of the EC hierarchy, 691 to the second, 465 to the third and 165 to the fourth. There are caveats to this approach, we applied no filter to the pLDTT (a measure of confidence) of the predicted structures and there are EC numbers which have no representation in the PDB and therefore would produce no matches.