NAR Genomics and Bioinformatics, 2024, **6**, lqae021
https://doi.org/10.1093/nargab/lqae021
Advance access publication date: 27 February 2024
Methods Article

Detecting anomalous proteins using deep representations

Tomer Michael-Pitschaze[1,†], Niv Cohen[1,†], Dan Ofer[2], Yedid Hoshen[1] and Michal Linial[2,*]

[1]The Rachel and Selim Benin School of Computer Science and Engineering, The Hebrew University of Jerusalem, Jerusalem, Israel
[2]Department of Biological Chemistry, Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

[*]To whom correspondence should be addressed. Tel: +972 54 8820035; Email: michall@cc.huji.ac.il
[†]The first two authors should be regarded as Joint First Authors.

Abstract

Many advances in biomedicine can be attributed to identifying unusual proteins and genes. Many of these proteins' unique properties were discovered by manual inspection, which is becoming infeasible at the scale of modern protein datasets. Here, we propose to tackle this challenge using anomaly detection methods that automatically identify unexpected properties. We adopt a state-of-the-art anomaly detection paradigm from computer vision, to highlight unusual proteins. We generate meaningful representations without labeled inputs, using pretrained deep neural network models. We apply these protein language models (pLM) to detect anomalies in function, phylogenetic families, and segmentation tasks. We compute protein anomaly scores to highlight human prion-like proteins, distinguish viral proteins from their host proteome, and mark non-classical ion/metal binding proteins and enzymes. Other tasks concern segmentation of protein sequences into folded and unstructured regions. We provide candidates for rare functionality (e.g. prion proteins). Additionally, we show the anomaly score is useful in 3D folding-related segmentation. Our novel method shows improved performance over strong baselines and has objectively high performance across a variety of tasks. We conclude that the combination of pLM and anomaly detection techniques is a valid method for discovering a range of global and local protein characteristics.

Introduction

The unprecedented growth in quality and quantity of molecular data (e.g. genomes, transcriptomes) in recent years raises the need for a systematic approach for functional annotation of new protein sequences (1). The consistent increase in the success of automatic functional prediction was mostly attributed to the expansion in the number and variety of solved structures and the overall increase in databases (2). The Gene Ontology (GO) framework is used as a gold standard for the assignment of biochemical function, biological process, and cellular localization terms to (3). Protein function is intrinsically complex and poorly defined. It is mostly indirectly studied through evolutionary conservation (e.g. protein homologous families and 3D structure (4). However, proteins carry numerous cellular functions that are context dependent. Examples are protein-protein interactions, cell signaling, and the regulatory network. For the genomic-based collection of UniProtKB/TrEMBL (230 M sequences, Release 4/2022), <1% have experimental evidence, and the majority of the database includes predicted proteins with no supporting evidence (5) With such constraints, direct experiments are the best way to determine high accuracy in functional prediction.

The assignment of functions to sequences is a challenging task. Even with known homologs, inheritance transitivity is a source of inaccuracy and ambiguity in multi-domain protein functions (6). Annotation efforts usually fail when function is rare and represented by orphan sequences (7). Occasionally, protein functions that were not previously observed are reported, which suggests that unique and unexpected functions exist and that methodologies for their systematic findings are needed. Examples are short peptides in humans that resemble snail cone toxins (8), heat resistant *hero* proteins (9) prion proteins that drive pathological aggregations (10), intrinsically disordered proteins that involve phase separation (11) and more. Distinguishing alterations in protein function due to mistakes in translation (12), and developing effective screening methods to identify novel functions is of ultimate importance to the field of protein design and engineering (13).

Deep learning approaches have led to protein fold prediction with extremely high accuracy, as implemented in AlphaFold. It also allowed populating the unstructured space of protein sequences with high confidence structures (14), with 36% of all amino acids in the human proteome predicted with high confidence (15). However, in instances where only a few homologs exist, or there is low divergence, AlphaFold predictions are of lower quality. Importantly, mapping a fold to its function is not always evident, as the same fold may account for a large number of unrelated functions (16). In addition, training models from sequence (17) using NLP methodologies (18,19), led to breakthroughs in protein function inference (e.g. ProteinBERT, ProtTrans, ESM) and more (20–23).

Most deep learning approaches, both for proteins and other data modalities, heavily rely on manually annotated samples. However, often the most exciting research tasks require discovering new phenomena rather than distinguishing between known data classes. Breaking from previous research, we explore a new setting for identifying novel, previously unknown protein types. As such protein types are unknown and unexpected, we do not assume that any annotated examples of such proteins are provided to us. To detect such novel samples automatically, we must rely on the ability to distinguish between samples similar to the training data ('*normal data*')

Received: March 14, 2023. **Revised:** November 17, 2023. **Editorial Decision:** January 29, 2024. **Accepted:** February 23, 2024

and novel data types (‘*anomalies*’). We refer to this setting here as *anomaly* detection, but the terms *‘novelty*‘ and *‘outlier*‘ are often used interchangeably in the literature (24). In the case of anomaly detection in tabular data, it was shown that *density estimation* of samples as vectors is a strong approach (25). Namely, a sample is deemed anomalous if its features are far away from any normal samples, so it is likely not to have come from the same distribution. With other data modalities, a key step for anomaly detection is to map sparse samples to a relatively dense embedding space (24). After mapping each sample to a descriptive vector, a density estimation approach can be applied to various data modalities, including images (26), time-series (27), tabular healthcare data (28,29).

In this study, we aim to establish a computational method for anomaly detection in proteins at a genomic scale. The methods we adapt were originally developed and successfully applied in the domain of computer vision. We adapt the pretrained approach of SPADE, by extracting residue level features from a protein embedding network and using them to detect residue level anomalies with density estimation. For the whole sequence anomaly detection case, we borrow an image-level approach (similar to the approaches presented in DN2 and PANDA). However, in order to detect full-length protein anomalies (i.e. whether the whole protein is an enzyme or not) with such approaches we have to summarize the residue level embeddings to representations of entire proteins. We describe our technical approach for this summarization in Methods, in section ‘*Whole sequence anomaly detection*’.

We tune the method to seek novelty within different subgroups of the protein sequence space. For this goal, we introduce protein function annotation terms from Gene Ontology (GO) and UniProtKB keywords as ground truth. We also discuss functions that describe structural segments within protein sequences. We rely on the notion that proteins with the same 3D fold might share only minimal sequence similarity (30,31). We present a method that highlights novel functions and discuss the predictive power needed to identify novel functions within the protein sequence database.

Materials and methods

We present a method for detecting anomalous proteins. Crucially, our method does not assume that we are able to characterize the anomalous proteins, as unusual and interesting proteins are often unexpected. We adapted recent breakthroughs in image anomaly detection (e.g. DN2 (32), SPADE (33), PANDA (26)) and applied them to the amino acid sequence of proteins. Our method tackles different categories in protein function detection: full-length proteins (i.e. whether the whole protein is an enzyme or not) and at the local residue level, denoted as *protein anomaly segmentation*. Our approach for anomaly detection consists of two stages: deep protein feature extraction and anomaly scoring. In section ‘Whole sequence anomaly detection’, we describe our embedding method for full-length protein sequences.

Feature extraction

Anomaly detection methods require powerful representations of the data. We desire representations that reflect the biological semantic similarity between proteins. While proteins are encoded as amino acid sequences, which are simply represented as a sequence of characters, this does not explicitly hold information about their structure, roles and interactions within the protein. Many approaches have been developed for protein representation in the field of protein structure, protein-protein interactions (PPI) and function in general (23,34).

In this work, we choose to represent protein sequences by deep neural network (NN) encoders. The encoders are first pretrained on huge protein datasets, to solve a bidirectional language modeling task (20,35). Specifically, the model is trained on a subsequence of amino acids in the protein, where some have been replaced with masked out tokens and its task are to predict these ‘hidden’ tokens. This paradigm has two main advantages: (i) it can exploit a massive, unlabeled protein dataset; (ii) a success on this task implicitly requires a high-level understanding of the protein, its function and structure. We assume that such powerful encoders have already been trained and provided to us. This assumption is realistic and based on the use of protein language models (pLM) that are readily available and were found highly effective for different downstream tasks. For example, ESM, ProtTrans, ProteinBERT (20,21,23). Following common practice, we use the penultimate layer of the pretrained encoders as our representation. We denote the activation map for protein P as $\psi(P)$.The model provides a vector representation for each amino acid. We denote the representation of the i-th amino acid as $F_i = \psi(P)_i$.

Anomaly scoring

We use a density-based anomaly scoring rule. The motivation is that normal proteins are common, and we often find multiple examples for each normal protein pattern. Conversely, anomalous proteins are anticipated to be rare, especially in model organisms that have been extensively studied. We will not expect to find many examples of each anomalous protein pattern. We therefore use k nearest neighbors (kNN) to measure the anomaly score of a particular protein or residue. Consider, for example, the representation of residue i in protein P,$\Psi(P)_i$. We compute its distance from each residue embedding in all proteins in the training set. The anomaly score of our target protein residue $s(P,i)$ is computed as the average of the distance to the k nearest residues. Let us denote the k nearest residues from the normal train data of our target residue $\Psi(P)_i$ as $N_k(\Psi(P)_i)$:

$$s\left(P,\, i\right) \;=\; \frac{1}{k}\Sigma_{\psi\in N_k(\Psi(P)_i)}\left|\left|\psi \;-\; \Psi(P)_i\right|\right|^2$$

Whole sequence anomaly detection

Proteins can be anomalous either due to an anomalous local region, or due to a protein-wide anomalous property. To detect anomalies in entire proteins, one might consider using the anomaly score of the most anomalous residue within the entire protein. Although this method sometimes achieves strong results, it often fails. The reason is that residue embeddings are relatively local, while anomalous properties might be global. We propose whole-protein anomaly scoring, which considers both local and global patterns. One simple way to account for all the protein positions is to represent an entire protein as the mean of each of its residue embeddings. The mean embedding is similar to average pooling, a standard way of summarizing a sequence of features. To detect anomalies using the mean embedding, we look for proteins whose mean embedding is far from the mean embedding of any protein in the normal train set. This is computed using the Euclidean distance to its

K nearest neighbors. Such proteins found in low density areas are likely to come from a different distribution than normal ones.

Using a single mean embedding for the entire protein may not provide a sufficient description of the local variations within the protein. An alternative way to summarize features, which can work better, is to use set features to represent the protein as a set of its segments. We adapt the method of Tzachor and Hoshen (36) in using set features to represent an entire sequence (Supplementary Text S1). Differently from their work, we operate on deep representations rather than on the raw data, as they are far more informative for proteins. We score a protein as normal or anomalous using nearest neighbors, with the Euclidean distance between the set features. We denote our whole-protein score (by either method) for protein *P*as $s(P)$.

Implementation details

We include here the key implementation details. Further details on the implementation our methodology and robustness to impurities in the training sets are in Supplementary Text S1.

The code is available in https://github.com/Tomer-Michael/prada.

The databased used: https://github.com/Tomer-Michael/prada/tree/main/compressed_datasets.

Training set subsampling and cleaning. To improve efficiency and mitigate the potential presence of anomalies in the training set, we subsample the training set. In order to select the samples that best represent the normal data, our method picks the most *typical* samples. Concretely, we randomly selected n (we used $n = 50\,000$) samples to be used as evaluators. For each of the remaining examples, we retrieve the K evaluators that are nearest to it and average the distance to them. Finally, we select the m samples (we used $m = 50\ 000$) that have the smallest K evaluator distance. For residue-level samples, nearby regions in the same protein are often embedded with very similar features. We ensure that evaluators are selected for a diverse set of proteins and locations. When fewer than n + m training samples are provided, we use random subsampling.

Preprocessing. Following ESM, all sequences were trimmed to the first 1022 amino acids.

Architecture. We used the ESM-1b (esm1b_t33_650M_UR50S) network, with weights provided by the authors of ESM. We used the 'esm1b_t33_650M_UR50S' model from https://github.com/facebookresearch/esm u.

Set features for protein sequences. We follow the method by (37). We use a window duration of 49 residues, 30 temporal dilations, 200 projections and 100-bin histograms. We did not use whitening.

The exact dataset used in this study can be found under our project repository: https://github.com/Tomer-Michael/prada/tree/main/compressed_datasets.

Results

Detecting contaminating viral proteins in human proteome

We evaluate the effectiveness of anomaly detection methods for discovering unknown protein types. In our experimental protocol, we provide training and test sets. The training set consists of normal proteins only, and the test set consists of normal and anomalous proteins. We train our method based only on the normal training data, and use it to compute anomaly scores for each test set protein. Our evaluation includes multiple anomaly types according to the different datasets used.

We tested the ability of the anomaly detection method to identify viral sequence contamination, specifically by human-infecting viruses. The task of identifying pathogenic viral sequences with respect to their hosts is of clinical relevance. We tested the ability of the method to detect viral proteins with respect to the host human proteome. A dataset was compiled from the curated SwissProt database (38). For this task, we further filtered the viruses to keep only those that are associated with humans as hosts. Out of 35K proteins, 27K remained following filtering by the host. See details in Supplementary Text S1 and Supplementary Table S1.

The accuracy of an anomaly detector is dependent on the desired tradeoff between false positive (FP) and false negative (FN) detections. In order to specify the desired tradeoff, one often chooses a threshold such that all samples with higher anomaly scores are considered anomalous. However, methods can have inconsistent ranking depending on the choice of threshold. For threshold-independent evaluation, most anomaly detection papers use the ROC–AUC metric which averages the true-positive rate for all possible false-negative rates (determined by different choices of threshold). An interpretation of this metric is that given a random normal sample x_{norm} and an anomalous sample x_{anom} (both from the test set) the ROC-AUC is equal to the probability that $s(x_{anom}) > s(x_{norm})$.

Figure 1. ROC-AUC performance in identifying contaminated viral proteins for the task of separating viral proteins from the host proteome. We only considered humans as viral hosts (and removed cases of broader virus-host tropism). Set and mean embedding methods achieved the best results.

As further evidence for the efficacy of our approach, we computed the proportion of true anomalies out of the top M% proteins with the highest anomaly scores. Our anomaly score allowed us to prioritize candidate proteins with a far higher probability of being anomalous than randomly sampling the test set. Table 1 summarizes the performance across multiple tasks. It is apparent that highly scored proteins have a far higher likelihood of being anomalous than randomly sampled ones.

Detecting unknown protein types

We evaluate different datasets that cover diverse attributes of proteins, including biochemical functions, cellular localization and protein interactions. We attempt to identify anomalies among these different datasets, considering one protein type as our normal class (seen during training), and any protein not in that class as the anomalous class. In each case, we aim to detect anomalous samples among the normal test data: enzymes (by their E.C. numbering system)—we consider enzymes as anomalies and non-enzymes as normal; extracellular/intracellular—we considered secreted (i.e. extracellular) proteins as anomalous and non-secreted as normal; ion/metal binding—proteins that bind ion/metal are considered anomalous and any other considered normal multiple interactions—proteins that interact with multiple other proteins we considered as anomalous and the rest as normal. The

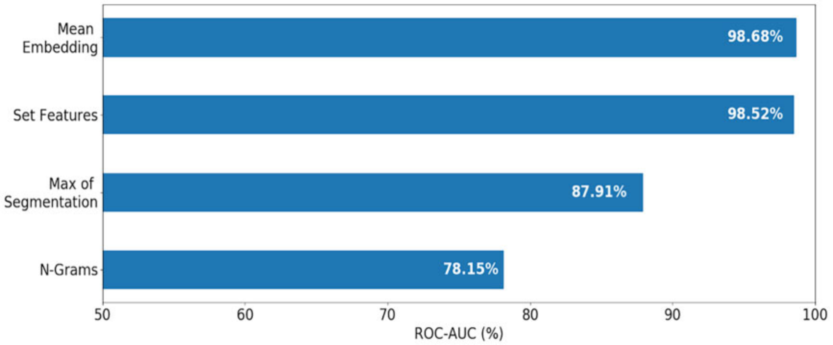

Figure 1. Anomaly detection accuracy for identifying human-specific viral proteins (% ROC–AUC).

Table 1. Frequency of anomalous samples for samples with high anomaly scores (% of the entire set)

Set features	Human/virus	Enzyme	Extracellular/intracellular	Ion/metal binding	Multiple interactions
Full dataset	75.00	21.55	17.33	20.34	45.81
Test set	85.76	35.46	29.54	33.80	62.83
Top 20%	99.85	78.01	50.67	68.23	82.07
Top 10%	99.70	80.10	56.00	76.64	81.85
Top 5%	99.57	80.21	60.68	83.60	79.75
Top 1%	99.57	82.20	69.62	91.35	72.44

Table 2. Evaluation of anomaly detection methodologies

Dataset	Source	N-Gram	Max of segmentation	Mean embedding	Set features
Human/virus	SwissProt	78.15	87.91	98.68	98.52
Enzyme (E.C.)	SwissProt	67.78	74.33	84.54	**85.32**
Extracellular/intracellular	SwissProt	60.87	57.61	**74.91**	70.59
Ion/metal binding	SwissProt	66.37	66.58	78.69	**79.93**
Multiple interactions	SwissProt	65.63	69.88	75.71	**76.26**
Prions	UniRef90	78.05	29.23	80.59	**85.76**

fraction of the minority group in the database varies drastically from 0.01%for prion proteins to 36.8% for task of extracellular/intracellular. For full details on the prediction model and the database used, see Supplementary Text S1. Recall that a property such as being an ion binding protein is based on the spatial arrangement of a small cluster of amino acids (i.e. local), while being involved in protein–protein interaction has a more global context.

We compare our method against two baseline anomaly detection approaches. Table 2 shows the results on these tasks. The ground truth for all the listed tasks is derived from SwissProt (unless stated otherwise). The *N-Gram* approach (39) relies on the raw amino-acid (encoded as one-hot vectors) counts rather than on deep representations of protein segments. As anticipated, it performs significantly worse, emphasizing the benefit of including contextualized embeddings from the pLMs (18). While the *Max of Segmentation* approach utilizes deep features, it underperforms the other approaches. The *mean embedding* method, which averages the features of protein segments, and the *set embedding* method, which uses set features, are the top performers on all datasets.

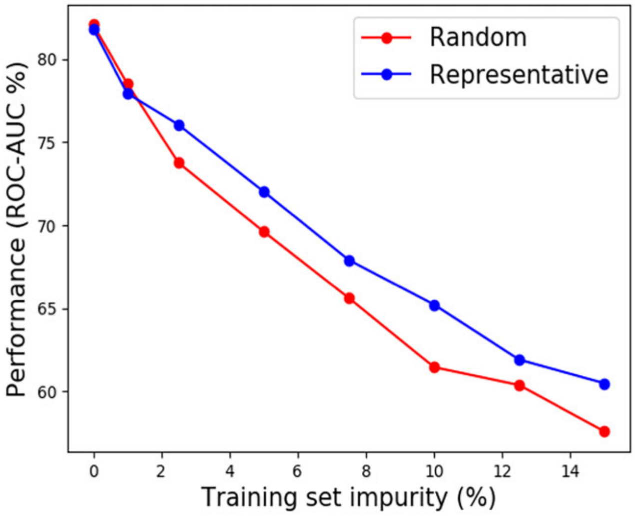

Figure 2. Anomaly detection accuracy for different impurity levels in the training set of the enzyme dataset (see details in Supplemental Text S1). While very large impurity rates naturally reduce performance, our representative subsampling method improves performance over random subsampling.

Robustness to training set impurity

As anomalies are hard to detect, the training set may include anomalous samples. We test the robustness of our method against the existence of (unknown) anomalies in the training set ('impurity'). Figure 2 summarizes the impact of different levels of impurity in the training by examining the enzyme dataset. First, we find that with a relatively small amount of data in the training dataset (up to 1% contamination), our results are not significantly degraded. When there is a larger degree of impurity in the training dataset (on the order of 10%), the degradation is more severe, but we are able to mitigate it using the selection of a representative subset described in Methods.

Identifying novel prion-like proteins

To test the ability of our method to detect anomalies due to a global functional property, we tested our method on discovering prion-like characteristics and achieved 85.76% accuracy (ROC–AUC, *set embedding*). The prion-anomaly detection dataset includes all non-fragment proteins from the UniRef90 database. We used UniRef90 proteins (i.e. representatives with <90% identity) in two manually annotated classes: known prion proteins that are labeled with the SwissProt molecular function keyword of Prion (KW-0640) versus all other (remaining) proteins. We focus on prions as these are rare with poorly defined biochemical properties. Prions undergo anomalous shifts in their 3D structure, which eventually leads to irreversible aggregation and physiopathology *in vivo* (40).

Our unsupervised model was used to create a scored, ranked list of 60 000 prion predictions on a diverse sample of UniRef90 proteins. The top 100 most likely predictions are listed in Supplementary Table S2). We observed that the proteins with the highest anomaly scores are relatively short (mean 205 amino acids; length is <80 amino acids for 50% of the top list). The proteins are signified as having non-standard taxonomy sampling with over-representation of proteins from fungi and slime mold (14% each), bacteria and viruses (12% each). These are poorly studied proteins, with over a third of them being named ‘uncharacterized’. Many of these top scoring proteins have compositional bias (28%).

Our results support the notion that prion-like proteins have low sequence similarity to other proteins, low sequence complexity, and low confidence structure (Figure 3). Most of the structures predicted as anomalous proteins are of low confidence (pLDDT orange/yellow color), and only very small fragments reach high confidence (dark blue). In many of these proteins, over-representation of specific amino acids is evident. For example, Q54VH6 and Q54QL5 from the *Dictyostelium discoideum* (Social amoeba) are composed of over 70% asparagine (N). Similarly, in Q54KT2, histidine (H) and proline (P) dominate the sequence. Extreme bias in the occurrence of amino acids signifies many of the identified prion-like proteins (41).

Correspondences between the anomaly score and structural-based segmentation

We further tested the ability of anomaly detection methods to assign local function and protein segmentation. To this end, we focused on 8035 proteins and extracted residue-level anomaly scores for them. We analyzed a number of such cases and illustrate our findings for two representative proteins with respect to structural predictions by AlphaFold2 (14). Both proteins are characterized by long, unstructured segments. The UniProt Q96DN6 protein (1033 amino acids) is encoded by the gene MBD6 (Methyl-CpG-binding domain protein 6; Figure 4A). It binds to heterochromatin indirectly (without interacting with methylated or unmethylated DNA). In addition, it is recruited to sites of induced DNA damage and potentially acts in chromatin organization. While a detailed knowledge of its 3D is unavailable, AlphaFold2 predicts that the minimal positional error (dark green, Figure 4A, middle) is limited to <100 amino acids at its N’-terminal domain. This region serves as an anchor site, with the rest of the 3D structure being of low confidence and a very large alignment positional error (Figure 4A, bottom).

The second example of protein Q5VUJ9 (Dynein regulatory complex protein 8) has a similar trend. This protein (269 amino acids) regulates ciliary motility and the microtubule sliding in motile axonemes. The second half of the protein also acts as an anchor site for the extended low-confidence unstructured region (Figure 4B). As shown in Figure 4B, the unfolded segment that includes the first 200 amino-acids is poorly determined by AlphaFold2. Note that these unstructured long segments match the very low anomaly score. We conclude that the anomaly score detects non-classical proteins with large segments of unstructured regions, where high score highlights anchor regions (Figure 4A and B, bottom). The anomaly score is seemingly less sensitive to domain boundaries. Although there is no direct information on the 3D structure or evolutionary conservation for many of the proteins marked as anomalous, the notion of folded region and non-structural regions is rediscovered by our methodology. Additional examples of the local anomaly score profiles are shown in Supplementary Figure S1.

Discussion

Studying the source of anomalous proteins is relevant for understanding the source of functional novelty (42). Identifying novel functions among the curated UniProtKB/SwissProt is a challenging task. Often it is restricted to proteins characterized by an accelerated evolutionary rate (i.e. under positive selection), enriched with mutations (i.e. polymorphic hotspot regions), or originating from a rapidly evolving phylogenetic lineage. Function annotations often rely on having homologues in model organisms. However, genuine anomalous proteins include orphan proteins, which are poorly annotated and over-represented in less studied organisms (43). While the function of most proteins is uniquely defined, moonlighting proteins have multiple (often unrelated) functions. For example, the SMN’s (survival motor neuron) main role is in the biogenesis of small nuclear ribonucleoproteins (snRNPs). However, when expressed in axonal projections, it acts to control local translation (44). Proteins may also alter their function according to their oligomerization state. The beta amyloid, which is the hallmark of Alzheimer’s disease, exhibits neurotoxicity when oligomerized. However, as a monomer, it actually acts to quench metal-inducible oxygen radicals, thereby inhibiting neurotoxicity (45). Many of these unexpected examples were identified sporadically. We present a systematic method that can be used for identifying new candidates that are anomalous on a genomic scale.

We frame this setting as anomaly detection and modify existing methods developed in the image anomaly detection community to detect anomalies in proteins. We note, that the term anomaly detection is often used interchangeably with novelty, outlier, or out-of-distribution detection. There are sometimes semantic variations between these terms, but the literature is typically inconsistent about them. In study we use ‘anomaly detection’ to highlight its relation to the one-class-classification setting in images.

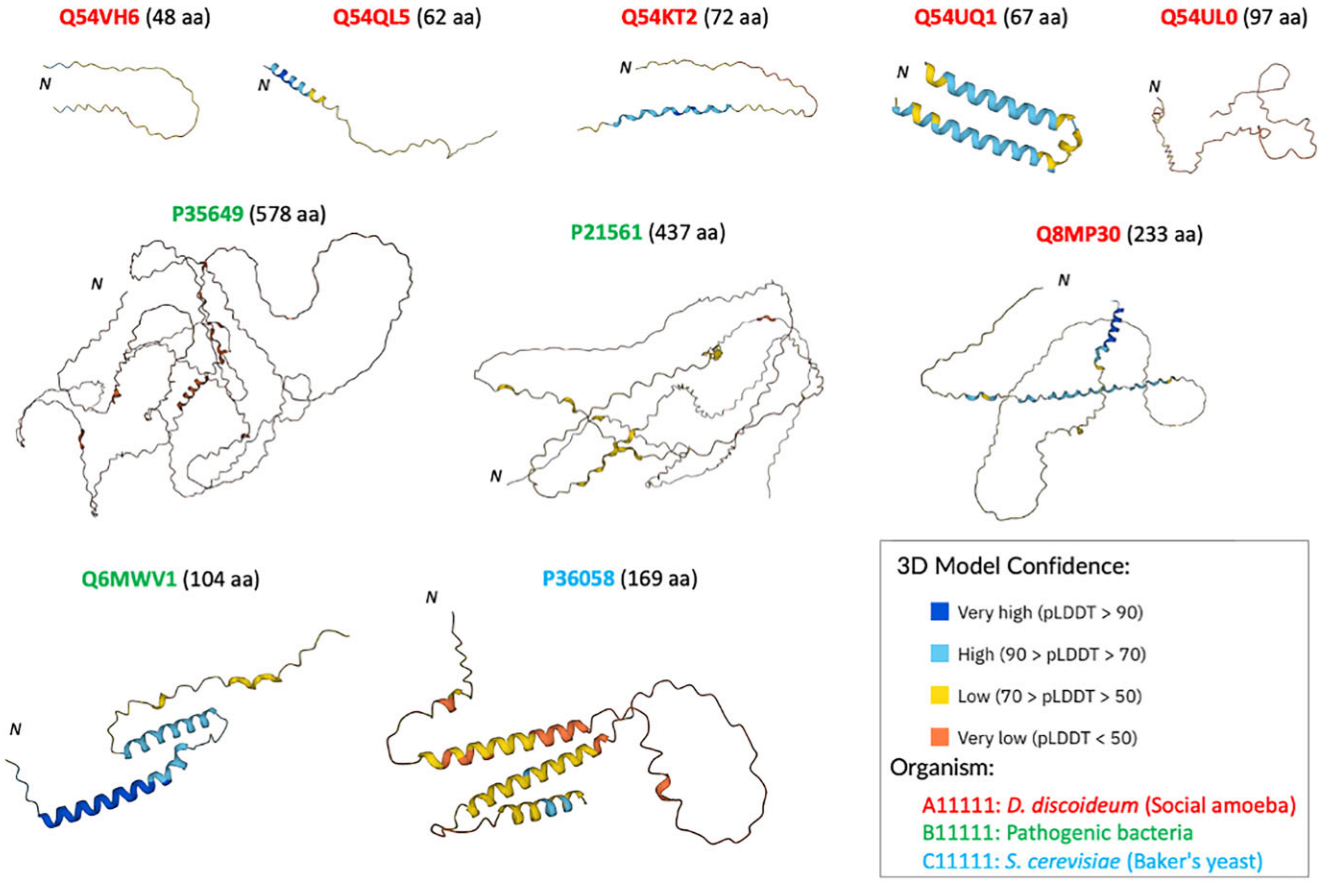

Figure 3. The 3D structures of the 10 proteins having the highest anomaly scores for the prion-like task. The AlphaFold2 structure predictions are shown and colored by the pLDDT confidence score. The organism/taxonomical groups are indicated by the font color of the protein ID. Note that only small portions within these representative sequences are colored by high confidence (dark blue, pLDDT > 90). Short proteins from the *Dictyostelium discoideum* dominate the anomalous prion-like sequences. N, indicates the N′-terminal of the protein.

Although our technical approach follows the leading paradigm for anomaly detection with deep features, future research may provide further improvements to our anomaly detection results. First, further improvement on deep protein embedding methods will directly improve the quality of our results. This is especially true, if we have some prior knowledge regarding some biological properties that may indicate anomalies. In this study, we applied a single approach to many different tasks (Table 2). However, it is possible that different embeddings may highlight different aspects of the proteins that will be useful for particular tasks. For example, proteins associated with functions that involve RNA or DNA must be quite long, often unstructured, and have abundant basic residues that are common in nucleic acid interactions from bacteria to humans. Incorporating such domain-specific knowledge can significantly improve performance.

There are other promising areas for improvement, particularly: (i) summarizing the residue representation to a whole sequence representation (ii) using the final representation for scoring. For example, other technical approaches for whole protein representation may include graph-based representations, i.e. building a graph for each protein based on the internal similarities of the protein residues. Having such a graph for each of the seen (normal) proteins, one could compare them to the graphs of test proteins; to detect protein abnormalities. To distinguish between the graphs of normal and anomalous test proteins, we may use anomaly detection methods for graphs (46). The scoring function we use for our representation may leave room for further exploration as well. We use the kNN density estimation as a strong and robust baseline, but other approaches, including density-estimation-based and others, could be explored as well (24,47).

Another possible direction for future improvements is using additional knowledge to adapt the used pre-trained representation to better describe normal variation in the data and avoid false positive (FP) detection. Such knowledge may consist of statistical assumptions regarding the distribution of the normal data that can be used to adapt the representation. This was done in the cases of previous methods: DeepSVDD (48), PANDA (49), Mahalanobis (50) and OOD-no-labels (29). Prior knowledge may also come in the form of additional auxiliary features (18) or labels. While still not assuming labeled anomalies, the normal samples may have semantic labels. For example, class labels for protein function in the normal data may allow us to better adapt the representation to detect novel, unlabeled protein functions (51,52). Another option is that the user may label some attributes they wish to ignore. E.g., wishing not to detect known protein types in unseen organisms as anomalies. In such cases, we may provide organism labels for the data in order to make our representation agnostic to the source organism (53). A limitation in our search for anomalies stems from the generic protocol

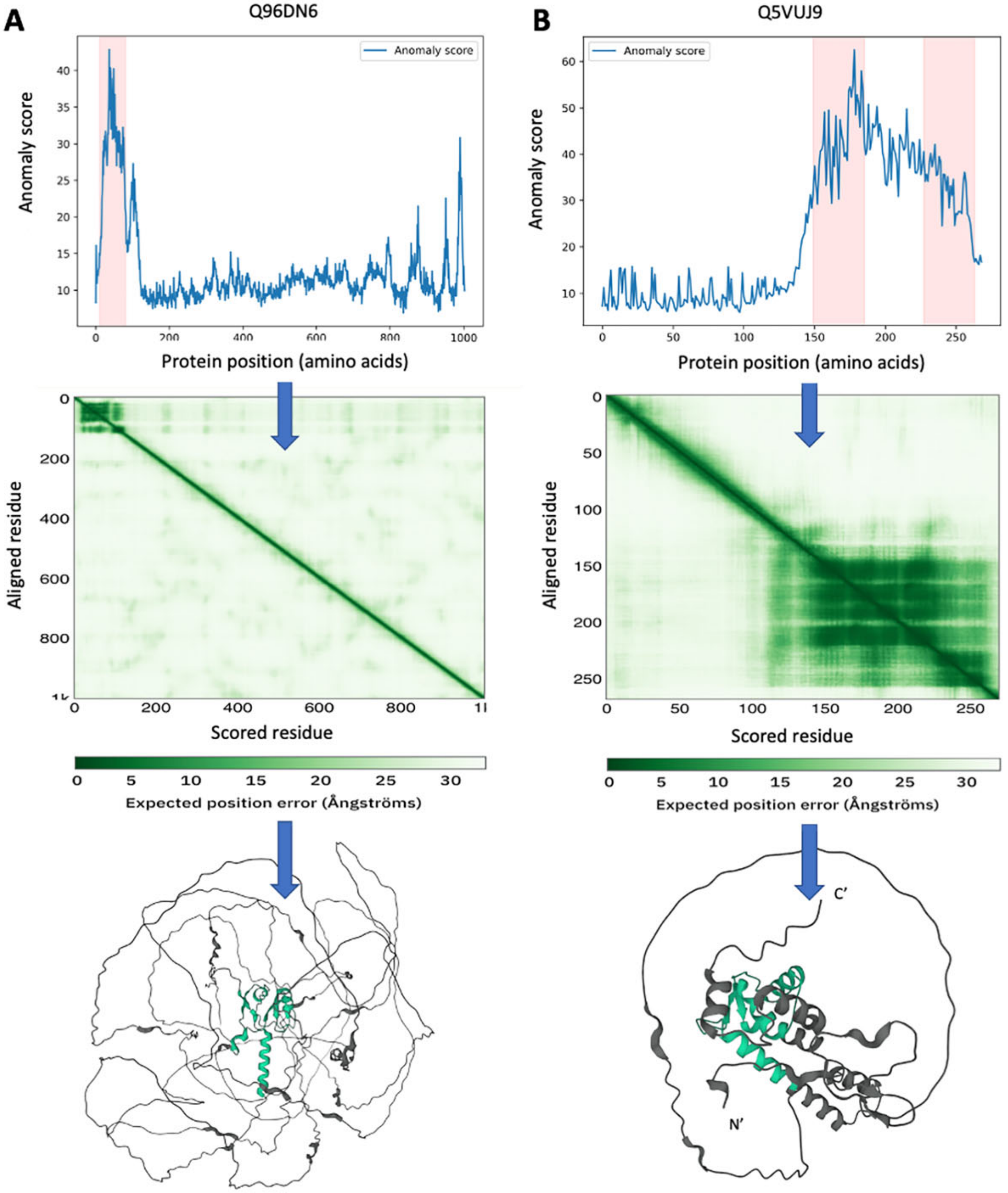

Figure 4. Residue-based anomaly detection protocol on represented proteins. (**A**) UniProtKB Q96DN6 (methyl-CpG-binding domain protein 6, gene symbol MDB6), and (**B**) Q5VUJ9 (Dynein regulatory complex protein 8; gene symbol EFCAB2). Top: The plots show the anomaly score along the position of the sequence (x-axis). Note that the y-axis for the anomaly score is not identical in A and B. Pfam domains are colored pink. Middle: Predicted Aligned Error (PAE) plots**.** Specifically, the green color indicates expected distance error in Å. The color corresponds to the expected distance error in residue x's position when the prediction and true structure are aligned on residue y. Bottom: AlphaFold2 predictions colored by the region with minimal PAE. Note that the protein regions with maximal values of residue-based anomaly overlap the regions with minimal PAE values.

for protein annotations. For example, the automatic pipeline of genome annotation often overlooks short proteins (54,55). Revisiting these short sequences revealed toxin-like function, novel antibiotics and unexpected immunological cell recognition proteins (8,54).

We tested the ability of the anomaly detection method to identify viral sequence contamination (Figure 1). From the results of this task, we were able to draw the following conclusions: (i) human–host viruses were more likely to be detected as anomalies than viruses with a more general tropism. Viral proteins that were mistakenly classified as human proteins overlap with cases of mimicry (56). (ii) Latent viruses such as the herpes virus were misclassified as anomalies. Notably, latent viruses provide a real difficulty to the immune recognition system, where the separation between self to non-self is blurred. (iii) Retrovirus sequences are of viral origin, which along evolution became endogenous to the human genome. These are often misclassified as anomalies. Retroviral-like proteins are evolutionary remnants, and once integrated, their duplication is identical to any other human gene (57).

In this study, we cover a wide range of protein functionalities, some of which are extremely rare (prions), while other functions are far more common (enzymes). Prion identification reached a high success rate. It may reflect the lack of contamination in the training, due to prion proteins' rarity. Prions are of great interest from structural and medical perspectives. They act as infectious agents with devastating outcomes. From a biochemical point of view, the pathogenic protein may form non-reversible aggregates that lead to a chain reaction that infects benign prion proteins. Prion propagation is a common concept shared between mammals and fungi but has been poorly studied in other organisms (58). Prion proteins may tilt the balance to accelerate the 'infectious' potential (59). It was debated whether the infectivity capacity of prions is a true anomaly to our biological understanding (60). Considering the unprecedented speed of determining protein sequences from poorly studied genomes, the unsupervised anomaly detection is an attractive approach for identifying functional novelty within the protein sequence database.

Data availability

The data underlying this article are available in the article and in its online supplementary material or will be shared on reasonable request to the corresponding author. The code and data are available on Zenodo at https://doi.org/10.1101/2023.04.03.535457.

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

Center for Interdisciplinary Data Science Research (CIDR) [3035000440].

Conflict of interest statement

None declared.

References

1. Friedberg,I. (2006) Automated protein function prediction—the genomic challenge. *Briefings Bioinf.*, **7**, 225–242.
2. Radivojac,P., Clark,W.T., Oron,T.R., Schnoes,A.M., Wittkop,T., Sokolov,A., Graim,K., Funk,C., Verspoor,K. and Ben-Hur,A. (2013) A large-scale evaluation of computational protein function prediction. *Nat. Methods*, **10**, 221–227.
3. Martin,D., Berriman,M. and Barton,G.J. (2004) GOtcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinf.*, **5**, 178.
4. Ofran,Y., Punta,M., Schneider,R. and Rost,B. (2005) Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discov. Today*, **10**, 1475–1482.
5. Ouzounis,C.A. and Karp,P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol.*, **3**, COMMENT2001.
6. Green,M. and Karp,P. (2005) Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res.*, **33**, 4035–4039.
7. Tautz,D. and Domazet-Lošo,T. (2011) The evolutionary origin of orphan genes. *Nat. Rev. Genet.*, **12**, 692–702.
8. Kaplan,N., Morpurgo,N. and Linial,M. (2007) Novel families of toxin-like peptides in insects and mammals: a computational approach. *J. Mol. Biol.*, **369**, 553–566.
9. Tsuboyama,K., Osaki,T., Matsuura-Suzuki,E., Kozuka-Hata,H., Okada,Y., Oyama,M., Ikeuchi,Y., Iwasaki,S. and Tomari,Y. (2020) A widespread family of heat-resistant obscure (Hero) proteins protect against protein instability and aggregation. *PLoS Biol.*, **18**, e3000632.
10. Halfmann,R., Alberti,S. and Lindquist,S. (2010) Prions, protein homeostasis, and phenotypic diversity. *Trends Cell Biol.*, **20**, 125–133.
11. Uversky,V.N. and Dunker,A.K. (2010) Understanding protein non-folding. *Biochim. Biophys. Acta (BBA) Proteins Proteomics*, **1804**, 1231–1264.
12. Drummond,D.A. and Wilke,C.O. (2009) The evolutionary consequences of erroneous protein synthesis. *Nat. Rev. Genet.*, **10**, 715–724.
13. Ufarte,L., Potocki-Veronese,G. and Laville,E. (2015) Discovery of new protein families and functions: new challenges in functional metagenomics for biotechnologies and microbial ecology. *Front. Microbiol.*, **6**, 563.
14. Varadi,M., Anyango,S., Deshpande,M., Nair,S., Natassia,C., Yordanova,G., Yuan,D., Stroe,O., Wood,G. and Laydon,A. (2022) AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.*, **50**, D439–D444.
15. Tunyasuvunakool,K., Adler,J., Wu,Z., Green,T., Zielinski,M., Žídek,A., Bridgland,A., Cowie,A., Meyer,C. and Laydon,A. (2021) Highly accurate protein structure prediction for the human proteome. *Nature*, **596**, 590–596.
16. Orengo,C.A., Jones,D.T. and Thornton,J.M. (1994) Protein superfamilles and domain superfolds. *Nature*, **372**, 631–634.
17. Wan,C. and Jones,D.T. (2020) Protein function prediction is improved by creating synthetic feature samples with generative adversarial networks. *Nat. Mach. Intell.*, **2**, 540–550.
18. Ofer,D., Brandes,N. and Linial,M. (2021) The language of proteins: NLP, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.*, **19**, 1750–1758.
19. Khurana,D., Koli,A., Khatter,K. and Singh,S. (2022) Natural language processing: state of the art, current trends and challenges. *Multimed. Tools Appl.*, **82**, 3713–3744.
20. Rives,A., Meier,J., Sercu,T., Goyal,S., Lin,Z., Liu,J., Guo,D., Ott,M., Zitnick,C.L. and Ma,J. (2021) Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Natl. Acad. Sci. U.S.A.*, **118**, e2016239118.
21. Elnaggar,A., Heinzinger,M., Dallago,C., Rehawi,G., Wang,Y., Jones,L., Gibbs,T., Feher,T., Angerer,C., Steinegger,M., *et al.* (2022) ProtTrans: toward understanding the language of life through self-supervised learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**, 7112–7127.
22. Chowdhury,R., Bouatta,N., Biswas,S., Floristean,C., Kharkar,A., Roy,K., Rochereau,C., Ahdritz,G., Zhang,J. and Church,G.M. (2022) Single-sequence protein structure prediction using a language model and deep learning. *Nat. Biotechnol.*, **40**, 1617–1623.
23. Brandes,N., Ofer,D., Peleg,Y., Rappoport,N. and Linial,M. (2022) ProteinBERT: a universal deep-learning model of protein sequence and function. *Bioinformatics*, **38**, 2102–2110.
24. Ruff,L., Kauffmann,J.R., Vandermeulen,R.A., Montavon,G., Samek,W., Kloft,M., Dietterich,T.G. and Müller,K.-R. (2021) A unifying review of deep and shallow anomaly detection. *Proc. IEEE*, **109**, 756–795.
25. Fischer,J., Mayer,C.E. and Söding,J. (2008) Prediction of protein functional residues from sequence by probability density estimation. *Bioinformatics*, **24**, 613–620.

26. Reiss,T., Cohen,N., Bergman,L. and Hoshen,Y. (2021) In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2806–2814.
27. Hoshen,Y. (2022) Time series anomaly detection by cumulative radon features. arXiv doi: https://arxiv.org/abs/2202.04067, 08 February 2022, preprint: not peer reviewed.
28. Gu,X., Akoglu,L. and Rinaldo,A. (2019) Statistical analysis of nearest neighbor methods for anomaly detection. In: Wallach,H., Larochelle,H., Beygelzimer,A., d'Alché-Buc,F., Fox,E. and Garnett,R. (eds.) *Advances in Neural Information Processing Systems*. Vol. **32**.
29. Cohen,N., Abutbul,R. and Hoshen,Y. (2023) Out-of-distribution detection without class labels. arXiv doi: https://arxiv.org/abs/, 22 September 2022, preprint: not peer reviewed.
30. Webb,B. and Sali,A. (2016) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Bioinformatics*, **54**, 5.6.1–5.6.37.
31. Marks,D.S., Hopf,T.A. and Sander,C. (2012) Protein structure prediction from sequence variation. *Nat. Biotechnol.*, **30**, 1072–1080.
32. Bergman,L., Cohen,N. and Hoshen,Y. (2020) Deep nearest neighbor anomaly detection. arXiv doi: https://arxiv.org/abs/2002.10445, 24 February 2020, preprint: not peer reviewed.
33. Cohen,N. and Hoshen,Y. (2020) Sub-image anomaly detection with deep pyramid correspondences. arXiv doi: https://arxiv.org/abs/2005.02357, 05 May 2020, preprint: not peer reviewed.
34. Ben-Hur,A., Ong,C.S., Sonnenburg,S., Schölkopf,B. and Rätsch,G. (2008) Support vector machines and kernels for computational biology. *PLoS Comput. Biol.*, **4**, e1000173.
35. Devlin,J., Chang,M.-W., Lee,K. and Toutanova,K. (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv doi: https://arxiv.org/abs/1810.04805, 11 October 2018, preprint: not peer reviewed.
36. Tzachor,I. and Hoshen,Y. (2023) Window projection features are all you need for time series anomaly detection. In: *ICLR*.
37. Cohen,N., Tzachor,I. and Hoshen,Y. (2023) Set features for fine-grained anomaly detection. arXiv doi: https://arxiv.org/abs/2302.12245, 02 March 2023, preprint: not peer reviewed.
38. UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
39. Cavnar,W.B. and Trenkle,J.M. (1994) In: *Proceedings of SDAIR-94, 3rd Annual symposium on Document Analysis and Information Retrieval*. Las Vegas, NV, Vol. **161175**.
40. Moore,R.A., Taubner,L.M. and Priola,S.A. (2009) Prion protein misfolding and disease. *Curr. Opin. Struct. Biol.*, **19**, 14–22.
41. Afsar Minhas,F.U.A., Ross,E.D. and Ben-Hur,A. (2017) Amino acid composition predicts prion activity. *PLoS Comput. Biol.*, **13**, e1005465.
42. Singh,U. and Syrkin Wurtele,E. (2020) How new genes are born. *eLife*, **9**, e55136.
43. Hanson,A.D., Pribat,A., Waller,J.C. and Crécy-Lagard,V.D. (2010) 'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list–and how to find it. *Biochem. J.*, **425**, 1–11.
44. Sanchez,G., Dury,A.Y., Murray,L.M., Biondi,O., Tadesse,H., El Fatimy,R., Kothary,R., Charbonnier,F., Khandjian,E.W. and Cote,J. (2013) A novel function for the survival motoneuron protein as a translational regulator. *Hum. Mol. Genet.*, **22**, 668–684.
45. Zou,K., Gong,J.-S., Yanagisawa,K. and Michikawa,M. (2002) A novel function of monomeric amyloid β-protein serving as an antioxidant molecule against metal-induced oxidative damage. *J. Neurosci.*, **22**, 4833–4841.
46. Ma,R., Pang,G., Chen,L. and van den Hengel,A. (2022) In: *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*. pp.704–714.
47. Pang,G., Shen,C., Cao,L. and Hengel,A.V.D. (2021) Deep learning for anomaly detection: a review. *ACM Comput. Surv. (CSUR)*, **54**, 1–38.
48. Ruff,L., Vandermeulen,R., Goernitz,N., Deecke,L., Siddiqui,S.A., Binder,A., Müller,E. and Kloft,M. (2018) In: *International Conference on Machine Learning*. PMLR, pp. 4393–4402.
49. Reiss,T., Cohen,N., Horwitz,E., Abutbul,R. and Hoshen,Y. (2023) Anomaly detection requires better representations. In: Karlinsky,L., Michaeli,T. and Nishino,K. (eds.) *Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science*. Vol. **13804**, Springer, Cham.
50. Rippel,O., Mertens,P. and Merhof,D. (2021), Modeling the distribution of normal data in pre-trained deep features for anomaly detection. In: *25th International Conference on Pattern Recognition (ICPR)*. IEEE, pp. 6726–6733.
51. Hendrycks,D. and Gimpel,K. (2016) A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv doi: https://arxiv.org/abs/1610.02136, 07 October 2016, preprint: not peer reviewed.
52. Fort,S., Ren,J. and Lakshminarayanan,B. (2021) Exploring the limits of out-of-distribution detection. *Adv. Neural Inform. Process. Syst.*, **34**, 7068–7081.
53. Cohen,N., Kahana,J. and Hoshen,Y. (2022) Red PANDA: disambiguating anomaly detection by removing nuisance factors. arXiv doi: https://arxiv.org/abs/2207.03478, 07 July 2022, preprint: not peer reviewed.
54. Linial,M., Rappoport,N. and Ofer,D. (2017) Overlooked short toxin-like proteins: a shortcut to drug design. *Toxins*, **9**, 350.
55. Hemm,M.R., Paul,B.J., Miranda-Ríos,J., Zhang,A., Soltanzad,N. and Storz,G. (2010) Small stress response proteins in Escherichia coli: proteins missed by classical proteomic studies. *J. Bacteriol.*, **192**, 46–58.
56. Rappoport,N. and Linial,M. (2012) Viral proteins acquired from a host converge to simplified domain architectures. *PLoS Comput. Biol.*, **8**, e1002364.
57. Escalera-Zamudio,M. and Greenwood,A.D. (2016) On the classification and evolution of endogenous retrovirus: human endogenous retroviruses may not be 'human'after all. *APMIS*, **124**, 44–51.
58. Tuite,M.F. and Serio,T.R. (2010) The prion hypothesis: from biological anomaly to basic regulatory mechanism. *Nat. Rev. Mol. Cell Biol.*, **11**, 823–833.
59. Chakrabortee,S., Byers,J.S., Jones,S., Garcia,D.M., Bhullar,B., Chang,A., She,R., Lee,L., Fremin,B. and Lindquist,S. (2016) Intrinsically disordered proteins drive emergence and inheritance of biological traits. *Cell*, **167**, 369–381.
60. Malinovska,L., Kroschwald,S. and Alberti,S. (2013) Protein disorder, prion propensities, and self-organizing macromolecular collectives. *Biochim. Biophys. Acta (BBA) Proteins Proteomics*, **1834**, 918–931.

Received: March 14, 2023. **Revised:** November 17, 2023. **Editorial Decision:** January 29, 2024. **Accepted:** February 23, 2024