EZLearn: Exploiting Organic Supervision in Large-Scale Data Annotation

Maxim Grechkin
Paul G. Allen School of Computer Science
University of Washington
Seattle, WA

Hoifung Poon
Microsoft Research
Redmond, WA

Bill Howe
Information School
University of Washington
Seattle, WA

Abstract

Many real-world applications require large-scale data annotation, such as identifying tissue origins based on gene expression profiles and classifying images into semantic categories. Annotation classes are often numerous and subject to changes over time, and annotating examples has become the major bottleneck for supervised learning methods. In science and other high-value domains, large repositories of data samples are often available, together with two sources of *organic supervision*: a lexicon for the annotation classes, and text descriptions that accompany some data samples. Distant supervision has emerged as a promising paradigm for exploiting such indirect supervision by automatically annotating examples where the text description contains a class mention in the lexicon. However, due to linguistic variations and ambiguities, such training data is inherently noisy, which limits the accuracy in this approach. In this paper, we introduce an auxiliary natural language processing system for the text modality, and incorporate co-training to reduce noise and augment signal in distant supervision. Without any manually labeled data, our *EZLearn* system learned to accurately annotate data samples in functional genomics and scientific figure comprehension, even substantially outperforming state-of-the-art supervised methods trained on tens of thousands of annotated examples.

Introduction

The confluence of technological advances and the open data movement [20] has led to an explosion of publicly available datasets, heralding an era of data-driven hypothesis generation and discovery in high-value applications [24]. A prime example is *open science*, which promotes open access to scientific discourse and data to facilitate large-scale data reuse and scientific collaboration [7]. In addition to enabling reproducibility, this trend has the potential to accelerate scientific discovery, reduce the cost of research, and facilitate automation [25, 16].

However, progress is hindered by the lack of consistent and high-quality annotations. For example, tissues from neurons to blood share the same genome, but vary in gene expression, which is crucial to understanding cell differentiation and cancer [10, 9]. The NCBI Gene Expression Omnibus (GEO) [3] contains over two million sample gene expression profiles, yet only a fraction of them have explicit tissue annotation. As a result, only 20% of the datasets have ever been reused, and tissue-specific expression studies are still being done at small scale [24]. Similarly, figures in scientific papers convey rich information, but there is no principled way to search them by semantics [14].

Annotating data samples with standardized classes is the canonical multi-class classification problem, but standard supervised approaches are difficult to apply. Hiring experts to annotate examples for thousands of classes such as tissue types is unsustainable. Crowd-sourcing is generally not applicable, as annotation requires expertise that most crowd workers do not possess. Moreover, the annotation standard is often revised over time, incurring additional cost for labeling new examples.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

While labeled data is expensive and difficult to create at scale, unlabeled data is usually in abundant supply. Many methods have been proposed to exploit it, but they typically still require labeled examples to initiate the process [1, 18, 6]. Even zero-shot learning, where the name implies learning with no labeled examples for *some* classes, still requires labeled examples for related classes [22, 26].

In this paper, we propose *EZLearn*, which makes annotation learning easy by exploiting two sources of *organic supervision*. First, the annotation classes generally come with a lexicon for standardized references (e.g., "liver", "kidney", "acute myeloid leukemia cell" for tissue types). While labeling individual data samples is expensive and time-consuming, it takes little effort for a domain expert to provide a few example terms for each class. In fact, in the sciences and other high-value applications, such a lexicon is often available as part of an existing domain ontology. For example, the Brenda Tissue Ontology specifies 4931 human tissue types, each with a list of standard names [8]. We call such indirect supervision "organic" to emphasize that it is readily available as an integral part of a given domain. Second, data samples are often accompanied by a text description, some of which directly or indirectly mention the relevant classes (e.g., the caption of a figure, or the description entered by a lab technician for a gene expression sample). Together with the lexicon, these descriptions present an opportunity for exploiting distant supervision by generating noisy labeled examples at scale [19].

In practice, however, there are serious challenges to enact this learning process. Descriptions are created for general human consumption, not as high-quality machine-readable annotations. They are provided voluntarily by data owners and lack consistency of any kind. Ambiguity, typos, abbreviations, and non-standard references abound [15, 25]. Additionally, annotation standard evolves over time, some terms become obsolete but were used in older samples. As a result, while there are potentially many data samples whose description contains class information, only a fraction of them can be identified using distant supervision, and noises are introduced due to reference ambiguity. This problem is particularly acute for domains with a large number of classes and/or frequent update.

To best exploit indirect supervision using all instances, *EZLearn* introduces an auxiliary text classifier for handling complex linguistic phenomena in descriptions. This auxiliary classifier first uses the lexicon to find exact matches to teach the main classifier. In turn, the main classifier helps the auxiliary classifier improve by annotating additional examples where class mentions are non-standard or ambiguous. This co-supervision continues until neither classifier can improve any further. Effectively, *EZLearn* represents the first attempt in combining distant supervision and co-training, using text as the auxiliary modality for learning. Figure 1 shows the architecture.

To investigate the effectiveness and generality of *EZLearn*, we applied it to two important applications in functional genomics and scientific figure comprehension, which differ substantially in domain characteristics such as sample input dimension and description length. In functional genomics, there are thousands of well-established classes. In scientific figure comprehension, prior work only considers three coarse classes, and we expand them to twenty-four finer-grained ones. In both scenarios, *EZLearn* successfully learned an accurate classifier with zero manually labeled examples.

EZLearn

Let $X = \{x_i : i\}$ be the set of data samples and C be the set of classes. Automating annotation amounts to learning a multi-class classifier $f : X \rightarrow C$. For example, x_i may be a gene expression profile, whereas C is the set of tissue types. Additionally, t_i denotes the text description that accompanies x_i. Sometimes, the description is not available, in which case t_i is the empty string. By default, there are no available labeled examples (x, y^*) where $y^* \in C$ is the true class for annotating $x \in X$. Instead, *EZLearn* assumes that a lexicon L_c is available with a set of example terms for referencing $c \in C$. Note that we do not assume that L_c is complete, nor that such terms are unambiguous. Rather, we simply require that L_c is non-empty for any c of interest.

To handle linguistic variations and ambiguities, *EZLearn* introduces an auxiliary classifier $f_T : T \rightarrow C$, where $T = \{t_i : i\}$ is the set of text descriptions that accompany the data samples. f_T is initialized using the initial labeled set D^0, which contains all (x_i, c) where t_i contains a class reference in lexicon L_c. At iteration k, we first train a new main classifier f^k using D^{k-1}. We then apply f^k to X and create a new labeled set D^k_T, which contains all (t_i, c) where $f^k(x_i) = c$. We then train a new text classifier f^k_T using D^k_T, and create the new labeled set D^k with all (x_i, c) where $f^k_T(t_i) = c$. This process continues until convergence, which is guaranteed given conditional independence of the two views [1]. Empirically, it happens quickly. Algorithm 1 shows the *EZLearn* algorithm.

Method	# Labeled	# All	AUPRC	Prec@0.5	Use expression	Use text	Use lexicon	Use EM
URSA	14510	0	0.40	0.52	yes	no	no	no
Co-EM	14510	116895	0.51	0.61	yes	yes	no	yes
Dist. Sup.	0	116895	0.59	0.63	yes	yes	yes	no
EZLearn	0	116895	**0.67**	**0.83**	yes	yes	yes	yes

Table 1: Comparison of test results between *EZLearn* and state-of-the-art supervised, semi-supervised, and distantly supervised methods on the Comprehensive Map of Human Gene Expression. We reported the area under the precision-recall curve (AUPRC) and precision at 0.5 recall.

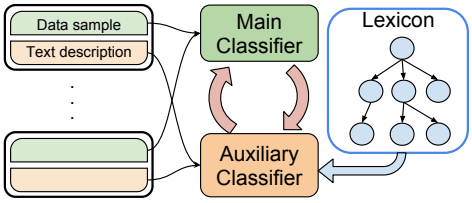

Figure 1: *EZLearn* architecture: an auxiliary text classifier is introduced to bootstrap from the lexicon (often available from an ontology) and co-teach the main classifier until convergence.

Algorithm 1 *EZLearn*

Input: Data samples X, text descriptions T, annotation classes C, and lexicon L_c containing example references for each class $c \in C$.
Output: Trained classifiers $f : X \rightarrow C$ (main) and $f_T : T \rightarrow C$ (auxiliary).
Initialize: Generate the initial training data D^0 by adding all (x_i, c) where $x_i \in X$ and its text description $t_i \in T$ mentions a term in L_c.
for $k = 1 : N_{iter}$ **do**
  $f \leftarrow \texttt{Train}_{\texttt{main}}(D^{k-1})$; $D^k_T \leftarrow f(X)$
  $f_T \leftarrow \texttt{Train}_{\texttt{aux}}(D^k_T)$; $D^k \leftarrow f_T(T)$
end for

In both the initialization step and later iterations, a labeled set might contain more than one class for a sample, which is not a problem for the learning algorithm and is useful when there is uncertainty about the correct class. We can use any classifier for $\texttt{Train}_{\texttt{main}}$ and $\texttt{Train}_{\texttt{aux}}$. Features for the main classifier are domain-specific and can be what any reasonable supervised approach might use. For the text classifier, we use standard n-gram features, which are effective in both applications we experimented on. It is possible to tailor them for specific domains. Generally, a classifier will output a score for each class, rather than predicting a single class. The score reflects the confidence in predicting the given class. *EZLearn* generates the labeled set by adding all (sample,class) pairs for which the score crosses a threshold, which is a hyperparameter. We chose 0.3 in preliminary experiments, which allows up to 3 classes to be assigned to a sample.

Application: Functional Genomics

Annotation task The goal is to annotate tissue types based one gene expressions. The input is a gene expression profile (a 20,000-dimension vector with a numeric value signifying the expression level for each gene). The output is a tissue type. We used the BRENDA Tissue Ontology [8], which contains 4931 human tissue types. For gene expression data, we used the Gene Expression Omnibus [5], a popular repository run by the National Center for Biotechnology Information. We focused on the most common data-generation platform (Affymetrix U133 Plus 2.0), and obtained a dataset of 116895 human samples. Each sample was processed using UPC to minimize batch effects and normalize expression values to [0,1] [23]. Text descriptions were obtained from GEOmetadb [31].

Main classifier We implemented $\texttt{Train}_{\texttt{main}}$ using deep denoising auto-encoder (DAE) with three LeakyReLU layers to convert the gene expression profile to a 128-dimensional vector [30], followed by multinomial logistic regression, trained in Keras [2], using L2 regularization with weight $1e-4$ and RMSProp optimizer [27] with default parameters.

Auxiliary classifier We implemented $\texttt{Train}_{\texttt{aux}}$ using the fastText classifier with their recommended parameters (25 epochs and starting learning rate of 1.0) [13]. The auxiliary classifier is initialized by simply predicting the most specific class in BRENDA with one of its standard terms appearing in the description. It is possible to have multiple matching classes, in which case all were added to the labeled set for training a new main classifier. In principle, we can continue the alternating training steps until convergence, when neither classifier's predictions change significantly. In practice, convergence usually comes quickly [21], and we simply ran all experiments with five iterations.

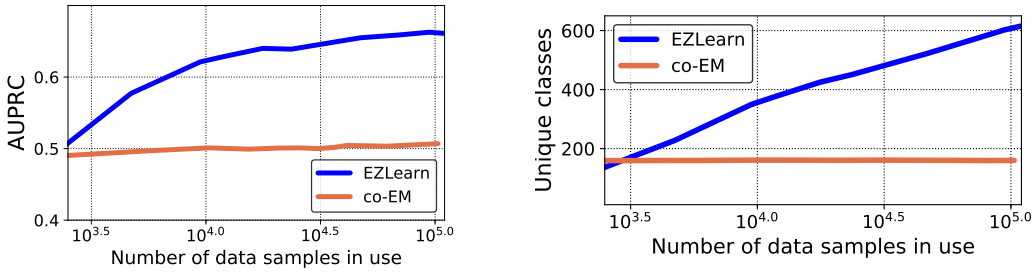

Figure 2: (a) Comparison of test accuracy with varying amount of unlabeled data.(b) Comparison of number of unique classes in high-confidence predictions with varying amount of unlabeled data.

Systems We compared *EZLearn* with URSA [15], the state-of-the-art supervised method that is trained on a large labeled dataset of 14,510 examples and used a sophisticated Bayesian method to refine SVM classification based on the ontology. We also compared it with co-training [1] and its variant co-EM [21], two representative methods for leveraging unlabeled data that also use an auxiliary view to support the main classification. Unlike *EZLearn*, they use labeled data to train their initial classifiers. After the first iteration, high-confidence predictions on the unlabeled data are added to the labeled examples. In co-training, once a unlabeled sample is added to the labeled set, it is not reconsidered again, whereas in co-EM, all of them are re-annotated in each iteration. We found that co-training and co-EM performed essentially the same, and so only report the co-EM results.

Evaluation We evaluated the classification results using *ontology-based precision and recall*. For each singleton class, predicted or gold, we expand it to include its ancestors other than the root (representing everything). We can then measure precision and recall in the standard way. Namely, precision is the proportion of correct predicted classes among all predicted classes, and recall is the proportion of correct predicted classes among gold classes, with ancestors included in all cases. This closely resembles the approach by [29], except that we are using the "micro" version (i.e., the predictions for all samples are first combined before measuring precision and recall), which is more appropriate in our applications. If the system predicts an irrelevant class in a different branch under the root, the overlap between the predicted and gold set is empty and the penalty is severe. If the predicted class is an ancestor (more general) or a descendent (more specific), there is overlap and the penalty is less severe, with overly general or specific predictions penalized more than close neighbors. We tested on the Comprehensive Map of Human Gene Expression (CMHGP), the largest expression dataset with manual tissue annotations [28]. CMHPG used tissue types from the Experimental Factor Ontology (EFO) [17], which can be mapped to the BRENDA Tissue Ontology. To make the comparison fair, 7,209 CMHGP samples that were in the supervised training set for URSA were excluded from the test set. The final test set contains 15,129 samples of 628 tissue types.

Results We report both the area under the precision-recall curve (AUPRC) and the precision at 0.5 recall. Table 1 shows the main classification results. All results were averaged over fifteen runs (except URSA). Remarkably, without using any labeled data, *EZLearn* outperformed the state-of-the-art supervised method by a wide margin, improving AUPRC by an absolute 27 points over URSA, and over 30 points in precision at 0.5 recall. Compared to distant supervision, the use of EM led to further significant gains of 8 points in AUPRC and 20 points in precision at 0.5 recall. Compared to co-EM, *EZLearn* improves AUPRC by 16 points and precision at 0.5 recall by 22 points. To investigate why *EZLearn* attained such a clear advantage even against co-EM, we compared their performance using varying amount of unlabeled data (averaged over fifteen runs). Figure 2(a) shows the results. Note that the x-axis (number of unlabeled examples in use) is in log-scale. Co-EM barely improves with more unlabeled data, whereas *EZLearn* improves substantially from 2% to 100% of unlabeled data. To understand why this is the case, we further compare the number of unique classes predicted by the two methods. See Figure 2(b). Co-EM is confined to the classes in its labeled data and its use of unlabeled data is limited to the extent of improving predictions for those classes. In contrast, by using organic supervision to generate noisy examples, *EZLearn* can expand the classes in its purview with more unlabeled data, while improving predictive accuracy for individual classes.

Application: Scientific Figure Comprehension

Figures communicate key results and provide visual explanations of complex concepts. However, while text understanding has been intensely studied, figures have received much less attention in the past. A notable exception is the Viziometrics project [14], which annotated a large number of examples for classifying scientific figures into semantic classes. Due to the considerable cost of labeling examples, they only used five coarse classes: `Plot`, `Diagram`, `Image`, `Table` and `Equation`. We exclude the last two as they do not represent true figures. In practice, figure-comprehension projects can be much more useful if they include larger set of more fine-grained classes. To explore this direction, we devised an ontology where `Plot`, `Diagram`, and `Image` are further refined into a total of twenty-four classes, such as `Boxplot`, `MRI` and `PieChart`. However, to cover these new classes, the supervised-learning approach adopted by Viziometrics will require annotating an even larger number of examples. *EZLearn*, on the other hand, does not require manually labeled data, and can be applied directly to learning the fine-grained classifier.

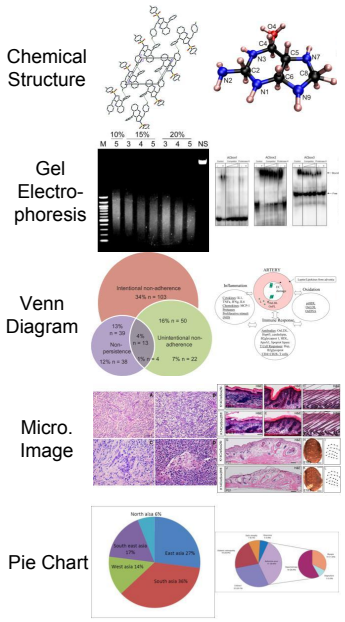

Figure 3: Example annotations by *EZLearn*, all chosen randomly among figures with no class information in their captions.

Annotation task The goal is to annotate figures with semantic types in the predefined ontology. The input is the image of a figure with varying size. The output is the semantic type. We obtained the data from the Viziometrics project [14] through its open API. For simplicity, we focused on single-pane figures, yielding 1,174,456 images along with free-text captions. As in the gene expression case, captions might be empty or missing.

Systems Each figure image was first resized and converted to a 2048-dimensional real-valued vector using a convolutional neural network [11] trained on ImageNet [4]. We follow [12] and use the ResNet-50 model with pre-trained weights provided by Keras [2]. We used the same classifiers and hyperparameters as in the functional genomics application. We used a lexicon that simply comprises of the names of the new classes, and compared *EZLearn* with the Viziometrics classifier. We also compared with a lexicon-informed baseline that annotates a figure with the most specific class whose name is mentioned in the caption (or root otherwise).

Evaluation We followed the functional genomics application and evaluated on ontology-based precision and recall. Since the new classes are direct refinement of the old ones, we can also evaluate the Viziometrics classifier using this metric. To the best of our knowledge, there is no prior dataset or evaluation for figure annotation with fine-grained semantic classes. Therefore, we manually annotated an independent test set of 500 examples.

Results *EZLearn* substantially outperformed both the lexicon-informed baseline and the Viziometrics classifier, scoring 79% in AUPRC compared to 44% (lexicon baseline) and 53% (Viziometric), and 88% precision at 0.5 recall compared to 31% (lexicon baseline) and 43% (Viziometric). The state-of-the-art Viziometrics classifier was trained on 3271 labeled examples, and attained an accuracy of 92% on the coarse classes. So the gain attained by *EZLearn* reflects its ability to extract a large amount of fine-grained semantic information missing in the coarse classes. Figure 3 shows example figure annotations by *EZLearn*, all chosen from figures with no class mention in their captions.

Discussion

We propose *EZLearn* for large-scale data annotation, which exploits two readily available sources of organic supervision: a lexicon containing standard class references and text descriptions provided by data owners. By introducing an auxiliary text classifier to co-teach the main classifier, *EZLearn* leverages co-training to reduce noise and amplify signal in distant supervision. *EZLearn* is well suited to the sciences and other high-value domains that contain a large number of classes and/or undergo frequent update. Experiments in functional genomics and scientific figure comprehension show that *EZLearn* is broadly applicable, robust to noise, and capable of learning accurate classifier without any manually labeled data, even outperforming state-of-the-art supervised systems by a wide margin. Future directions include: incorporate word embedding and other known semantic similarity; leverage hierarchical relations among annotation classes; apply *EZLearn* to other domains.

References

[1] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, 1998.

[2] François Chollet. Keras. `https://github.com/fchollet/keras`, 2015.

[3] Emily Clough and Tanya Barrett. The gene expression omnibus database. *Methods Mol Biol*, 1418, 2016.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR 2009*, pages 248–255. IEEE, 2009.

[5] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210, 2002.

[6] Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE TPAMI*, 28(4):594–611, 2006.

[7] Sascha Friesike, Bastian Widenmayer, Oliver Gassmann, and Thomas Schildhauer. Opening science: towards an agenda of open science in academia and industry. *J. of Tech. Transfer*, Aug 2015.

[8] Marion Gremse, Antje Chang, Ida Schomburg, Andreas Grote, Maurice Scheer, Christian Ebeling, and Dietmar Schomburg. The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, 2011.

[9] Maria Gutierrez-Arcelus, Halit Ongen, Tuuli Lappalainen, Stephen B Montgomery, Alfonso Buil, Alisa Yurovsky, Julien Bryois, Ismael Padioleau, Luciana Romano, Alexandra Planchon, et al. Tissue-specific effects of genetic and epigenetic variation on gene regulation and splicing. *PLoS Genet*, 11(1):e1004958, 2015.

[10] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[12] Bill Howe, Po-shen Lee, Maxim Grechkin, Sean T Yang, and Jevin D West. Deep mapping of the visual literature. In *WWW Companion*, pages 1273–1277, 2017.

[13] Armand Joulin, Edouard Grave, and Piotr Bojanowski Tomas Mikolov. Bag of tricks for efficient text classification. *EACL 2017*, page 427, 2017.

[14] Po-shen Lee, Jevin D West, and Bill Howe. Viziometrics: Analyzing visual information in the scientific literature. *IEEE Transactions on Big Data*, 2017.

[15] Young-suk Lee, Arjun Krishnan, Qian Zhu, and Olga G. Troyanskaya. Ontology-aware classification of tissue and cell-type signals in gene expression profiles across platforms and technologies. *Bioinformatics*, 29(23):3036–3044, 2013.

[16] Maxwell W. Libbrecht and William Stafford Noble. Machine learning applications in genetics and genomics. *Nat. Rev. Genetics*, 2015.

[17] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling sample variables with an experimental factor ontology. *Bioinformatics*, 2010.

[18] David McClosky and Eugene Charniak. Self-training for biomedical parsing. In *ACL*, 2008.

[19] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL*, pages 1003–1011, 2009.

[20] Jennifer C Molloy. The open knowledge foundation: open data means better science. *PLoS Biol*, 9(12):e1001195, 2011.

[21] Kamal Nigam and Rayid Ghani. Analyzing the effectiveness and applicability of co-training. In *CIKM*, 2000.

[22] Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009.

[23] Stephen R. Piccolo, Michelle R. Withers, Owen E. Francis, Andrea H. Bild, and W. Evan Johnson. Multiplatform single-sample estimates of transcriptional activation. *PNAS*, 2013.

[24] Heather A. Piwowar and Todd J. Vision. Data reuse and the open data citation advantage. *PeerJ*, 1:e175, October 2013.

[25] Johan Rung and Alvis Brazma. Reuse of public genome-wide gene expression data. *Nature Reviews Genetics*, 2013.

[26] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *NIPS*, 2013.

[27] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural nets for ML*, 2012.

[28] Aurora Torrente, Margus Lukk, Vincent Xue, Helen Parkinson, Johan Rung, and Alvis Brazma. Identification of cancer related genes using a comprehensive map of human gene expression. *PLOS ONE*, 11(6):1–20, 06 2016.

[29] Karin Verspoor, Judith Cohn, Susan Mniszewski, and Cliff Joslyn. A categorization approach to automated ontological function annotation. *Protein Science*, 15(6):1544–1549, 2006.

[30] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, pages 1096–1103. ACM, 2008.

[31] Yuelin Zhu, Sean Davis, Robert Stephens, Paul S. Meltzer, and Yidong Chen. GEOmetadb: powerful alternative search engine for the gene expression omnibus. *Bioinformatics*, 2008.