RESEARCH ARTICLE

Heuristic-enabled active machine learning: A case study of predicting essential developmental stage and immune response genes in *Drosophila melanogaster*

Olufemi Tony Aromolaran[1,2]*, Itunu Isewon[1,2], Eunice Adedeji[2,3], Marcus Oswald[4,5], Ezekiel Adebiyi[1,2], Rainer Koenig[4,5], Jelili Oyelade[1,2]*

1 Department of Computer & Information Sciences, Covenant University, Ota, Ogun State, Nigeria, **2** Covenant University Bioinformatics Research (CUBRe), Covenant University, Ota, Ogun State, Nigeria, **3** Department of Biochemistry, Covenant University, Ota, Ogun State, Nigeria, **4** Integrated Research and Treatment Center, Center for Sepsis Control and Care (CSCC), Jena University Hospital, Am Klinikum, Jena, Germany, **5** Institute of Infectious Diseases and Infection Control, Jena University Hospital, Am Klinikum, Jena, Germany

* ola.oyelade@covenantuniversity.edu.ng (JO); olufemi.aromolaran@stu.edu.cu.ng (OTA)

Abstract

Computational prediction of absolute essential genes using machine learning has gained wide attention in recent years. However, essential genes are mostly conditional and not absolute. Experimental techniques provide a reliable approach of identifying conditionally essential genes; however, experimental methods are laborious, time and resource consuming, hence computational techniques have been used to complement the experimental methods. Computational techniques such as supervised machine learning, or flux balance analysis are grossly limited due to the unavailability of required data for training the model or simulating the conditions for gene essentiality. This study developed a heuristic-enabled active machine learning method based on a light gradient boosting model to predict essential immune response and embryonic developmental genes in *Drosophila melanogaster*. We proposed a new sampling selection technique and introduced a heuristic function which replaces the human component in traditional active learning models. The heuristic function dynamically selects the unlabelled samples to improve the performance of the classifier in the next iteration. Testing the proposed model with four benchmark datasets, the proposed model showed superior performance when compared to traditional active learning models (random sampling and uncertainty sampling). Applying the model to identify conditionally essential genes, four novel essential immune response genes and a list of 48 novel genes that are essential in embryonic developmental condition were identified. We performed functional enrichment analysis of the predicted genes to elucidate their biological processes and the result evidence our predictions. Immune response and embryonic development related processes were significantly enriched in the essential immune response and embryonic developmental genes, respectively. Finally, we propose the predicted essential genes for future experimental studies and use of the developed tool accessible at http://heal.covenantuniversity.edu.ng for conditional essentiality predictions.

OPEN ACCESS

Citation: Aromolaran OT, Isewon I, Adedeji E, Oswald M, Adebiyi E, Koenig R, et al. (2023) Heuristic-enabled active machine learning: A case study of predicting essential developmental stage and immune response genes in *Drosophila melanogaster*. PLoS ONE 18(8): e0288023. https://doi.org/10.1371/journal.pone.0288023

Editor: Jian Xu, East China Normal University School of Life Sciences, CHINA

Received: April 3, 2023

Accepted: June 18, 2023

Published: August 9, 2023

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: https://doi.org/10.1371/journal.pone.0288023

Data Availability Statement: The data used for model evaluation is publicly available on UCI machine learning repository. The source code for the heal application is available at https://github.com/phemmy2k2/conditional-essentiality. The data can be accessed through the link to the data repository as shown below: https://zenodo.org/record/8117236.

Funding: 1. Deutsche Forschungsgemeinschaft (https://www.dfg.de/) within the project KO 3678/5-1, and the German Federal Ministry of Education and Research (BMBF, Fkz 01EO1002, 01EO1502 and 13N15711) 2. World Bank awarded to Covenant Applied Informatics and Communication Africa Centre of Excellence (CApIC-ACE) through the ACE Impact Project (2019 – 2024) The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

Introduction

A gene is defined as absolute essential if its loss of function causes infertility or death in an organism or cell. There are a few computational approaches for predicting gene essentiality including homology search and evolutionary analysis-based approach [1], constraint-based methods [2], and machine learning (ML) approaches [3, 4]. Conditionally essential genes are genes that are essential in a particular condition but non-essential in another condition.

Conditional essentiality has predominantly been defined in terms of growth conditions [5, 6]. Recent systematic studies of gene essentiality revealed that two sets of essential genes exist; core essential genes that are always required for viability, and conditional essential genes that vary in essentiality in different genetic and environmental contexts [7]. The variability in essentiality depends on the phenotype being assessed (lethality, reproduction, growth and/or development), the species in which the gene is encoded and environmental/growth conditions [8, 9]. Costanzo and colleagues posited that environments often affect genes with a close functional relation to the pathways that are perturbed by a condition [10].

Another cause of variability in gene essentiality is experimental conditions such as temperature, pH, nutrient availability and/or, potentially, exposure to pathogens or microbes. Conditional essentiality has been linked to genetic factors. Some studies that systematically compared gene essentiality among closely related yeast isolates identified modifier loci that alter gene essentiality [11, 12]. Genetic factors also give rise to a phenomenon known as synthetic lethality where the loss of one of two genes that perform similar functions could render non-essential genes essential and essential genes dispensable [13]. More recently, it has become evident that gene essentiality also depends on the ability of cells to adaptively evolve and proliferate despite the inactivation of an essential gene, suggesting that essentiality is not a property of genes, but of cellular functions [13].

Curation of experimentally identified essential genes in developmental stages has been documented in some model organisms such as *Drosophila melanogaster* [14]. A set of essential genes are also required when a foreign body invades a host organism, this results in the immune response condition [15, 16]. Identifying a comprehensive set of essential genes in both developmental stage and immune response condition will be beneficial for identifying potential novel drug and insecticidal targets that overcomes the current drug and insecticide resistance in the fight against some diseases such as malaria [17].

Constraint-based methods such as Flux balance analysis have been used to identify conditionally essential genes [18]. However, the constraint-based methods have some drawbacks, which includes inability to identify non-metabolic genes and cannot be used to investigate genome-scale metabolic reactions under transient dynamic states without including data on enzyme kinetics [19, 20].

TnseqDiff is another commonly used computational approach for conditional essentiality prediction [21]. TnseqDiff utilizes two steps to estimate the conditional essentiality for each gene in the genome. First, it collects evidence of conditional essentiality for each insertion by comparing read counts of that insertion between conditions. Second, it combines insertion-level evidence to infer the essentiality for the corresponding gene. [21]. One of the major limitations of this approach is that transposon sequencing (Tn-seq) data is only available for a few model organisms thereby limiting the approach to bacteria. Owing to the genetic similarities and conserved pathways between *D. melanogaster* and mammals, the use of the Drosophila model as a platform to unveil novel mechanisms of infection and disease progression has been widely investigated [22] including host-pathogen interaction studies [23, 24].

Manimaran and others made the first attempt to explain the conditional essentiality of genes using the ML approach [25]. They obtained the protein interaction dataset from predictions of genome-wide functional linkages in E. coli which contains 3,682 proteins and 78,048 interactions. Three centrality features (Degree, closeness, and betweenness) were used with an SVM model. The study focused on growth conditions determined based on the expression of the genes in a microarray dataset. They predicted 1192 bacteria genes to be conditionally essential across 61 growth conditions. An extensive experiment was conducted to obtain the essentiality status of the genes used to train the ML model which is a very expensive, challenging, and time-consuming approach.

Computational prediction of conditional essentiality research is an open problem that is gaining wide attraction in recent years. Recent reviews have identified prediction of conditionally essential genes as a major limitation of the ML approaches so far [19, 20]. This is a major limitation because there is no sufficient labelled data to train ML techniques for predicting conditional essentiality. This study is motivated by the challenge posed by data that could not be manually annotated by experts or require experiments for annotation as found in experimental studies, an example is identifying gene function from sequence information or predicting conditionally essential genes. Therefore, we sought to develop a ML technique that is capable of reliably predicting conditionally essential genes in both model and non-model organisms.

Active machine learning (AL) techniques have been used for annotating unlabelled data based on the limited data as seen in image recognition [26], activity recognition [27], and text labelling [28]. A traditional AL algorithm was presented by [29, 30]. Active learning algorithms are iterative sampling schemes, where a classification model is adapted regularly by feeding it with new labelled samples corresponding to the ones that are most beneficial for the improvement of the model performance. The new labelled samples used to improve the model performance are obtained using a sample selection strategy. A sample selection strategy describes the techniques used by the active learning procedure to select the most valuable points to be manually labelled. Some of the commonly used strategies are Uncertainty-based [31], Committee-based [32, 33] and Expected Impact [34] selection strategies.

Uncertainty-based AL technique is the most widely used. The challenge with the application of the existing AL techniques in bioinformatics is that most biological data require extensive literature search and experiments for annotation which is time-consuming and very expensive. Therefore, the use of AL techniques for annotating biological data when there is limited label data requires an innovative approach. This study replaces the human component of AL with a heuristic component to enable the application of AL to predict conditionally essential genes and pave a way for broad application of heuristic-enabled active learning to solve challenging problems in biomedical research. To our knowledge, this study is the first to apply machine learning method to conditional essentiality prediction.

Methods

Defining benchmark models and datasets

To benchmark the HEAL technique, the sampling query strategy used was replaced with a random selection technique which is hereafter referred to as the RandAL technique. In addition, the traditional AL technique that uses the uncertainty query strategy was implemented hereafter referred to as the UncAL technique. The RandAL technique trains the base classifier using the labelled data and the trained classifier was used to pre-label all the samples in the unlabelled set. Subsequently, a specified batch (n = 20) of pre-labelled samples were randomly selected and presented to the expert for manual correction of the pre-labelled samples. The manually corrected labelled samples are then added to the labelled dataset for the next iteration of the annotation until a stop criterion is reached. The stop criterion in this experiment is

Table 1. Description of the real-world datasets for model validation.

Dataset	Number of Instances	Positive samples size	Negative samples size	Number of Attributes
Breast cancer	699	241	458	10
Credit rating-A	690	307	383	14
Credit rating-G	1000	700	300	24
Diabetes	768	268	500	8
Heart disease	270	150	120	13

https://doi.org/10.1371/journal.pone.0288023.t001

when 90% of the entire dataset has been labelled. For the UncAL technique, the random selection of query samples was replaced by the uncertainty technique. The three techniques (HEAL, UncAL, and RandAL) were evaluated based on five real-world datasets retrieved from the UCI Machine Learning Repository [35]. The real-world datasets are diabetes, breast cancer, heart disease, and credit scoring (Australia and German) datasets. Table 1 presents the description of the real-world datasets. For each of the five datasets, 20 percent of the dataset was randomly selected as the labelled set and the class label was excluded from the remaining 80 percent which was designated as the unlabelled set.

Datasets for conditional essentiality

This study sought to develop a technique for annotating unlabelled data when there is limited label data to train an ML model. Annotating genes that are essential in a given condition presents a problem of sparsity of labelled data [36]. For the purpose of evaluation of the developed technique, two categories of conditions in *D. melanogaster* were evaluated, these are the embryonic developmental condition and immune response condition. A total of 161 and 343 genes were collected from FlyBase as essential in embryonic developmental condition and immune response conditions respectively, while 12,058 and 11,993 genes were assigned as the negative class label samples in embryonic developmental condition and immune response conditions respectively. The essential embryonic genes were queried from Flybase using the term “lethal—all die during embryonic stage” which implies the organism died when the genes were mutated during the embryonic stage. For immune response labelling, Drosophila phenotypic data “allele_phenotypic_data_fb_2020_02.tsv” was downloaded from Flybase. Lethal and immune response were used to filter the phenotype of the genes. The most common technique used for identifying the physical trait of the genes was transposon mutagenesis. 21 genes were found to be essential in both conditions as shown in Fig 1B.

The training data for embryonic developmental condition comprised of 80 positive and 1600 negative samples which were randomly selected to represent the labelled data while the remaining samples represent the set of unlabelled data for the active learning analysis. 144 of the 343 essential immune response genes and 2880 of the non-essential immune genes were also randomly selected to represent the labelled data while the remaining samples represent the set of unlabelled data.

Feature generation

Feature quality is a major factor in the development of a ML model for predicting essential genes. A total of 50,334 features were generated based on broad range of features derived from (1) gene sequence, (2) protein sequence, (3) functional domains of the proteins, (4) gene sets from Gene Ontology (GO), (5) the number of homologous sequences, (6) topology properties from protein-protein interaction networks, and (7) subcellular localization of the protein (Fig 1A). Protein and gene sequences were downloaded from Ensembl [37, 38] using BioMart [39].

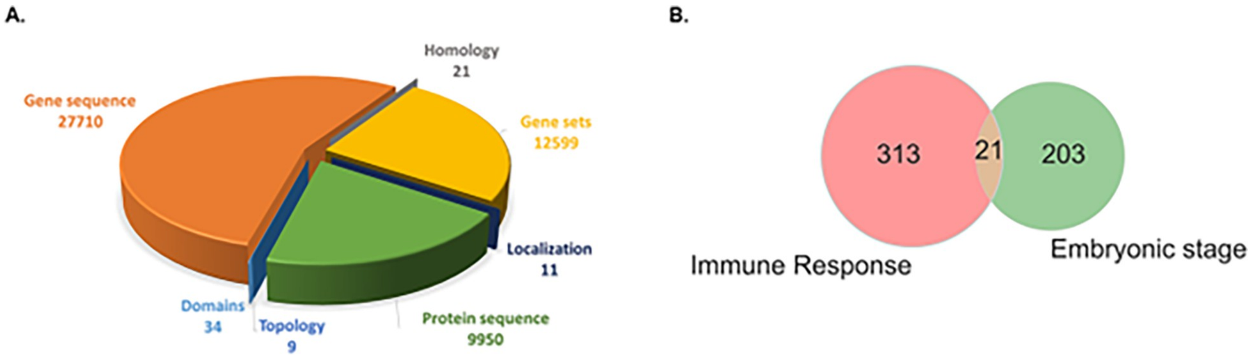


Fig 1. Distribution of the features and class label for conditional essentiality prediction in *D. melanogaster*. A. The generated features included intrinsic (e.g. protein and DNA sequence) and extrinsic features (e.g. topology of co-expression and protein-protein interaction networks). The number of features derived from individual categories is shown below the various categories. **B.** Venn diagram shows the total number of essential genes in each condition obtained from FlyBase and the number of genes essential in both conditions.

https://doi.org/10.1371/journal.pone.0288023.g001

For deriving the protein and gene sequence features (features in categories 1 and 2), various numerical representations characterizing the nucleotide and amino acid sequences and compositions of the query genes were calculated using seqinR [40], protr [41], CodonW [42] and rDNAse [43]. Using seqinR [40] the number and fraction of individual amino acids and other protein sequence features including the number of residues, the percentage of physico-chemical classes and the theoretical isoelectric point were calculated. Further protein sequence features were obtained using protr [41] including autocorrelation, Conjoint Triad Descriptors (CTD), quasi-sequence order and pseudo amino acid composition. CodonW [42] was used to calculate gene characteristics like gene length and GC content but also frequencies of optimal codons (frequency of codons favored by natural selection, see [44]) and the effective number of codons. Using rDNAse [43] gene descriptors like auto covariance or pseudo nucleotide composition, and *kmer* frequencies (n = 2–7) were calculated.

For deriving domain features (feature category 3), BioMart was used to obtain protein family (*pfam*) domains, number of coiled coils, the prediction of membrane helices, post-translational modifications, β-turns, cofactor binding, acetylation and glycosylation sites, trans membrane helices and signal peptides. In addition, the number and lengths of UTRs were obtained using BioMart. For features obtained from gene sets defined by Gene Ontology (feature category 4), gene sets of all GO terms including biological process, cellular localization and molecular function were obtained from Ensembl (version 102, released in Nov, 2020) [37, 38]. Gene sets were removed if they showed high redundancy according to the following method. The gene overlap of each pair of gene sets A and B was quantified by Jaccard similarity coefficients,

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

Pairs with J(A, B) above a threshold (threshold = 0.3) were included in the model and represented as an undirected graph, G = (X, E), with the gene sets as vertices X and the pairs above the threshold as edges E. A linear model was set up with a constraint to select at most one of the vertices of an edge:

$$Xi + Xj \leq 1, \qquad for\ every\ \{i, j\} \in E \tag{2}$$

$$Xi = 0,\ or\ Xi = 1, \quad for\ 1 \leq i \leq n \tag{3}$$

with the objective function to *maximize*

$$\sum w_i X_i$$

where, w_i is the weight of a gene set. The weight is derived from its significance (p-value) and calculated as $1 - log10(p\text{-}value)/100$. This maximization was done employing linear integer programming solved using Gurobi (version 7.5.1, https://www.gurobi.com). With this, we formulated the optimization problem to select at most one gene set from each pair in such a way that the overall number of non-redundant gene sets was maximized. This optimization problem was formulated as a mixed integer linear programming problem and solved using Gurobi (version 7.5.1, https://www.gurobi.com). A gene list was generated for each query gene according to a protein association network obtained from the STRING database [45]. The gene list for a gene is the set of all adjacent genes in the protein association network. A gene set enrichment test was performed employing Fisher's exact test and the negative *log10* of the *p*-value was used as a feature.

The number of homologous proteins (feature category 5) was obtained by blasting the protein sequence of the query protein against the complete RefSeq database [46] using PSI-BLAST [47]. The number of proteins found with e-value cutoffs from 1e–5 to 1e-100 were used as features. Topology features (feature category 6) were computed using the NetworkX [48] library in Python. Protein association data was downloaded from STRING [45] and an undirected network was constructed. From this, degree, degree distribution, closeness centrality, eigenvalue centrality, betweenness centrality, harmonic centrality, subgraph centrality, load centrality and Page rank as topological features were computed for each gene. To note, the harmonic centrality of a node *g* is the sum of the reciprocal of the shortest path distances from all other nodes to *g*. The higher the value, the higher the centrality [49]. The subcellular localization of proteins (feature category 7) was derived using DeepLoc [50]. DeepLoc predicts the likely location of a protein within a cell by assigning probability scores to eleven eukaryotic cell compartments (cytoplasm, nucleus, extracellular, mitochondria, plasma membrane, ER, chloroplast, Golgi apparatus, lysosome, vacuole and peroxisome). In total we generated 50,334 features.

Data normalization and feature selection

The dataset for conditional essentiality prediction consists of thousands of features from different categories with different range of values. Therefore, the data requires to be normalized and prepared for ML. All the features were merged into a single table followed by a *z*-score transformation of each feature to normalize the data. In addition, redundant highly correlating features with Pearson correlation coefficients > 0.70 were removed to avoid multicollinearity which introduces a bias in the analysis and extrapolation is likely to be seriously erroneous [51, 52]. If more than two features are highly correlated, then the one with the highest correlation with the target class was selected.

To overcome the class imbalance problem when training the classifier, Synthetic Minority Over-Sampling Technique (SMOTE) was used. SMOTE is a technique that creates synthetic, non-duplicated samples of the minority class, thereby making the total samples in both minority and majority classes to be equal [53]. For each minority class observation, SMOTE calculates the *k* nearest neighbours and randomly creates multiple synthetic samples between the observation and the nearest neighbours depending on the number of oversampling needed.

For each iteration and based on the labelled set, we performed two steps for feature selection prior to training of the machines. First, we applied an embedded approach based on Random Forests as suggested by [54] for feature selection. Each tree in the forest was initialized by bootstrapping from the training data to train a baseline model. Its performance was estimated using the out-of-bag (OOB) samples from the training data. Then, the values of one feature were randomly shuffled, keeping all other features the same, yielding permutated data. The permutated dataset was applied to the learned model and its performance was evaluated. Finally, the difference between the benchmark score from the baseline model and the score from the permutated model was calculated to determine the importance of the feature [55]. By this, we ranked all features and selected the top 400 features for training the downstream classifier.

Heuristic-enabled active machine learning

In this study, Light GBM, an ensemble model was used as the classifier for the active learner due to its high prediction accuracy and fast execution time [55]. Also, in recent studies, ensemble models such as Random forest and Extreme gradient boosting have shown to have a good performance on numerical data from biological sources [3, 56]. Due to the small size of the labelled data, 5-fold CV was used during the training of the classifiers. The hyper-parameter settings for the classifier was set according to the optimal settings obtained in our previous study [56] where *n_estimators = 600*, *learning_rate = 0.05*, *num_leaves = 32*, *colsample_bytree = 0.2*, *reg_alpha = 3*, *reg_lambda = 1*, *min_split_gain = 0.01 and min_child_weight = 40*.

The traditional AL algorithm was modified by replacing the human component with a heuristic function that uses a threshold specified by the user to filter queried samples. The sampling query function was also modified to use the *certainty* sampling technique proposed by this study instead of the widely used uncertainty technique. The certainty technique is the reverse of the uncertainty method. Unlike the uncertainty method that selects samples close to the classification or decision boundary as queries for the human expert, the certainty technique selects samples with high prediction confidence, these are samples with high prediction probability for the positive class and very low prediction probability for the negative class. Typically, the prediction probability is between 0 and 1 and by default, ML algorithms set their classification boundary as 0.5. It classifies all samples with a prediction probability below 0.5 as negative samples while those with a level of 0.5 and above are classified as positive samples. However, the classification boundary was set to 0.6 for this analysis when it was observed that the data was biased towards the positive samples and resulting in a high false positive rate.

The sample selection strategy described in the heuristic function introduced by this study is based on a cut-off that is dynamically assigned at each iteration according to the distribution of the classes in the pre-labelled dataset. The distribution of the prediction probability of the positive (0.6–1.0) and negative (0.0–0.59) samples as obtained from the automatic annotation is represented as quantiles. Samples in the first quartile for the negative distribution (Q^{-1}, closer to 0) and samples within the fourth quartile of the positive distribution (Q^{+4}, closer to 1) were selected by the heuristic function for further filtering. The heuristic function contains a threshold set by the user to exclude samples below the threshold and append the samples with values above the threshold to the labelled data. This further increases the classifier's prediction power which has a similar impact as selecting samples closer to the classification boundary, requiring humans to refine the automatic annotation by the classifier manually. A threshold of 0.9 was chosen by this study which ensures samples very close to the positive samples in the labelled data are annotated as positive. The complete implementation is described in Algorithm 1 and 2 and the schematic workflow of the Heuristic-Enabled Active Learning (**HEAL)** is shown in Fig 2.

```
Algorithm 1: Heuristic-Enabled Active Learning Algorithm
Inputs
- Initial training set $X^{\alpha} = \{x_i, y_i\}_{i=1}^{l}$ (X∈χ, α =1)
- Pool of candidates $U^{\alpha} = \{x_i\}_{i=l+1}^{l+u}$ (U∈χ, α =1)
- Threshold = confidence cut-off specified by the user.
repeat
  Train a model with the current training set $X^{\alpha}$.
  for each candidate in $U^{\alpha}$ do
    Score = Evaluate a user-defined classification model
  end for
  $P^{\alpha}$ = $U^{\alpha}$ + Score
  Rank the candidates in $P^{\alpha}$ according to the score of the classification model
  $S^{\alpha}$ = Heuristic component ($P^{\alpha}$, threshold)
  Add the batch to the training set $X^{\alpha+1} = X^{\alpha} \cup S^{\alpha}$
  Remove the batch from the pool of candidates $U^{\alpha+1} = U^{\alpha} \setminus S^{\alpha}$
  α = α+1
until a stopping criterion is met.
```

```
Algorithm 2: Heuristic component ($P^{\alpha}$, threshold):
  Divide $P^{\alpha}$ into $P^{+}$ and $P^{-}$ sets according to ranked score
  Compute quartiles for $P^{+}$ and $P^{-}$: $Q_i = \frac{i}{4}(n+1)^{th}\ term|(1 \le i \le 4, n = |U|)$
  Dynamically estimate the filtering cut-off for $P^{+}$ and $P^{-}$
    if min ($P^{+Q4}$) > threshold
    sup = min ($P^{+Q4}$)
    else:
    sup = threshold
  Set cut-off for $P^{-}$: inf = max ($P^{-Q1}$)
  Filter $P^{+}$ and $P^{-}$ based on sup and inf respectively where $q$ = the batch size of selected points $S$.
  Return filtered sample points $S^{\alpha} = \{x_k, y_k\}_{k=1}^{q}$ to the main function
```

Gene set enrichment analyses of the predicted essential immune response genes

To discover the biological and functional knowledge of the genes predicted to be essential for embryo developmental stage and immune response conditions in *D. melanogaster*, gene set enrichment analysis was performed using g:Profiler based on the Ensembl database version 102 [57]. The SCS algorithm with default settings as described by [57] was used to correct for multiple testing and the significance threshold was set to $p = 0.05$. The term size of the selected enriched gene sets was set between 3 and 500 to filter out too specific and too general gene sets.

Results

Evaluation of HEAL and benchmark models

A main hypothesis of this study was that a comparable performance could be achieved by replacing the human component in the AL model with a heuristic function to eliminate the high cost and time involved in using the traditional AL model. To evaluate the performance of HEAL to existing traditional AL models, we implemented Uncertainty AL, which is the traditional AL model based on uncertainty sample selection method with human component and Random AL, which is also a traditional AL model based on random sample selection method. Five publicly available datasets were applied to the three AL models. The results show that HEAL performed comparatively better than UncAL and has superior performance when compared with RandAL (Fig 3). HEAL has a significantly lower running time compared to both UncAL and RandAL across the five datasets (Fig 3D). The reason for the low running time is because HEAL does not require an expert for its annotation which is associated with

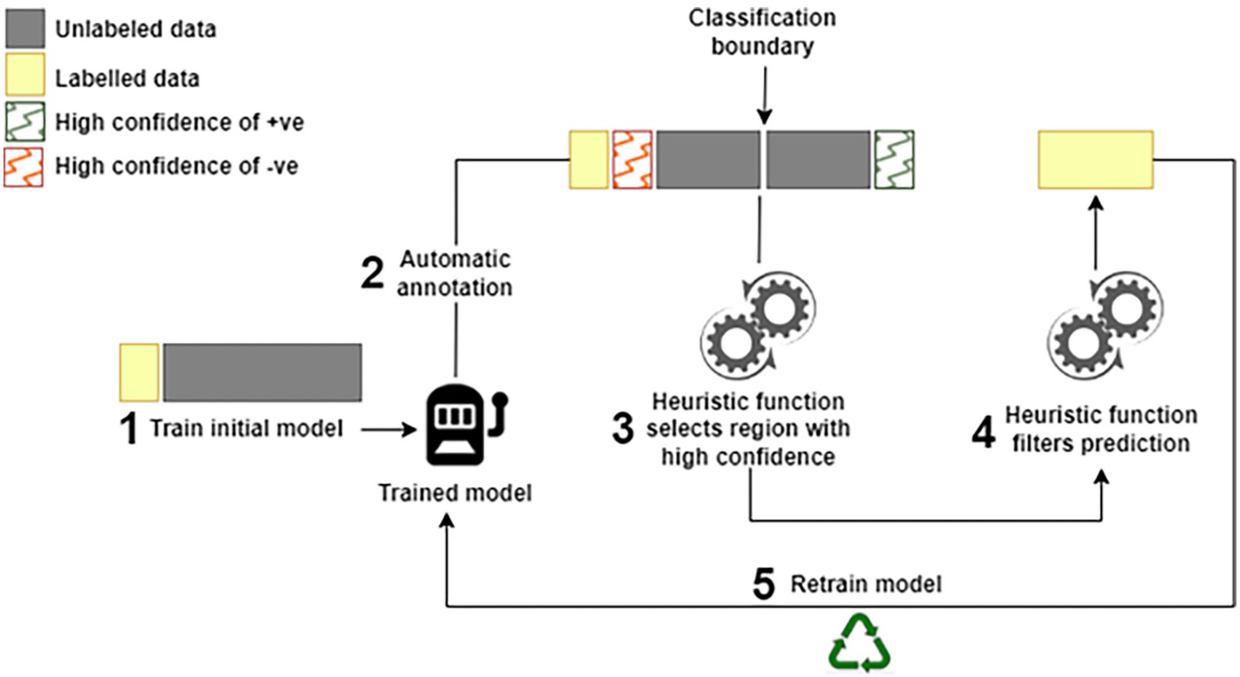


Fig 2. Heuristic-enabled active learning implemented to label conditionally essential genes.

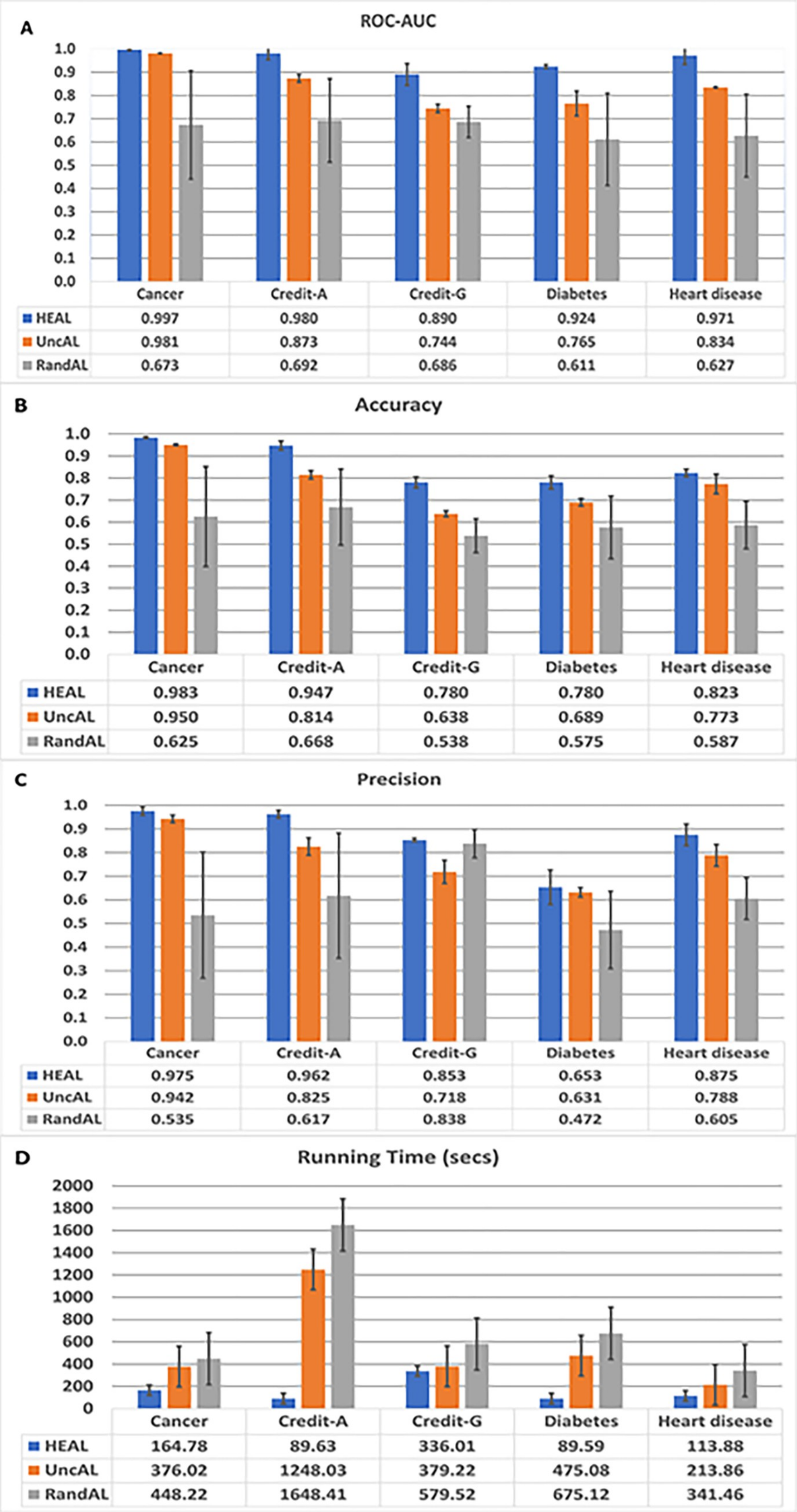
https://doi.org/10.1371/journal.pone.0288023.g002

computational delay prior to each iteration in the process. A stratified view of the predictions from the three techniques across the five datasets based on the confusion matrix is presented in S1 Fig.

Prediction of conditionally essential genes in *Drosophila melanogaster*

To evaluate the proposed model for conditional essentiality prediction, two conditions based on immune response and developmental stage conditions were examined. For the developmental stage conditions, a total of 53 genes were predicted as essential in the embryonic stage of *D. melanogaster* after five iterations with five of the predicted genes annotated as such in FlyBase. We performed gene set enrichment analysis to elucidate the biological processes enriched in the predicted genes. We found several growth and morphogenesis processes, such as post-embryonic animal morphogenesis and post-embryonic animal organ development (Fig 4A) indicating the need for these specific growth processes for the organism to develop from the embryonic stage into the larva stage. Table 2 shows the list of top 10 genes predicted to play essential roles during the embryonic stage of *D. melanogaster* and the complete list of predicted genes is shown in S1 Table.

For the immune response conditions, a total of 10 genes were predicted as essential for these conditions in *D. melanogaster* after eight iterations with 6 of the predicted genes annotated as such in FlyBase. Strikingly, the enrichment analysis of the predicted immune response genes revealed that immune and defense response related processes are significantly enriched in the predicted genes (Fig 4B) which implies that the four novel immune response genes would be good candidates for further experimental validation. Table 3 shows the list of the 10 genes predicted to play essential roles during the embryonic stage of *D. melanogaster*. Genes highlighted in red were found to be annotated as essential immune response genes in FlyBase.



	Cancer	Credit-A	Credit-G	Diabetes	Heart disease
■ HEAL	0.997	0.980	0.890	0.924	0.971
■ UncAL	0.981	0.873	0.744	0.765	0.834
■ RandAL	0.673	0.692	0.686	0.611	0.627

	Cancer	Credit-A	Credit-G	Diabetes	Heart disease
■ HEAL	0.983	0.947	0.780	0.780	0.823
■ UncAL	0.950	0.814	0.638	0.689	0.773
■ RandAL	0.625	0.668	0.538	0.575	0.587

	Cancer	Credit-A	Credit-G	Diabetes	Heart disease
■ HEAL	0.975	0.962	0.853	0.653	0.875
■ UncAL	0.942	0.825	0.718	0.631	0.788
■ RandAL	0.535	0.617	0.838	0.472	0.605

	Cancer	Credit-A	Credit-G	Diabetes	Heart disease
■ HEAL	164.78	89.63	336.01	89.59	113.88
■ UncAL	376.02	1248.03	379.22	475.08	213.86
■ RandAL	448.22	1648.41	579.52	675.12	341.46

Fig 3. Comparative analysis of HEAL and other techniques on five real-world datasets. Results from HEAL technique show superior performance in terms: **A.** ROC-AUC **B.** Accuracy and **C.** Precision. **D.** HEAL has the lowest running time compared to other methods.

https://doi.org/10.1371/journal.pone.0288023.g003

Implementation of HEAL technique as a web resource

Lack of labelled data has been a challenge in bioinformatics research that has persisted for decades. Prediction of conditionally essential genes using a machine learning approach is an example of the numerous bioinformatics problems that remained intractable. Notably, this study developed the HEAL technique into a web application that provides a tool for bioinformatics analyses to annotate biological and non-biological data when there is limited labelled data for training a machine learning model. Django framework was used for the development of the web version of the HEAL technique. Django is a high-level Python framework that encourages rapid development and clean, pragmatic design. The framework integrates Hypertext mark-up language (HTML), Cascading stylesheet (CSS), and JavaScript seamlessly with Python. This application can be found on heal.covenantuniversity.edu.ng. Fig 5 presents the HEAL web application portal. The portal provides fields for uploading the labelled and unlabelled data in CSV format. After uploading the input data, the user has the option to first preview the statistical information about the input data or perform data annotation directly. The data statistics returned include the number of features, the ratio of labelled to unlabelled, and the ratio of the class label of the initial training data. The data statistics are displayed on the right-hand side of the screen. If an error occurs during the prediction, a log will appear specifying the error. The server was configured to reject the processing of files with sizes more than 4Mb to avoid overloading the server with large data. This implies that feature selection should be performed before using this tool for data with a large feature set.

A dropdown element provides a list of threshold values for the active learning component for selection by the user. The threshold determines the stringency level of the active learning

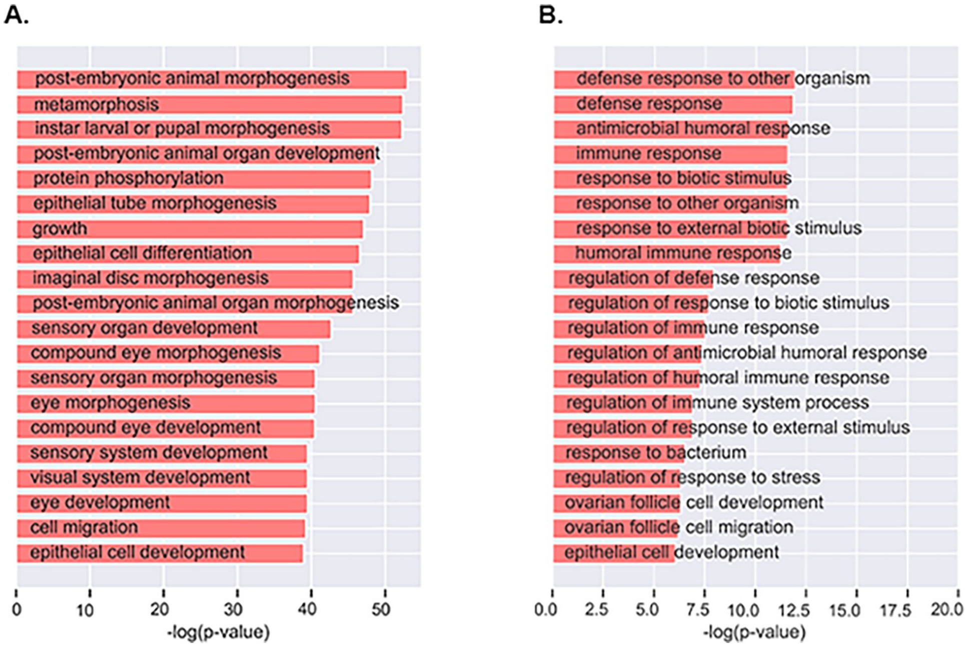


Fig 4. Functional enrichment analysis of genes predicted by the HEAL model. A. Biological process enriched in the predicted essential embryonic genes. B. Biological process enriched in the predicted essential immune response genes.

https://doi.org/10.1371/journal.pone.0288023.g004

Table 2. Top 10 predicted essential embryonic stage genes.

FlyBase ID	Gene Name	Gene Description
FBgn0265623	Su(z)2	Suppressor of zeste 2
FBgn0283427	FASN1	Fatty acid synthase 1
FBgn0287184	FASN3	Fatty acid synthase 3
FBgn0286784	TER94	Transitional endoplasmic reticulum 94
FBgn0003315	satDNA	satellite DNA, unknown function
FBgn0286785	scb	scab
FBgn0286786	hoip	hoi-polloi
FBgn0036448	mop	myopic
FBgn0265434	zip	zipper
FBgn0036980	RhoBTB	Rho-related BTB domain containing

https://doi.org/10.1371/journal.pone.0288023.t002

component with the default threshold value set to 0.9. After the server successfully finishes the processing, the output from the analysis which contains the dynamic cut-off for selecting high confidence samples, current classifier performance scores, and the number of samples added to the labelled set for each iteration are displayed in the result page shown in Fig 6. The base classifier performance indicates the confidence of the prediction. It is recommended to have a minimum of 90% base model accuracy for a reliable prediction.

Discussion

As at the time of conducting this research, there are no studies found from literature reviewed that have successfully applied machine learning techniques to predict conditionally essential genes responsible for any condition. A related study used a semi-supervised ML approach to predict HIV dependency factors in humans using only network-based features from protein interaction databases. They reported a precision score of 85% at 60% recall [58]. Some of the top-ranked genes predicted as essential in immune response conditions are discussed below along with their functions with respect to their importance to the organism's immune response conditions.

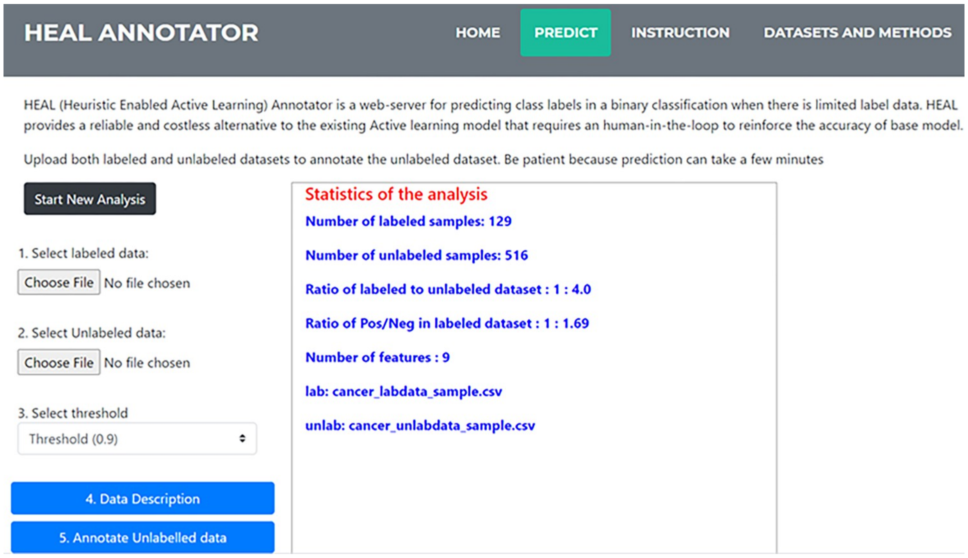
The state-of-the-art AL techniques have consistently used the uncertainty method for sampling selection and present the queried samples to the expert for manual correction of the pre-labelled samples. For the next iteration, the manually corrected samples were added to the

Table 3. Predicted essential immune response genes.

FlyBase ID	Gene Name	Gene Description
FBgn0035976	PGRP-LC	Peptidoglycan recognition protein LC
FBgn0016917	Stat92E	Signal-transducer and activator of transcription protein at 92E
FBgn0041184	Socs36E	Suppressor of cytokine signaling at 36E
FBgn0043903	dome	domeless
FBgn0000250	cact	cactus
FBgn0034476	Toll-7	Toll-7
FBgn0004364	18w	18 wheeler
FBgn0002930	nec	necrotic
FBgn0086358	Tab2	TAK1-associated binding protein 2
FBgn0000229	bsk	basket

# Genes in bold typeface are annotated as immune genes in FlyBase

https://doi.org/10.1371/journal.pone.0288023.t003



HEAL (Heuristic Enabled Active Learning) Annotator is a web-server for predicting class labels in a binary classification when there is limited label data. HEAL provides a reliable and costless alternative to the existing Active learning model that requires an human-in-the-loop to reinforce the accuracy of base model.

Upload both labeled and unlabeled datasets to annotate the unlabeled dataset. Be patient because prediction can take a few minutes

1. Select labeled data: No file chosen

2. Select Unlabeled data: No file chosen

3. Select threshold: Threshold (0.9)

Statistics of the analysis

Number of labeled samples: 129

Number of unlabeled samples: 516

Ratio of labeled to unlabeled dataset : 1 : 4.0

Ratio of Pos/Neg in labeled dataset : 1 : 1.69

Number of features : 9

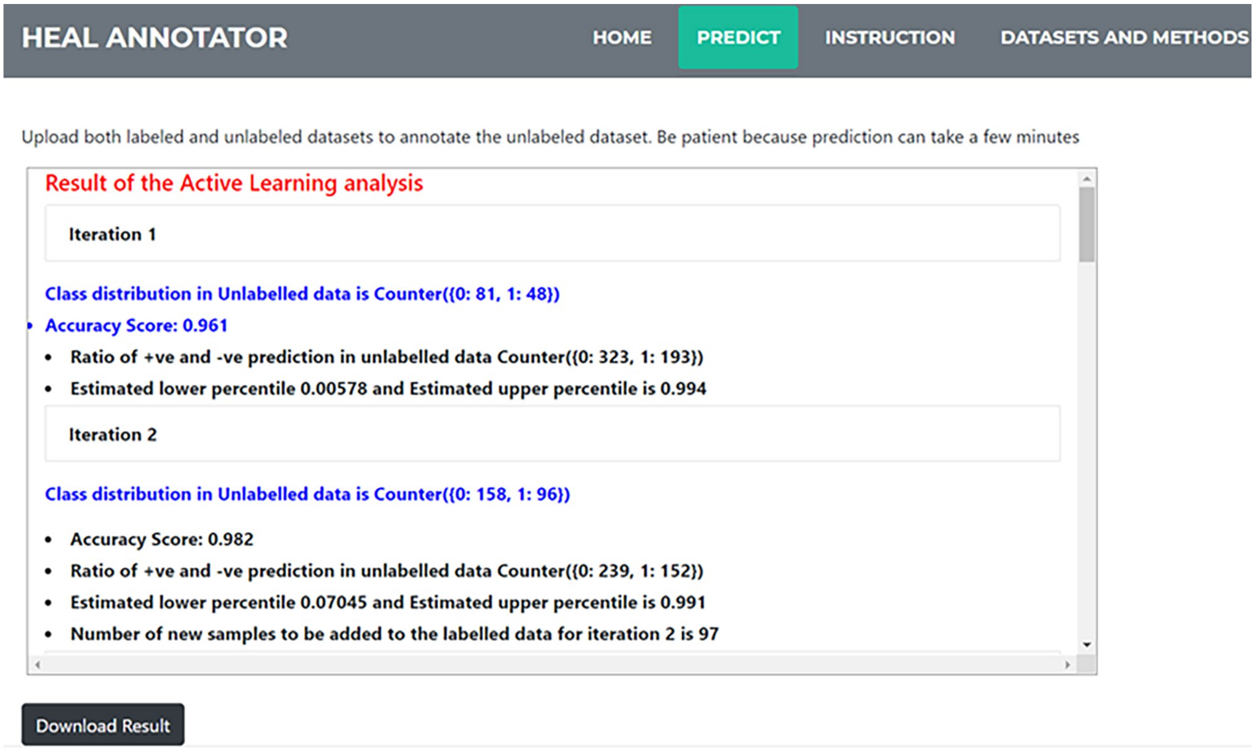
lab: cancer_labdata_sample.csv

unlab: cancer_unlabdata_sample.csv

Fig 5. Data description page of the HEAL annotator web application.

https://doi.org/10.1371/journal.pone.0288023.g005

labelled set and removed from the unlabelled set. The iteration was repeated until all samples in the unlabelled set have been completely labelled. In this study, heuristic-enabled active learning (HEAL) model, which replaces the human component of the traditional AL with a heuristic function, was developed. To benchmark the HEAL technique, the state-of-the-art AL



Upload both labeled and unlabeled datasets to annotate the unlabeled dataset. Be patient because prediction can take a few minutes

Result of the Active Learning analysis

Iteration 1

Class distribution in Unlabelled data is Counter({0: 81, 1: 48})

- Accuracy Score: 0.961
- Ratio of +ve and -ve prediction in unlabelled data Counter({0: 323, 1: 193})
- Estimated lower percentile 0.00578 and Estimated upper percentile is 0.994

Iteration 2

Class distribution in Unlabelled data is Counter({0: 158, 1: 96})

- Accuracy Score: 0.982
- Ratio of +ve and -ve prediction in unlabelled data Counter({0: 239, 1: 152})
- Estimated lower percentile 0.07045 and Estimated upper percentile is 0.991
- Number of new samples to be added to the labelled data for iteration 2 is 97

Fig 6. Result page of the HEAL annotator. User uploads their input files, and the program annotates the unlabelled dataset and presents the result to the user.

https://doi.org/10.1371/journal.pone.0288023.g006 technique that uses the uncertainty method (UncAL) and random sampling method (RandAL) were implemented. Five publicly available datasets were applied to the three techniques and the HEAL technique performs better when compared to the UncAL technique. The RandAL technique outputted the least performance which means that informed selection of samples from the unlabelled set to be added to the labelled set is critical for the good performance of the AL techniques. The *certainty* technique introduced by this study also seeks to increase the prediction power of the ML model by selecting samples with high confidence based on the pre-labelling by the base classifier to be added to the labelled set. The ambiguity at the decision boundary of the model is gradually resolved as the prediction power of the ML model increases. This accounts for the good performance recorded by HEAL technique.

In comparing the running time for the evaluated techniques, the HEAL technique showed a significantly reduced running time. The low running time recorded by HEAL is a result of the replacement of the human component from the AL process with a heuristic function. The human expert is required to manually go over all the selected pre-labelled samples and correct them one after the other which will be cumbersome if the size of the unlabelled set is large. Replacement of the human expert with the heuristic function provided by the HEAL technique also eliminates the financial cost associated with employing an expert for manual annotation thereby making it a preferred choice for AL techniques in future studies. During the development of HEAL, only binary classification was considered which implies that HEAL cannot be directly applied to a multiclass classification problem. HEAL performed well on mixed (Categorical, Boolean, and Continuous) data types.

The choice of training samples for conditional essentiality prediction is a potential limitation of this computational approach. The chances of getting sufficient positive and negative samples of a specific condition to train an active learning model are very low because most experimental studies focus on identifying "what genes do" and not "what they did not do" and the function of several genes are yet to be completely known. The use of wrong training data will affect the accuracy of the active learning model. When validated using Flybase, the proposed HEAL model performs significantly well in immune response condition compared to the embryonic developmental stage condition. The choice of features used to build the model determines how well the model performs in varying conditions.

In the following, we discuss some of the top-ranked genes predicted as essential for embryonic stage and immune response conditions in *D. melanogaster* along with their functions based on findings from our literature study. Ten proteins were predicted by HEAL as important for immune response. Six of these genes: Peptidoglycan recognition protein LC (*PGRP-LC*, FBgn0035976); Signal-transducer and activator of transcription protein at 92E (*Stat92E*, FBgn0016917); Suppressor of cytokine signaling at 36E (*Socs36E*, FBgn0041184); Domeless (*dome*, FBgn0043903); cactus (*cact*, FBgn0000250); Toll-7 (*Toll-7*, FBgn0034476) are already annotated as immune response genes in the literature [59–63]. For example, Peptidoglycan recognition protein LC (*PGRP-LC*, FBgn0035976) encodes a transmembrane receptor that is recognized and bounded to diaminopimelic acid (DAP)- containing peptidoglycan [64]. DAP- containing peptidoglycan is a cell wall component found on Gram-negative bacteria and certain Gram-positive bacteria. Its binding to *PGRP-LC* during bacterial infection activates the immune deficiency signalling pathway [65]. This leads to the induction of antibacterial genes and phagocytosis [60, 66]. Mutations in *PGRP-LC* leading to a loss of function increases susceptibility to gram-negative bacterial infection [67]. Knockdown of *PGRP-LC* also increased the copy number of sigma virus and reduced the survival rate of Drosophila infected with sigma virus when treated with CO2 [68], very likely to be linked to the cellular immune response to sigma virus infection. *Stat92E* is important in sustaining an effective balance between immune responses and also in inhibiting transcription of diverse immune effector genes activated by Relish. *Stat92E* mutant flies have been reported to have higher bacterial clearance activities compared to the wild type. However, these mutant flies die upon bacterial infection [69]. This reveals the importance of *Stat92E* in regulating a balanced immune response.

The four other genes predicted as immune response genes were 18 wheeler (*18w*, FBgn0004364 or FBgn0287775); necrotic (*nec*, FBgn0002930); TAK1-associated binding protein 2 (*Tab2*, FBgn0086358), and basket (*bsk*, FBgn00002290). The *18w* encodes a member of the Toll-like receptor family involved in antibacterial humoral response. *18w* mutant flies (larvae) have been reported to have reduced expression of antimicrobial peptide genes and suffered increased lethality upon bacterial challenge [70]. However, in adult flies, *18w* mutant flies had expressed antimicrobial peptide genes at levels similar to the wild type [71]. This suggests that the role of *18w* in immune response is age or developmental stage-specific. [72] reported increased expression levels of *18w* in 4-week-old flies infected with *E. coli* compared to their 1-week-old infected counterparts. Also, the transcript level of *18w* was significantly correlated (r = 0.80) with the ability of the flies to clear the bacteria compared to their 1-week-old counterparts. This further emphasizes the age-specific importance of *18w* in immune response to bacterial challenge.

Similarly, *nec*, encodes a hemolymphatic serine protease inhibitor (*serpin*, *spn*)—*Spn43Ac* that negatively regulates the Toll immune signalling pathway [73, 74]. The *nec* mutant flies constitutively express *Drosomycin*, in response to fungal infection [74]. *Tab2* participates in the activation of the immune deficiency (*Imd*) signalling pathway through its interaction with the product of transforming growth factor (*TGF*) beta-activated kinase 1 (*Tak1*) [75]. dsRNA silencing of *Tab2* has been noted to block expression of an antibacterial peptide produced by *Imd* activation, and *JNK* activation by peptidoglycans [76]. Likewise, *Tab2* RNAi eliminated the induction of a broad range of immune response genes in S2 Drosophila cells [77]. Further to this, *bsk* encodes a serine/threonine-protein kinase, a key component of the *JNK* signalling pathway. Drosophila *bsk* RNAi knockdown mutants have been reported to completely lack clot melanization [78]. Also, *bsk* mutant Drosophila larvae failed to melanise eggs from the parasitoid *Leptopilina boulardi* [79]. These studies showed *bsk* as an important mediator for cellular immune response through melanisation.

HEAL predicted 53 genes as essential for development. The top ten genes include Suppressor of zeste 2 (*Su(z)2*, FBgn0265623); Fatty acid synthase 1 (*Fasn1*, FBgn0283427); Fatty acid synthase 3 (*Fasn3*, FBgn0287184); Transitional elements of the endoplasmic reticulum 94 kDa (*Ter94*, FBgn0286784); Scab (*scb*, FBgn0286785); Hoi-polloi (*hoip*, FBgn0286786); myopic (*mop*, FBgn0036448); Zipper (*zip*, FBgn0265434 or FBgn0287873); Rho-related BTB domain containing (*RhoBTB*, FBgn0036980); Shibire (*shi*, FBgn0003392). These are discussed as follows:

Su(z)2 encodes a protein that is a functionally redundant homologue of the Polycomb Group (*PcG*) gene Posterior sex combs (*Psc*) protein [80]. *PcG* proteins are epigenetic regulators crucial in maintaining cell fate and stem cell function [81]. *Psc/Su(z)2* alongside Polyhomeotic (*PH*), Polycomb (*PC*), and *dRING* make up the Polycomb repressor complex 1 (*PRC1*) which play a role in ubiquitination of *H2A* [82]. *Su(z)2* restricts the proliferation and maintains the identity of the Cyst Stem Cell (*CySC*) in testis samples of Drosophila. It is also important for germline stem cell (GSC) maintenance and germ cell development, observed to act as a tumor suppressor [83]. *Su(z)2* disrupts dmyc auto-repression, Hence, it provides and maintains *Myc* levels required for embryonic growth and proliferation [84]. Similarly, [85] reported that only 17.4% of embryos from *Su(z)2* mutant flies emerged as adults compared to 91.5% adult emergence observed in wild type. These studies reveal the importance of *Su(Z)2* in the development of Drosophila from embryo to adults.

Fasn1 and *Fasn3* encode fatty acid synthase involved in the biosynthesis of saturated fatty acids [86]. [87] reported that *Fasn*–/–mutant mice embryos died before implantation and the *Fasn*+/–embryos died at various stages of their development, hence, the importance of *fasn* in embryonic development. In Drosophila, *Fasn1* levels have been observed to steadily increase during embryogenesis (peaks at 13.5–18 h), and then decline at the end of the embryonic life [88]. While *Fasn1* is present in all larvae tissues, *Fasn3* is expressed in the cuticle, epidermis, muscle, and oenocytes of larvae [89]. [90] noted that Drosophila with RNAi targeting *Fasn3* in their oenocytes, produced embryo that did not mature into adults. Lethality in the offspring was observed either at the second/third larval transition stage 4–5 days after egg deposition, at the third larval stage or at the pupa stage. However, flies with RNAi targeting *Fasn1* in oenocytes produced viable offspring. This might be due to an incomplete RNAi effect, although *Fasn3* is oenocyte-specific in adult flies but *Fasn1* is not [91]. These studies reveal the importance of fatty acid synthase 1 in the development of Drosophila.

Ter94 is a regulator of the ubiquitin proteasome system [92]. It is expressed in the embryo, in pupae, and in imago, but suppressed in the larvae stage of Drosophila [93]. Overexpression of *Ter94* RNAi in Drosophila third instar wing imaginal discs has been observed to cause pupal lethality [94]. Similarly, *Ter94* is important for oogenesis. [95] reported that embryos laid by female flies with germ-line clones of weak loss-of-function alleles of *Ter94* have reduced hatchability compared to the wild type. In turn, female flies with germ-line clones of a strong loss-of-function allele of *Ter94*, do not produce egg chambers [96]. *Ter94* regulates Bone morphogenetic proteins (BMPs) signalling during embryogenesis [97]. It also positively regulates Notch signalling [98]. Also, maternal knockdown of Ter94 caused significant 86% arrest in early stage 2 embryogenesis [99]. Hence, it is important for developmental events in the fly.

Scb encodes the α-PS3 Integrin. It regulates cell adhesion, signalling, polarity, and migration [100]. It is required for heart lumen formation [101]. S*cb* mutant flies have abnormal salivary glands, mislocalized pericardial cells and interrupted trachea [102]. It regulates pupal wing vein formation [103]. Also, mutations in *scb* reportedly resulted in impaired phagocytosis of apoptotic cells in Drosophila embryos [104]. These studies allude to the importance of *scb* during the development of Drosophila.

Hoip in Drosophila encodes a highly conserved RNA-binding protein [105]. *Hoip* mutant embryos have been reported to have aberrant myogenesis preventing them from emerging from the chorion after embryogenesis [106]. Hence, *hoip* is necessary for the initiation and maintenance of muscle structural gene expression during embryogenesis. Deficiency of *hoip* in mice has also been noted to cause embryonic lethality [107]. These studies portray *hoip* to be important during development in flies.

In turn, mop encodes a His domain protein-tyrosine phosphatase [108]. Depletion of mop impairs border cell cluster integrity and cell adhesion during oogenesis in Drosophila [109]. In Drosophila, *zip* encodes the non-muscle myosin II heavy chain. Zip mutant embryos have abnormal cell shape changes in the epidermis and incomplete dorsal closure [110]. Dorsal close in Drosophila embryo involves reorganization and contractions of the actin-myosin cytoskeleton within epithelial cells, thereby leading to the shaping of the embryo [111]. [112] reported that zip RNAi embryos had aberrant elongation (about 80% of zip RNAi embryo had $<$ 50% of egg length, compared to the wild type which all had $\geq$70% egg length). Similarly, none of the zip RNAi embryo hatched 1 day after compared to the wild type in which $>$80% of the embryos hatched. The study revealed that the absence of zip leads to embryonic lethality. Hence, it is essential for embryo development.

RhoBTB is an atypical Rho GTPase. It is important for dendritic development in Drosophila with its knockdown in dendritic arborization neurons leading to a reduced number of dendrites [113]. *Shi* is Drosophila's dynamin, a GTPase necessary for endocytosis and vesicle recycling [114, 115]. It regulates endocytosis throughout its development [116]. Temperature-sensitive *shi* mutants *shits1*, *shits3* and *shits6* have been reported to display phenotypes of embryonic lethality, continuous larval, and adult paralysis at 29˚C [117, 118]. Similarly, loss of function of *shi* results in disruption of the tracheal network with ectopic branching and misallocation of dorsal trunk cells, implicating *shi* in tracheal development [119].

In summary, HEAL was able to predict important genes involved in development or immune response conditions, which were not previously identified in Drosophila melanogaster. This discovery will provide more insight into the immune response factors and the growth mechanism in Drosophila. Furthermore, the success of the HEAL model has provided a viable solution to the challenge of limited class labelled data to train a ML classifier often encountered in bioinformatics predictive analysis.

Conclusion

We developed a heuristic-enabled active machine learning model that eliminates the human component in the active learning pipeline and possessing a superior prediction performance compared to the state-of-the-art AL models based on five public datasets. The HEAL model was implemented as a web tool for annotating biological and non-biological data when there is limited labelled data for training a machine learning model. The HEAL model was also applied to address the problem of predicting conditionally essential genes which is an intractable problem in bioinformatics. Essential immune response and embryonic developmental stage genes in *D. melanogaster* were predicted. Four of the 10 predicted immune response genes were novel and 53 genes were identified as important in the embryonic developmental stage in *D. melanogaster*. These predicted genes are proposed for future experimental studies.

Supporting information

S1 Fig. Confusion matrix from the comparative analysis of HEAL and other techniques on five real-world datasets. A. HEAL results show a significantly lower false-positive rate except on the cancer dataset. B. UncAL produces higher true positives except on the Cancer and Credit-A data. C. RandAL performed well on the Credit-G data with a higher true positive and lower false negative.
(DOCX)

S1 Table. The complete list of predicted genes that play essential roles during the embryonic stage of *D. melanogaster*.
(CSV)

Author Contributions

Conceptualization: Olufemi Tony Aromolaran, Marcus Oswald, Ezekiel Adebiyi, Rainer Koenig, Jelili Oyelade.

Data curation: Olufemi Tony Aromolaran, Eunice Adedeji, Rainer Koenig.

Formal analysis: Olufemi Tony Aromolaran, Itunu Isewon, Marcus Oswald, Rainer Koenig, Jelili Oyelade.

Funding acquisition: Ezekiel Adebiyi, Rainer Koenig.

Investigation: Jelili Oyelade.

Methodology: Olufemi Tony Aromolaran, Marcus Oswald, Ezekiel Adebiyi, Jelili Oyelade.

Project administration: Itunu Isewon, Ezekiel Adebiyi, Rainer Koenig, Jelili Oyelade.

Resources: Ezekiel Adebiyi, Rainer Koenig.

Software: Olufemi Tony Aromolaran.

Supervision: Rainer Koenig, Jelili Oyelade.

Validation: Olufemi Tony Aromolaran, Eunice Adedeji, Jelili Oyelade.

Visualization: Olufemi Tony Aromolaran, Jelili Oyelade.

Writing – original draft: Olufemi Tony Aromolaran, Itunu Isewon, Eunice Adedeji, Marcus Oswald, Ezekiel Adebiyi, Rainer Koenig, Jelili Oyelade.

Writing – review & editing: Olufemi Tony Aromolaran, Itunu Isewon, Eunice Adedeji, Marcus Oswald, Ezekiel Adebiyi, Rainer Koenig, Jelili Oyelade.

References

1. Wei W, Ning L-W, Ye Y-N, Guo F-B. Geptop: a gene essentiality prediction tool for sequenced bacterial genomes based on orthology and phylogeny. PLoS One 2013; 8:e72343. https://doi.org/10.1371/journal.pone.0072343 PMID: 23977285
2. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. Nat Protoc 2010; 5:93–121. https://doi.org/10.1038/nprot.2009.203 PMID: 20057383
3. Aromolaran O, Beder T, Oswald M, Oyelade J, Adebiyi E, Koenig R. Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features. Comput Struct Biotechnol J 2020; 18:612–21. https://doi.org/10.1016/j.csbj.2020.02.022 PMID: 32257045
4. Campos TL, Korhonen PK, Gasser RB, Young ND. An evaluation of machine learning approaches for the prediction of essential genes in eukaryotes using protein sequence-derived features. Comput Struct Biotechnol J 2019; 17:785–96. https://doi.org/10.1016/j.csbj.2019.05.008 PMID: 31312416
5. DeJesus MA, Ambadipudi C, Baker R, Sassetti C, Ioerger TR. TRANSIT-a software tool for Himar1 TnSeq analysis. PLoS Comput Biol 2015; 11:e1004401. https://doi.org/10.1371/journal.pcbi.1004401 PMID: 26447887
6. Saha S, Heber S. In silico prediction of yeast deletion phenotypes. Genet Mol Res 2006; 5:224–32. PMID: 16755513
7. Bosch-Guiteras N, van Leeuwen J. Exploring conditional gene essentiality through systems genetics approaches in yeast. Curr Opin Genet Dev 2022; 76:101963. https://doi.org/10.1016/j.gde.2022.101963 PMID: 35939967
8. Campos TL, Korhonen PK, Hofmann A, Gasser RB, Young ND. Combined use of feature engineering and machine-learning to predict essential genes in Drosophila melanogaster. NAR Genomics Bioinforma 2020; 2:lqaa051. https://doi.org/10.1093/nargab/lqaa051 PMID: 33575603
9. Campos TL, Korhonen PK, Hofmann A, Gasser RB, Young ND. Harnessing model organism genomics to underpin the machine learning-based prediction of essential genes in eukaryotes–Biotechnological implications. Biotechnol Adv 2022; 54:107822. https://doi.org/10.1016/j.biotechadv.2021.107822 PMID: 34461202
10. Costanzo M, Hou J, Messier V, Nelson J, Rahman M, VanderSluis B, et al. Environmental robustness of the global yeast genetic interaction network. Science (80-) 2021; 372:eabf8424. https://doi.org/10.1126/science.abf8424 PMID: 33958448
11. Hou J, Tan G, Fink GR, Andrews BJ, Boone C. Complex modifier landscape underlying genetic background effects. Proc Natl Acad Sci 2019; 116:5045–54. https://doi.org/10.1073/pnas.1820915116 PMID: 30804202
12. Dowell RD, Ryan O, Jansen A, Cheung D, Agarwala S, Danford T, et al. Genotype to phenotype: a complex problem. Science (80-) 2010;328:469. https://doi.org/10.1126/science.1189015 PMID: 20413493
13. Larrimore KE, Rancati G. The conditional nature of gene essentiality. Curr Opin Genet Dev 2019; 58:55–61. https://doi.org/10.1016/j.gde.2019.07.015 PMID: 31470233
14. Thurmond J, Goodman JL, Strelets VB, Attrill H, Gramates LS, Marygold SJ, et al. FlyBase 2.0: the next generation. Nucleic Acids Res 2019; 47:D759–65. https://doi.org/10.1093/nar/gky1003 PMID: 30364959

15. Myllymäki H, Valanne S, Rämet M. The Drosophila imd signaling pathway. J Immunol 2014; 192:3455–62. https://doi.org/10.4049/jimmunol.1303309 PMID: 24706930

16. Bonagura VR, Rosenthal DW. Infections that cause secondary immune deficiency. Stiehm's immune Defic., Elsevier; 2020, p. 1035–58.

17. World Health Organization. WHO World Malaria Report 2020. 2020.

18. Basler G. Computational prediction of essential metabolic genes using constraint-based approaches. Gene Essentiality, New York, NY: Humana Press; 2015, p. 183–204.

19. Mobegi FM, Zomer A, De Jonge MI, Van Hijum SAFT. Advances and perspectives in computational prediction of microbial gene essentiality. Brief Funct Genomics 2017; 16:70–9.

20. Peng C, Lin Y, Luo H, Gao F. A comprehensive overview of online resources to identify and predict bacterial essential genes. Front Microbiol 2017; 8:2331. https://doi.org/10.3389/fmicb.2017.02331 PMID: 29230204

21. Zhao L, Anderson MT, Wu W, Mobley HLT, Bachman MA. TnseqDiff: identification of conditionally essential genes in transposon sequencing studies. BMC Bioinformatics 2017; 18:1–11.

22. Younes S, Al-Sulaiti A, Nasser EAA, Najjar H, Kamareddine L. Drosophila as a Model Organism in Host–Pathogen Interaction Studies. Front Cell Infect Microbiol 2020;10.

23. Akimana C, Al-Khodor S, Kwaik YA. Host factors required for modulation of phagosome biogenesis and proliferation of Francisella tularensis within the cytosol. PLoS One 2010; 5:e11025. https://doi.org/10.1371/journal.pone.0011025 PMID: 20552012

24. Ragab A, Buechling T, Gesellchen V, Spirohn K, Boettcher A, Boutros M. Drosophila Ras/MAPK signalling regulates innate immune responses in immune and intestinal stem cells. EMBO J 2011; 30:1123–36. https://doi.org/10.1038/emboj.2011.4 PMID: 21297578

25. Manimaran P, Hegde SR, Mande SC. Prediction of conditional gene essentiality through graph theoretical analysis of genome-wide functional linkages. Mol Biosyst 2009; 5:1936–42. https://doi.org/10.1039/B905264j PMID: 19763329

26. Ahmad K, Mekhalfi ML, Conci N. Event recognition in personal photo collections: An active learning approach. IS&T Int Symp Electron Imaging 2018; 2018:171–3.

27. Hossain HMS, Khan MAAH, Roy N. Active learning enabled activity recognition. Pervasive Mob Comput 2017; 38:312–30.

28. Miller B, Linder F, Mebane WR Jr. Active Learning Approaches for Labeling Text. Technical report, University of Michigan, Ann Arbor, MI; 2018.

29. Tuia D, Volpi M, Copa L, Kanevski M, Munoz-Mari J. A survey of active learning algorithms for supervised remote sensing image classification. IEEE J Sel Top Signal Process 2011; 5:606–17.

30. Baur T, Heimerl A, Lingenfelser F, Wagner J, Valstar MF, Schuller B, et al. Explainable cooperative machine learning with NOVA. KI-Künstliche Intelligenz 2020:1–22.

31. Wang G, Hwang J-N, Rose C, Wallace F. Uncertainty-based active learning via sparse modeling for image classification. IEEE Trans Image Process 2018; 28:316–29. https://doi.org/10.1109/TIP.2018.2867913 PMID: 30176591

32. Burbidge R, Rowland JJ, King RD. Active learning for regression based on query by committee. Int. Conf. Intell. data Eng. Autom. Learn., Springer; 2007, p. 209–18.

33. Freund Y, Seung HS, Shamir E, Tishby N. Selective sampling using the query by committee algorithm. Mach Learn 1997; 28:133–68.

34. Fu W, Wang M, Hao S, Wu X. Scalable active learning by approximated error reduction. Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discov. data Min., 2018, p. 1396–405.

35. Asuncion A, Newman D. UCI machine learning repository 2007. https://doi.org/ http://archive.ics.uci.edu/ml.

36. Aromolaran O, Aromolaran D, Isewon I, Oyelade J. Machine learning approach to gene essentiality prediction: a review. Brief Bioinform 2021; 22:bbab128. https://doi.org/10.1093/bib/bbab128 PMID: 33842944

37. Yates AD, Achuthan P, Akanni W, Allen J, Allen J, Alvarez-Jarreta J, et al. Ensembl 2020. Nucleic Acids Res 2020; 48:D682–8.

38. Howe KL, Contreras-Moreira B, De Silva N, Maslen G, Akanni W, Allen J, et al. Ensembl Genomes 2020—enabling non-vertebrate genomic research. Nucleic Acids Res 2020; 48:D689–95. https://doi.org/10.1093/nar/gkz890 PMID: 31598706

39. Smedley D, Haider S, Ballester B, Holland R, London D, Thorisson G, et al. BioMart–biological queries made easy. BMC Genomics 2009; 10:1–12.

40. Charif D, Lobry JR. SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. Struct. approaches to Seq. Evol., Springer; 2007, p. 207–32.

41. Xiao N, Cao D-S, Zhu M-F, Xu Q-S. protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences. Bioinformatics 2015; 31:1857–9. https://doi.org/10.1093/bioinformatics/btv042 PMID: 25619996

42. Peden J. CodonW. Univ Nottingham 1997.

43. Zhu M, Dong J, Cao D-S. rDNAse: R package for generating various numerical representation schemes of DNA sequences 2016.

44. Hershberg R, Petrov DA. General rules for optimal codon choice. PLoS Genet 2009; 5:e1000556. https://doi.org/10.1371/journal.pgen.1000556 PMID: 19593368

45. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, et al. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res 2019; 47:D607–13. https://doi.org/10.1093/nar/gky1131 PMID: 30476243

46. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2007; 35:D61–5. https://doi.org/10.1093/nar/gkl842 PMID: 17130148

47. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997; 25:3389–402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694

48. Hagberg A, Swart P, S Chult D. Exploring network structure, dynamics, and function using NetworkX. Los Alamos National Lab.(LANL), Los Alamos, NM (United States); 2008.

49. Boldi P, Vigna S. Axioms for centrality. Internet Math 2014; 10:222–62.

50. Almagro Armenteros JJ, Sønderby CK, Sønderby SK, Nielsen H, Winther O. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics 2017; 33:3387–95. https://doi.org/10.1093/bioinformatics/btx431 PMID: 29036616

51. Dawoud I. A new improved estimator for reducing the multicollinearity effects. Commun Stat—Simul Comput 2021:1–12. https://doi.org/10.1080/03610918.2021.1939374.

52. Kim JH. Multicollinearity and misleading statistical results. Korean J Anesthesiol 2019; 72:558. https://doi.org/10.4097/kja.19087 PMID: 31304696

53. Chawla N V, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002; 16:321–57.

54. Breiman L. Random forests. Mach Learn 2001; 45:5–32.

55. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst 2017; 30:3146–54.

56. Aromolaran O, Beder T, Adedeji E, Ajamma Y, Oyelade J, Adebiyi E, et al. Predicting host dependency factors of pathogens in Drosophila melanogaster using machine learning. Comput Struct Biotechnol J 2021; 19:4581–92. https://doi.org/10.1016/j.csbj.2021.08.010 PMID: 34471501

57. Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic Acids Res 2019; 47: W191–8.

58. Murali TM, Dyer MD, Badger D, Tyler BM, Katze MG. Network-based prediction and analysis of HIV dependency factors. PLoS Comput Biol 2011; 7:e1002164. https://doi.org/10.1371/journal.pcbi.1002164 PMID: 21966263

59. Agaisse H, Petersen U-M, Boutros M, Mathey-Prevot B, Perrimon N. Signaling role of hemocytes in Drosophila JAK/STAT-dependent response to septic injury. Dev Cell 2003; 5:441–50. https://doi.org/10.1016/s1534-5807(03)00244-2 PMID: 12967563

60. Choe K-M, Werner T, Stoven S, Hultmark D, Anderson K V. Requirement for a peptidoglycan recognition protein (PGRP) in Relish activation and antibacterial immune responses in Drosophila. Science (80-) 2002; 296:359–62. https://doi.org/10.1126/science.1070216 PMID: 11872802

61. Chowdhury M, Li C-F, He Z, Lu Y, Liu X-S, Wang Y-F, et al. Toll family members bind multiple Spätzle proteins and activate antimicrobial peptide gene expression in Drosophila. J Biol Chem 2019; 294:10172–81.

62. Nakamoto M, Moy RH, Xu J, Bambina S, Yasunaga A, Shelly SS, et al. Virus recognition by Toll-7 activates antiviral autophagy in Drosophila. Immunity 2012; 36:658–67. https://doi.org/10.1016/j.immuni.2012.03.003 PMID: 22464169

63. Nicolas E, Reichhart JM, Hoffmann JA, Lemaitre B. In vivo regulation of the IκB homologue cactus during the immune response of Drosophila. J Biol Chem 1998; 273:10463–9.

64. Kaneko T, Yano T, Aggarwal K, Lim J-H, Ueda K, Oshima Y, et al. PGRP-LC and PGRP-LE have essential yet distinct functions in the drosophila immune response to monomeric DAP-type peptidoglycan. Nat Immunol 2006; 7:715–23. https://doi.org/10.1038/ni1356 PMID: 16767093

65. Gottar M, Gobert V, Michel T, Belvin M, Duyk G, Hoffmann JA, et al. The Drosophila immune response against Gram-negative bacteria is mediated by a peptidoglycan recognition protein. Nature 2002; 416:640–4. https://doi.org/10.1038/nature734 PMID: 11912488

66. Rämet M, Manfruelli P, Pearson A, Mathey-Prevot B, Ezekowitz RAB. Functional genomic analysis of phagocytosis and identification of a Drosophila receptor for E. coli. Nature 2002; 416:644–8. https://doi.org/10.1038/nature735 PMID: 11912489

67. Takehana A, Yano T, Mita S, Kotani A, Oshima Y, Kurata S. Peptidoglycan recognition protein (PGRP)-LE and PGRP-LC act synergistically in Drosophila immunity. EMBO J 2004; 23:4690–700. https://doi.org/10.1038/sj.emboj.7600466 PMID: 15538387

68. Liao J-F, Wu C-P, Tang C-K, Tsai C-W, Rouhová L, Wu Y-L. Identification of regulatory host genes involved in sigma virus replication using RNAi knockdown in Drosophila. Insects 2019; 10:339. https://doi.org/10.3390/insects10100339 PMID: 31614679

69. Kim LK, Choi UY, Cho HS, Lee JS, Lee W, Kim J, et al. Down-regulation of NF-κB target genes by the AP-1 and STAT complex during the innate immune response in Drosophila. PLoS Biol 2007; 5:e238.

70. Williams MJ, Rodriguez A, Kimbrell DA, Eldon ED. The 18-wheeler mutation reveals complex antibacterial gene regulation in Drosophila host defense. EMBO J 1997; 16:6120–30. https://doi.org/10.1093/emboj/16.20.6120 PMID: 9321392

71. Ligoxygakis P, Bulet P, Reichhart J-M. Critical evaluation of the role of the Toll-like receptor 18-Wheeler in the host defense of Drosophila. EMBO Rep 2002; 3:666–73. https://doi.org/10.1093/embo-reports/kvf130 PMID: 12101100

72. Felix TM, Hughes KA, Stone EA, Drnevich JM, Leips J. Age-specific variation in immune response in Drosophila melanogaster has a genetic basis. Genetics 2012; 191:989–1002. https://doi.org/10.1534/genetics.112.140640 PMID: 22554890

73. Green C, Levashina E, McKimmie C, Dafforn T, Reichhart J-M, Gubb D. The necrotic gene in Drosophila corresponds to one of a cluster of three serpin transcripts mapping at 43A1. 2. Genetics 2000; 156:1117–27. https://doi.org/10.1093/genetics/156.3.1117 PMID: 11063688

74. Levashina EA, Langley E, Green C, Gubb D, Ashburner M, Hoffmann JA, et al. Constitutive activation of toll-mediated antifungal defense in serpin-deficient Drosophila. Science (80-) 1999; 285:1917–9. https://doi.org/10.1126/science.285.5435.1917 PMID: 10489372

75. Kleino A, Valanne S, Ulvila J, Kallio J, Myllymäki H, Enwald H, et al. Inhibitor of apoptosis 2 and TAK1-binding protein are components of the Drosophila Imd pathway. EMBO J 2005; 24:3423–34. https://doi.org/10.1038/sj.emboj.7600807 PMID: 16163390

76. Zhuang Z-H, Sun L, Kong L, Hu J-H, Yu M-C, Reinach P, et al. Drosophila TAB2 is required for the immune activation of JNK and NF-kappaB. Cell Signal 2006; 18:964–70. https://doi.org/10.1016/j.cellsig.2005.08.020 PMID: 16311020

77. Valanne S, Kleino A, Myllymäki H, Vuoristo J, Rämet M. Iap2 is required for a sustained response in the Drosophila Imd pathway. Dev Comp Immunol 2007; 31:991–1001. https://doi.org/10.1016/j.dci.2007.01.004 PMID: 17343912

78. Bidla G, Dushay MS, Theopold U. Crystal cell rupture after injury in Drosophila requires the JNK pathway, small GTPases and the TNF homolog Eiger. J Cell Sci 2007; 120:1209–15. https://doi.org/10.1242/jcs.03420 PMID: 17356067

79. Williams MJ, Wiklund M-L, Wikman S, Hultmark D. Rac1 signalling in the Drosophila larval cellular immune response. J Cell Sci 2006; 119:2015–24. https://doi.org/10.1242/jcs.02920 PMID: 16621891

80. Lo SM, Ahuja NK, Francis NJ. Polycomb group protein Suppressor 2 of zeste is a functional homolog of Posterior Sex Combs. Mol Cell Biol 2009; 29:515–25. https://doi.org/10.1128/MCB.01044-08 PMID: 18981224

81. Polycomb Orlando V., epigenomes, and control of cell identity. Cell 2003; 112:599–606.

82. Francis NJ, Saurin AJ, Shao Z, Kingston RE. Reconstitution of a functional core polycomb repressive complex. Mol Cell 2001; 8:545–56. https://doi.org/10.1016/s1097-2765(01)00316-1 PMID: 11583617

83. Morillo Prado JR, Chen X, Fuller MT. Polycomb group genes Psc and Su (z) 2 maintain somatic stem cell identity and activity in Drosophila. PLoS One 2012; 7:e52892. https://doi.org/10.1371/journal.pone.0052892 PMID: 23285219

84. Khan A, Shover W, Goodliffe JM. Su (z) 2 antagonizes auto-repression of Myc in Drosophila, increasing Myc levels and subsequent trans-activation. PLoS One 2009; 4:e5076.

85. Dasari V, Srivastava S, Khan S, Mishra RK. Epigenetic factors Polycomb (Pc) and Suppressor of zeste (Su (z) 2) negatively regulate longevity in Drosophila melanogaster. Biogerontology 2018; 19:33–45. https://doi.org/10.1007/s10522-017-9737-1 PMID: 29177687

86. Vrablik TL, Watts JL. Emerging roles for specific fatty acids in developmental processes. Genes Dev 2012; 26:631–7. https://doi.org/10.1101/gad.190777.112 PMID: 22474257

87. Chirala SS, Chang H, Matzuk M, Abu-Elheiga L, Mao J, Mahon K, et al. Fatty acid synthesis is essential in embryonic development: fatty acid synthase null mutants and most of the heterozygotes die in utero. Proc Natl Acad Sci 2003; 100:6358–63. https://doi.org/10.1073/pnas.0931394100 PMID: 12738878

88. Fabre B, Korona D, Groen A, Vowinckel J, Gatto L, Deery MJ, et al. Analysis of Drosophila melanogaster proteome dynamics during embryonic development by a combination of label-free proteomics approaches. Proteomics 2016; 16:2068–80. https://doi.org/10.1002/pmic.201500482 PMID: 27029218

89. Chintapalli VR, Wang J, Dow JAT. Using FlyAtlas to identify better Drosophila melanogaster models of human disease. Nat Genet 2007; 39:715–20. https://doi.org/10.1038/ng2049 PMID: 17534367

90. Parvy J-P, Napal L, Rubin T, Poidevin M, Perrin L, Wicker-Thomas C, et al. Drosophila melanogaster acetyl-CoA-carboxylase sustains a fatty acid–dependent remote signal to waterproof the respiratory system 2012.

91. Ghosh AC, Tattikota SG, Liu Y, Comjean A, Hu Y, Barrera V, et al. Drosophila PDGF/VEGF signaling from muscles to hepatocyte-like cells protects against obesity. Elife 2020; 9:e56969. https://doi.org/10.7554/eLife.56969 PMID: 33107824

92. Santhanam A, Peng W-H, Yu Y-T, Sang T-K, Chen G-C, Meng T-C. Ecdysone-induced receptor tyrosine phosphatase PTP52F regulates Drosophila midgut histolysis by enhancement of autophagy and apoptosis. Mol Cell Biol 2014; 34:1594–606. https://doi.org/10.1128/MCB.01391-13 PMID: 24550005

93. Pintér M, Jékely G, Szepesi RJ, Farkas A, Theopold U, Meyer HE, et al. TER94, a Drosophila homolog of the membrane fusion protein CDC48/p97, is accumulated in nonproliferating cells: in the reproductive organs and in the brain of the imago. Insect Biochem Mol Biol 1998; 28:91–8. https://doi.org/10.1016/s0965-1748(97)00095-7 PMID: 9639875

94. Reim G, Hruzova M, Goetze S, Basler K. Protection of armadillo/β-Catenin by armless, a novel positive regulator of wingless signaling. PLoS Biol 2014; 12:e1001988.

95. Ruden DM, Sollars V, Wang X, Mori D, Alterman M, Lu X. Membrane fusion proteins are required for oskar mRNA localization in the Drosophila egg chamber. Dev Biol 2000; 218:314–25. https://doi.org/10.1006/dbio.1999.9583 PMID: 10656772

96. León A, McKearin D. Identification of TER94, an AAA ATPase protein, as a Bam-dependent component of the Drosophila fusome. Mol Biol Cell 1999; 10:3825–34. https://doi.org/10.1091/mbc.10.11.3825 PMID: 10564274

97. Zeng Z, de Gorter DJJ, Kowalski M, ten Dijke P, Shimmi O. Ter94/VCP is a novel component involved in BMP signaling. PLoS One 2014; 9:e114475. https://doi.org/10.1371/journal.pone.0114475 PMID: 25469707

98. Li Y, Liu T, Zhang J. The ATPase TER94 regulates Notch signaling during Drosophila wing development. Biol Open 2019; 8:bio038984. https://doi.org/10.1242/bio.038984 PMID: 30530809

99. Zhang Z, Krauchunas AR, Huang S, Wolfner MF. Maternal proteins that are phosphoregulated upon egg activation include crucial factors for oogenesis, egg activation and embryogenesis in Drosophila melanogaster. G3 Genes, Genomes, Genet 2018; 8:3005–18. https://doi.org/10.1534/g3.118.200578 PMID: 30012668

100. Dinkins MB, Fratto VM, LeMosy EK. Integrin alpha chains exhibit distinct temporal and spatial localization patterns in epithelial cells of the Drosophila ovary. Dev Dyn an Off Publ Am Assoc Anat 2008; 237:3927–39. https://doi.org/10.1002/dvdy.21802 PMID: 19035354

101. Vanderploeg J, Vazquez Paz LL, MacMullin A, Jacobs JR. Integrins are required for cardioblast polarisation in Drosophila. BMC Dev Biol 2012; 12:1–12.

102. Stark KA, Yee GH, Roote CE, Williams EL, Zusman S, Hynes RO. A novel alpha integrin subunit associates with betaPS and functions in tissue morphogenesis and movement during Drosophila development. Development 1997; 124:4583–94. https://doi.org/10.1242/dev.124.22.4583 PMID: 9409675

103. Araujo H, Negreiros E, Bier E. Integrins modulate Sog activity in the Drosophila wing 2003.

104. Nonaka S, Nagaosa K, Mori T, Shiratsuchi A, Nakanishi Y. Integrin αPS3/βv-mediated phagocytosis of apoptotic cells and bacteria in Drosophila. J Biol Chem 2013; 288:10374–80.

105. Williams J, Boin NG, Valera JM, Johnson AN. Noncanonical roles for Tropomyosin during myogenesis. Development 2015; 142:3440–52. https://doi.org/10.1242/dev.117051 PMID: 26293307

106. Johnson AN, Mokalled MH, Valera JM, Poss KD, Olson EN. Post-transcriptional regulation of myotube elongation and myogenesis by Hoi Polloi. Development 2013; 140:3645–56. https://doi.org/10.1242/dev.095596 PMID: 23942517

107. Peltzer N, Rieser E, Taraborrelli L, Draber P, Darding M, Pernaute B, et al. HOIP deficiency causes embryonic lethality by aberrant TNFR1-mediated endothelial cell death. Cell Rep 2014; 9:153–65. https://doi.org/10.1016/j.celrep.2014.08.066 PMID: 25284787

108. Jia D, Soylemez M, Calvin G, Bornmann R, Bryant J, Hanna C, et al. A large-scale in vivo RNAi screen to identify genes involved in Notch-mediated follicle cell differentiation and cell cycle switches. Sci Rep 2015; 5:1–14. https://doi.org/10.1038/srep12328 PMID: 26205122

109. Chen D-Y, Li M-Y, Wu S-Y, Lin Y-L, Tsai S-P, Lai P-L, et al. The Bro1-domain-containing protein Myopic/HDPTP coordinates with Rab4 to regulate cell adhesion and migration. J Cell Sci 2012; 125:4841–52. https://doi.org/10.1242/jcs.108597 PMID: 22825871

110. Young PE, Richman AM, Ketchum AS, Kiehart DP. Morphogenesis in Drosophila requires nonmuscle myosin heavy chain function. Genes Dev 1993; 7:29–41. https://doi.org/10.1101/gad.7.1.29 PMID: 8422986

111. Jacinto A, Wood W, Woolner S, Hiley C, Turner L, Wilson C, et al. Dynamic analysis of actin cable function during Drosophila dorsal closure. Curr Biol 2002; 12:1245–50. https://doi.org/10.1016/s0960-9822(02)00955-7 PMID: 12176336

112. Kasza KE, Supriyatno S, Zallen JA. Cellular defects resulting from disease-related myosin II mutations in Drosophila. Proc Natl Acad Sci 2019; 116:22205–11. https://doi.org/10.1073/pnas.1909227116 PMID: 31615886

113. Straub J, Konrad EDH, Grüner J, Toutain A, Bok LA, Cho MT, et al. Missense variants in RHOBTB2 cause a developmental and epileptic encephalopathy in humans, and altered levels cause neurological defects in Drosophila. Am J Hum Genet 2018; 102:44–57. https://doi.org/10.1016/j.ajhg.2017.11.008 PMID: 29276004

114. Chen MS, Obar RA, Schroeder CC, Austin TW, Poodry CA, Wadsworth SC, et al. Multiple forms of dynamin are encoded by shibire, a Drosophila gene involved in endocytosis. Nature 1991; 351:583–6. https://doi.org/10.1038/351583a0 PMID: 1828536

115. Van der Bliek AM, Meyerowrtz EM. Dynamin-like protein encoded by the Drosophila shibire gene associated with vesicular traffic. Nature 1991; 351:411–4. https://doi.org/10.1038/351411a0 PMID: 1674590

116. Peters NC, Thayer NH, Kerr SA, Tompa M, Berg CA. Following the 'tracks': Tramtrack69 regulates epithelial tube expansion in the Drosophila ovary through Paxillin, Dynamin, and the homeobox protein Mirror. Dev Biol 2013; 378:154–69. https://doi.org/10.1016/j.ydbio.2013.03.017 PMID: 23545328

117. Grigliatti TA, Hall L, Rosenbluth R, Suzuki DT. Temperature-sensitive mutations in Drosophila melanogaster. Mol Gen Genet MGG 1973; 120:107–14.

118. Poodry CA, Hall L, Suzuki DT. Developmental properties of shibirets1: A pleiotropic mutation affecting larval and adult locomotion and development. Dev Biol 1973; 32:373–86.

119. Dammai V, Adryan B, Lavenburg KR, Hsu T. Drosophila awd, the homolog of human nm23, regulates FGF receptor levels and functions synergistically with shi/dynamin during tracheal development. Genes Dev 2003; 17:2812–24. https://doi.org/10.1101/gad.1096903 PMID: 14630942