PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

AWGE-ESPCA: An edge sparse PCA model based on adaptive noise elimination regularization and weighted gene network for *Hermetia illucens* genomic data analysis

Rui Miao☯, Hao-Yang Yu☯, Bing-Jie Zhong, Hong-Xia Sun, Qiang Xia*

Basic Teaching Department, Zhuhai Campus of Zunyi Medical University, Zhu Hai, China

☯ These authors are co-first authors who contributed equally to this work.
* xiaqiang1973@126.com

Citation: Miao R, Yu H-Y, Zhong B-J, Sun H-X, Xia Q (2025) AWGE-ESPCA: An edge sparse PCA model based on adaptive noise elimination regularization and weighted gene network for *Hermetia illucens* genomic data analysis. PLoS Comput Biol 21(2): e1012773. https://doi.org/10.1371/journal.pcbi.1012773

Editor: Dexiong Chen, Max-Planck-Institute of Biochemistry: Max-Planck-Institut fur Biochemie, GERMANY

Received: June 28, 2024

Accepted: January 6, 2025

Published: February 13, 2025

Data availability statement: All data and code used for running experiments, model fitting, and plotting is available on a GitHub repository at https://github.com/yhyresearcher/AWGE_ESPCA.

Funding: This work was supported by the Macau Science and Technology Development

Abstract

Hermetia illucens is an important insect resource. Studies have shown that exploring the effects of Cu^{2+}-stressed on the growth and development of the *Hermetia illucens* genome holds significant scientific importance. There are three major challenges in the current studies of *Hermetia illucens* genomic data analysis: firstly, the lack of available genomic data which limits researchers in *Hermetia illucens* genomic data analysis. Secondly, to the best of our knowledge, there are no Artificial Intelligence (AI) feature selection models designed specifically for *Hermetia illucens* genome. Unlike human genomic data, noise in *Hermetia illucens* data is a more serious problem. Third, how to choose those genes located in the pathway enrichment region. Existing models assume that each gene probe has the same priori weight. However, researchers usually pay more attention to gene probes which are in the pathway enrichment region. Based on the above challenges, we initially construct experiments and establish a new Cu^{2+}-stressed *Hermetia illucens* growth genome dataset. Subsequently, we propose AWGE-ESPCA: an edge Sparse PCA model based on adaptive noise elimination regularization and weighted gene network. The AWGE-ESPCA model innovatively proposes an adaptive noise elimination regularization method, effectively addressing the noise challenge in *Hermetia illucens* genomic data. We also integrate the known gene-pathway quantitative information into the Sparse PCA(SPCA) framework as a priori knowledge, which allows the model to filter out the gene probes in pathway-rich regions as much as possible. Ultimately, this study conducts five independent experiments and compared four latest Sparse PCA models as well as representative supervised and unsupervised baseline models to validate the model performance. The experimental results demonstrate the superior pathway and gene selection capabilities of the AWGE-ESPCA model. Ablation experiments validate the role of the adaptive regularizer and network weighting module. To summarize, this paper presents an innovative unsupervised model for *Hermetia illucens* genome analysis, which can effectively help researchers identify potential biomarkers. In addition, we also provide a working AWGE - ESPCA model code in the address: https://github.com/yhyresearcher/AWGE_ESPCA.

Fund (No. 0158/2019/A3), the Guangdong Provincial Department of Education youth innovative talent project Fund (No. 2023KQNCX155 to RM), the Postdoctoral training project of Zunyi Medical University Fund (No. 2023F-ZH-019 to RM), the National Natural Science Foundation of China (No. 32060127; No. 31260528), and the Key Construction Discipline of Immunology and Pathogen Biology Fund, Zunyi Medical University Zhuhai Campus, China (No. ZHGF2024-1). 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.

Author summary

Hermetia illucens is an insect of high economic value, which is widely used in the field of feed. Existing research suggests that Cu^{2+}-stressed can significantly affects the growth of *Hermetia illucens*. Therefore, the identification of genetic target information affecting the growth and development of *Hermetia illucens* is crucial for food safety. However, due to the lack of high-quality data sets, high data noisy and low sample number. None of the existing genomic analysis models can handle the *Hermetia illucens* data well. Based on the above problems, a novel unsupervised *Hermetia illucens* genomic analysis model (AWGE-ESPCA) is proposed in this paper. The AWGE-ESPCA model proposes a daptive noise elimination regularization to solve noise challenges in data and uses weighted gene network to enhance the biological interpretability capability of the model. The experimental results show that the AWGE-ESPCA model can well select potential target genes and key pathways. In addition, we demonstrate that the AWGE-ESPCA model can be extended to other insect genome analysis tasks.

1. Introduction

Hermetia illucens is a globally important resource insect that plays a great role in many fields [1–3]. For example, *Hermetia illucens* larvae feed on decaying organic matter and animal feces, and are widely used in various countries in the field of environmentally sound treatment. In addition, its by-products are rich in nutrients and antimicrobial peptides, which have a wide range of applications in the fields of feed and medicine [4,5]. Studies have shown that high concentration of Cu^{2+}-stressed not only seriously affects the growth and development of *Hermetia illucens* [6,7], but also reduces the content of nutrients such as total sugars and proteins in the hemolymph of the larvae [6,7], thus decreasing its value in feed and medicine. More importantly, if the concentration of Cu^{2+}-stressed in the feed of *Hermetia illucens* is too high, it may also poses a risk to human beings through bioconcentration [8,9]. Therefore, studying the effects of Cu^{2+}-stressed in the growth and development of *Hermetia illucens* can improve its application value in environmental harmless treatment, animal feed safety and medicine [10–12].

Currently, there are three major challenges for the Cu^{2+}-stressed *Hermetia illucens* genome analysis [13]: Firstly, there are few available data on Cu^{2+}-stressed *Hermetia illucens* [14]. Although some genomic datasets of Cu^{2+}-stressed *Hermetia illucens* growth have been constructed by researchers [14,15]. However, the sample sizes of the existing datasets are very small, and no researchers have constructed genomic datasets of *Hermetia illucens* under different Cu^{2+}-stressed environments. It prevents researchers from building high-performance models for *Hermetia illucens* genomics analysis. Secondly, to the best of our knowledge, there are still no AI feature selection models specifically for insect genomes. The noise problem of insect genomic data is even more serious compared to human genomic data [13,16]. In the case of *Hermetia illucens* data, the dataset contains many gene probes with excessive expression values. These gene probes are not relevant to sample classification but can significantly image the feature selection results of the AI model. Therefore, it is essential to construct an AI feature selection model specifically designed to analyze insect genomic data and address the noise challenge of the data. Thirdly, choose those genes located in the pathway enrichment region. The existing models assume that each gene probe has the same priori weight. Based on biological common sense, it is obvious that researchers will be more

concerned with regions that are more enriched in biological pathways [17,18]. In other words, gene probes that are associated with a larger number of gene pathways should obviously receive more attention. This is because these gene probes may play more important roles in the upstream and downstream processes of biological actions [19]. Therefore, how to rationally weight the gene probes in the model based on the known information about the number of pathways, so as to filter out the pathway-enriched regions is the third major challenge currently faced.

Based on the above challenges, we design experiments and build a Cu^{2+}-stressed *Hermetia illucens* genomics dataset. Then, we propose AWGE-ESPCA: An Edge Sparse PCA Model Based on Adaptive Noise Elimination Regularization and Weighted Gene Network (Fig 1). AWGE-ESPCA proposes an Adaptive Noise Elimination Regularizer for solving the noise problem of *Hermetia illucens* genome data. The AWGE-ESPCA model also uses the known gene-pathway quantity information as a priori knowledge and integrates it into the gene network of the Sparse PCA model. This weighted gene network allows the model to pay more attention to the pathway-rich regions in the gene network, and these gene probes are often the key sites in the pathway.

In this paper, we conduct five independent experiments. Firstly, we set up a noisy dataset for simulation experiments to verify the noise resistance of AWGE-ESPCA. The results show that AWGE-ESPCA can still accurately identify target genes even in the presence of noise. Secondly, we conduct experiments in the *Hermetia illucens* dataset. The experimental results show that the AWGE-ESPCA model is superior to existing models. Pathway enrichment analysis shows that AWGE-ESPCA model can effectively select the pathway information related to growth and development. Thirdly, we select a public insect dataset for our experiments, which shows that the AWGE-ESPCA model can be generalized to analyze data from other insects as well. Fourthly, we perform ablation experiments to validate the role of each component for AWGE-ESPCA. Finally, we perform bioenrichment analyses to identify biomarkers that might influence *Hermetia illucens* growth.

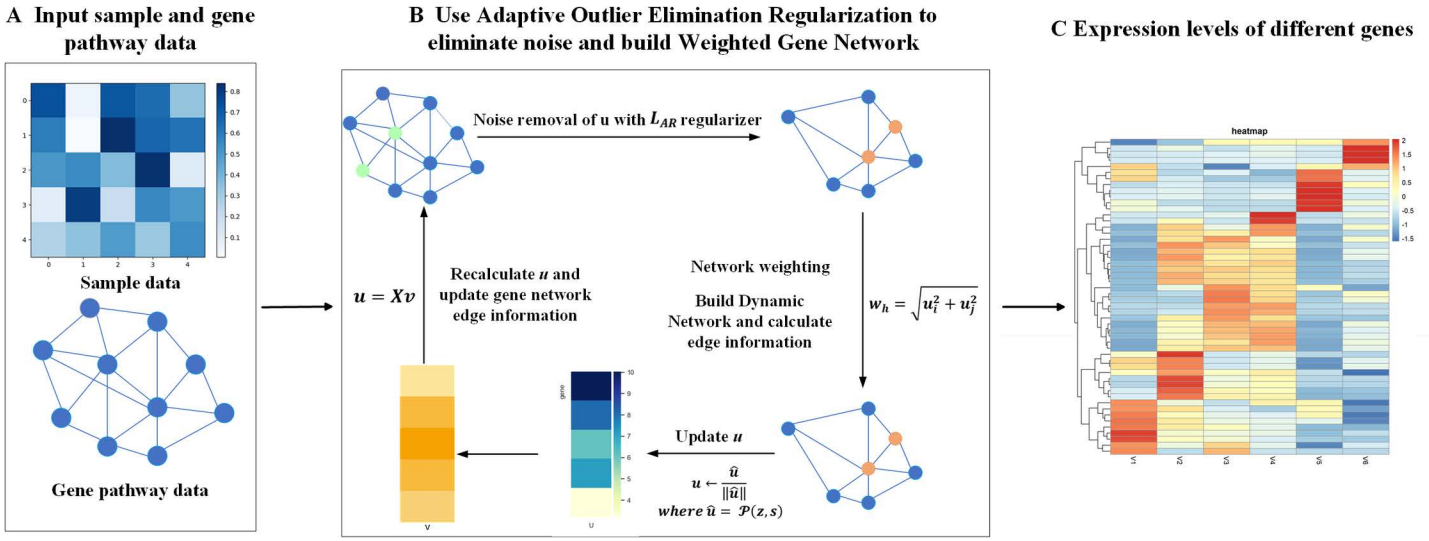

Fig 1. The algorithm of the AWGE-ESPCA model. The steps are as follows: (A) Input Data: It includes sample data and gene pathway data. (B) Data Processing: It involves two core modules. First, Adaptive Noise Elimination Regularization is used to eliminate noise; then, Weighted Gene Network is constructed. (C) Result Output: It obtains gene expression values with different weights.

https://doi.org/10.1371/journal.pcbi.1012773.g001

2. Related work

In terms of dataset construction, researchers have created a few datasets on the connection between Cu^{2+}-stressed and the growth and development of *Hermetia illucens*. For instance, Wang et al. researches the concentration of metal ions in the feces of larvae of *Hermetia illucens* grown under different concentrations of Cu^{2+}-stressed [6]. This experiment provides compelling evidence that *Hermetia illucens* will be enriched in excess Cu^{2+}-stressed, which will impact the larvae's growth and development. Deng et al. [20] constructs experiments to verify that Cu^{2+}-stressed and Cu^{2+}-stressed directly affect the body weight and pupation rate of *Hermetia illucens* larvae. However, no researchers have established genomic datasets of *Hermetia illucens* growth under different concentrations of Cu^{2+}-stressed. This prevents the researchers from studying the important genes and pathways affected by Cu^{2+}-stressed on the growth of *Hermetia illucens*, and how these genes might affect poultry and humans as they become enriched.

In genome analysis models, the common machine learning models to study the genome analysis of insects can be divided into two categories currently [21,22]. One category is supervised learning models [21]. Theoretically, supervised learning models have high prediction accuracy and good performance, but the model needs a lot of samples for training [22,23]. For insect data, the sample size is relatively limited [24]. Therefore, although supervised learning models perform well in many specific genomic data analysis tasks [25], it is difficult to effectively analyze insect genomic data. Another category is unsupervised learning models [26]. The advantages of unsupervised learning models include low sample size requirements, the ability to learn from unlabeled data and discover underlying patterns and structures [27,28]. They are excellent in processing high-dimensional data, dimensionality reduction, clustering and outlier detection [29,30]. For insect data with a relatively limited sample size, unsupervised learning models show greater potential and advantages in insect genome analysis [31]. Among the unsupervised models, the PCA model is one of the most commonly used models for dimensionality reduction and feature extraction [32,33]. However, the traditional PCA model arithmetic process is not able to remove the noise contained in the high-dimensional gene data [34], which reduces the credibility of the PCA model results. Therefore, some researchers proposes the Sparse PCA model (SPCA) [35], which introduces the L_0 regularizer to remove the noise in the data, characterized by strong biological interpretability and fast operation speed. The SPCA model achieves good performance in many biological application scenarios [36,37]. Although the SPCA model has a great improvement in the ability of noise removal and dimensionality reduction analysis compared with the PCA model, the SPCA model is not able to utilize the known biological network structure information [36]. In 2018, Min et al. [38] proposed the Sparse Edge Group PCA model (ESPCA) for feature extraction of gene data. In 2022, Miao et al. [39] proposed a Sparse PCA model based on meta-learning and dynamic networks (DM-ESPCA) to further improve the feature selection ability of the Sparse PCA model. However, the existing Sparse PCA models are still not a good solution to the noise problem in insect datasets. Since the existing Sparse PCA models are based on the L_0 regularizer for solving [40], the models will almost certainly select gene probes with too large expression values even if these gene probe are not related to the target of the experiment. Finally, all current Sparse PCA models assume that all gene probes have the same priori information in the pathway, which is also clearly not in line with biological common sense. Researchers usually prioritize those gene probes that are located in pathway-enriched regions [17,18]. This is because these gene probes may be the key probes that influence the epigenetic characteristics of organisms.

3. Materials and models

3.1. Ethics statement

The experimental populations were maintained under standard laboratory conditions following established protocols for insect welfare. All Cu^{2+}-stressed treatments were performed within approved safety guidelines for heavy metal exposure studies.

The RNA-seq experiments were performed by BGI Shenzhen Co., Ltd. in compliance with standard biosafety protocols and quality control procedures. The Drosophila melanogaster dataset (GSE243439) was obtained from the Gene Expression Omnibus (GEO) database and used in accordance with their data usage policies. The pathway information for Hermetia illucens was provided by Shenzhen UW Genome Science and Technology Service Co. under appropriate data sharing agreements.

3.2. Dataset

3.2.1. Hermetia illucens datasets.

***3.2.1.1. Experimental design and population establishment*:** First, we establish an experimental population of *Hermetia illucens* under copper ion (Cu^{2+}) stressed. *Hermetia illucens* eggs are placed in a plastic rearing box in an artificial climate chamber for incubation. Following hatching, we begin raising the larvae by providing artificial diet initially, but without the addition of Cu^{2+}-stressed. Upon reaching 4 days of age, they are divided into groups for Cu^{2+}-stressed experiments. To create experimental populations under Cu^{2+}-stressed, artificial feeds with varying Cu^{2+}-stressed concentrations (0 mg/kg, 75 mg/kg, 150 mg/kg, 300 mg/kg, 600 mg/kg, and 1,200 mg/kg) are fed to the groups for five generations. Three replicates are set up for each treatment group.

***3.2.1.2. High-throughput sequencing*:** Total RNA is isolated from the 5th instar larvae of *Hermetia illucens* in each treatment group using Trizol reagent (Invitrogen, Carlsbad, CA, USA). The purity and concentration of the RNA samples are determined using a Nanodrop spectrophotometer and a Qubit® 2.0, and the integrity of the RNA samples is determined using an Agilent 2100. After the total RNA samples pass the inspection, the RNA of *Hermetia illucens* larvae from each treatment group is sent to BGI Shenzhen Co., Ltd. for transcriptome sequencing using the Illumina Hiseq4000 sequencing platform. The sequencing results are RNA-seq expression values, with each sample analyzed for expression levels across 19,512 gene probes.

3.2.2. Drosophila melanogaster dataset. A dataset from the Gene Expression Omnibus (GEO) database is used this time: GSE243439 [https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE243439].

The core goal of this dataset is to find genes associated with the development of Chronic Myeloid Leukaemia (CML) in Drosophila melanogaster, with nine samples.

These nine samples are divided into three groups: the first is a control (W1118), the second is BCR-ABLp210, and the third is BCR-ABLT315I. The control (W1118) is a normal Drosophila, whereas BCR-ABLp210 and BCR-ABLT315I represent two different forms of BCR-ABL1 fusion proteins respectively.

The BCR-ABL1 fusion protein is produced by the fusion of the BCR and ABL1 genes and has aberrant tyrosine kinase activity, leading to CML. BCR-ABLp210 is the common form of the fusion protein, whereas BCR-ABLT315I is a drug-resistant mutant form.

Among them, we choose the control group and the BCR-ABLp210 group for our study because BCR-ABLp210 is the main causative gene of CML, and the study of its molecular characterisation can provide basic data that are more widely applicable to most CML patients.

3.2.3. Gene pathway data sets. Pathway information for *Hermetia illucens* gene is provided by Shenzhen UW Genome Science and Technology Service Co. Total 65829

pathway information. The Pathway of Drosophila melanogaster dataset comes from Pathway Commons dataset at http://www.pathwaycommons.org/ and https://www.genome.jp/kegg/.

3.3. Models

3.3.1. SPCA. Suppose there is an existing gene matrix $X \in R^{m,n}$, where m denotes the number of gene probes and n denotes the number of samples. The researcher usually uses Singular Value Decomposition (SVD) framework to solve genome sparsification problems based on L_0 regularizer. The optimization problem is formulated as equation (1):

$$\underset{\|u\|_2 \le 1, \|v\|_2 \le 1}{\text{maximize}} \; u^T X v, \text{s.t.} \, \| u \|_0 \le s \tag{1}$$

where u is an $m \times 1$ vector represents the first principal component (PC) loadings, while v represents the corresponding $n \times 1$ PC. The value of s represents the count of genes that the model retains. The $\| u \|_0$ and $\| u \|_2$ represent the L_0 and L_2 norms. Since the L_0 norm is not directly minimizable due to its non-convex nature, researchers generally adopt a greedy principle for the solution, i.e. $\| u \|_0 = \mathcal{P}(z,s)$,where $\mathcal{P}(z,s)$ achieves sparse projection. For vector u, its k -th entry is defined by equation (2):

$$[P(z,s)]_k = \begin{cases} z_k, \text{ if } k \in \text{supp}(z,s) \\ 0, \text{otherwise} \end{cases} \tag{2}$$

where $supp(z,s)$ denotes the set of indexes of the largest s absolute element of z.

3.3.2. ESPCA. In a 2018 study, Min et al. introduced the edge group Sparse PCA (ESPCA) [18]. The ESPCA framework evolves from the conventional probe sparse to a group sparse configuration, significantly enhancing the capability of feature discrimination in Sparse PCA. Let's define G as a group structure within the gene interaction network, where two connected genes form a group. Obviously, such edge-groups are overlapping. We denote $\mathcal{G} = \{e_1, \ldots, e_m\}$ as an edge set with all edges from a given gene interaction network. Here, The ESPCA formulation is presented as equation (3):

$$\| u \|_{ES} = \underset{\forall \mathcal{G}' \in \mathcal{G},\, support(u) \subseteq V(\mathcal{G}')}{minimize} |\mathcal{G}'| \tag{3}$$

where G' is a subset of G, $V(\mathcal{G}')$ is a vertex (gene) set induced from the edge set G', $|\mathcal{G}'|$ denotes the number of elements of G', and $support(u)$ denotes the set of indexes of nonzero elements of u [41]. Drawing from equation (4), this sparse model can be delineated as equation (4):

$$\underset{\|u\|_2 \le 1, \|v\|_2 \le 1}{maximize} \; u^T X v, s.t. \| u \|_{ES} \le s \tag{4}$$

where s represents the count of edges. The model also uses the greedy principle for sparse solving.

3.3.3. AWGE-ESPCA. Here, we propose the AWGE-ESPCA model (Fig 2). Compared with the existing models, the AWGE-ESPCA model has two main improvements. First, the AWGE-ESPCA model proposes a new adaptive Noise Elimination regularizer. The noise problem in insect datasets is manifested as extremely uneven data distribution, with some data probes having extremely large values, while others have extremely small values. This data noise greatly reduces the credibility and biological interpretability of the algorithm

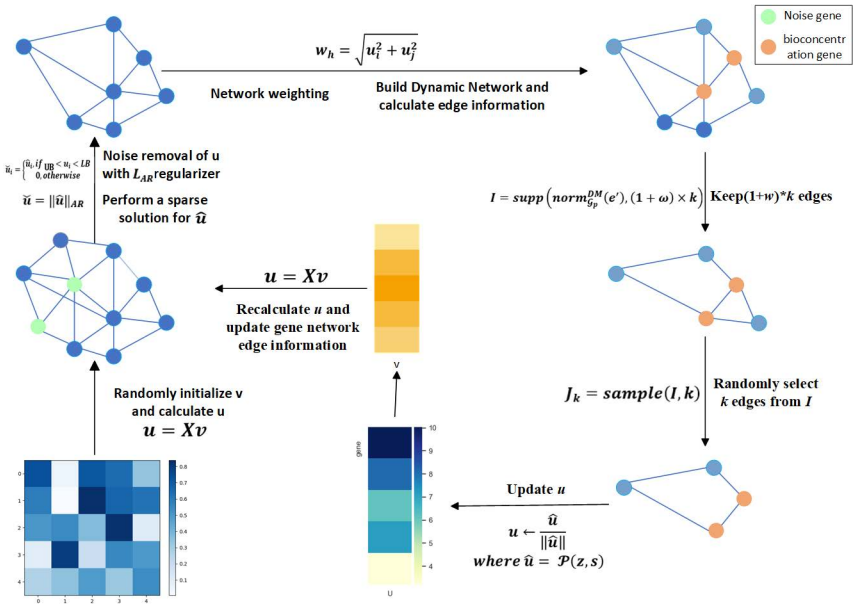

Fig 2. Flow chart of the AWGE-ESPCA model. Fig 2 shows the specific flow of the AWGE-ESPCA algorithm. First, randomly initialize v and calculate *u*; then, identify L_{AR} based on the modified regularizer and remove the data noise to get the new $\breve{u}$; then, calculate the edge information based on the gene network and the weighted information; finally, retain the important genes and edge information based on the $\breve{u}$ and edge information and continue to loop through the process based on the current result.

https://doi.org/10.1371/journal.pcbi.1012773.g002

and increases the complexity of data analysis. The adaptive Noise Elimination regularizer proposed in this paper can adaptively remove outliers from the data. It is more efficient and accurate than the previous manual selection of data. Second, the AWGE-ESPCA model proposes a new Weighted Gene Interaction Network (WGIN) module, which allows the model to focus on selecting key genes related to the pathway enrichment and target experiment function region. For example, the main goal of our *Hermetia illucens* gene data analyses is to find the pathway region related to growth and development, so WGIN's weighting module will make the model pay more attention to the corresponding pathway region.

***3.3.3.1. Adaptive regularization for noise probes elimination*:** Firstly, the model will randomly initialize the sample weight feature vector *v*. Next, the gene probe weight feature vector u is calculated based on equation (5):

$$u = Xv \tag{5}$$

here, the weight u_i of gene probes with excessive gene expression values will be very large, and we believe that they should be considered as potential noise in the data. Therefore, we propose the L_{AR} regularizer to perform noise removal for *u*. The specific steps are shown in equations (6)–(12).

To begin, we sort *u* from smallest to largest on equation (6):

$$\hat{u} = \mathrm{sort}(u) \tag{6}$$

Next, we compute the parameters $Q1_{\mathrm{sorted}}$, $Q3_{\mathrm{sorted}}$ according to equations (7) and (8):

$$Q1_{\text{sorted}} = u_k, k = \| m * lq \|_{in} \tag{7}$$

$$Q3_{\text{sorted}} = u_k, k = \| m * (1 - lq) \|_{in} \tag{8}$$

where lq is the regularization parameter of the model input, within the range (0, 0.5), and $\| * \|_{in}$ _ in denotes the integerized regularizer, which is obtained by rounding up to the next integer if the decimal portion of $m * (1 - lq)$ is greater than or equal to 0.5, otherwise it is rounded down to the current integer. Next, we calculate the upper edge weight parameter LB and the lower edge weight UB on equations (9) and (10):

$$\text{LB} = Q1_{\text{sorted}} - (1.5 \times \text{IQR}) \tag{9}$$

$$\text{UB} = Q3_{\text{sorted}} + (1.5 \times IQR) \tag{10}$$

where, $IQR = Q3_{\text{sorted}} - Q1_{\text{sorted}}$.

Finally, we sparsely solve for $\hat{u}$ using the sparse projection of equation (11), with:

$$\check{u}_i = \begin{cases} \hat{u}_i, if\ UB < u_i < LB \\ 0, otherwise \end{cases} \tag{11}$$

That is, the L_{AR} regularizer can be finally expressed as equation (12):

$$\check{u} = \| \hat{u} \|_{AR} \tag{12}$$

***3.3.3.2. Weighted gene interaction network*:** After computing $\check{u}$, we restore the order of the gene probes in $\check{u}$ to the initial state based on their original positions in the vector u, $\check{u} = returne(\check{u})$. Next, we introduce the weighted gene network sparsification module as shown in equations (13)–(22):

We assume that $e_h = (u_i, u_j) \in \mathcal{G}, u_i, u_j \in R^m$, and the weight w_h of e_h is define as formula (13):

$$w_h = \sqrt{u_i^2 + u_j^2} \tag{13}$$

where u_i and u_j are the left and right gene probes of e_h.

First, we calculate a weight vector $Wq = \{wq_1, \ldots, wq_m\}$ for each gene probe based on the known gene-pathway relationship information, where wq_i represents the number of pathways associated with the i -th gene probe that are relevant to the experiment target. In the absence of an explicit experimental goal, wq_i represents the number of pathways corresponding to the i -th gene probe. For example, in the *Hermetia illucens* dataset, Wq represents the number of gene pathways related to growth and development. Since the number of pathways for each probe in Wq varies greatly, we normalize W to $[a - b]$ using equations (14) and (15):

$$k = \frac{b - a}{\text{Max} - \text{Min}} \tag{14}$$

$$wq_i = a + k \times (wq_i - \text{Min}) \tag{15}$$

where a,b are the input parameters, $\text{Max} = \text{Max}\{\text{W}\}$, $\text{Min} = \text{Min}\{\text{W}\}$. The edge weights of the $h-th$ gene network can be represented as equation (16):

$$w_h = \sqrt{wq_i u_i^2 + wq_j u_j^2} \tag{16}$$

Finally, we can get the edges set of the gene network $\mathcal{G}(i)$. We use the Greedy-based Random Sampling Algorithm as proposed by DM-ESPCA for sparse solving of $\mathcal{G}(i)$. That is, Algorithm 1. $\mathcal{P}(z,k)$ is sparse projection, $\left[\mathcal{P}_{\mathcal{G}}(z,k)\right]_i (i=1,\ldots,m)$ meets the condition in equation (17):

$$\left[\mathcal{P}_{\mathcal{G}}(z,k)\right]_i = \begin{cases} z_i, if\ G(i) \cap sample(I,k) \neq \varnothing \\ 0, otherwise \end{cases} \tag{17}$$

where $I = supp\left(norm_{\mathcal{G}}^{AW}(e'), (1+\omega)\times k\right)$. If gene i is selected, $\left[\mathcal{P}_{\mathcal{G}}(z,k)\right]_i = z_i$, otherwise $\left[\mathcal{P}_{\mathcal{G}}(z,k)\right]_i = 0$. k represents the number of edges expected to be retained. ω is a parameter that controls the random ratio.

Finally, we use formula (18)–(20) to update vectors u and v until the algorithm convergence:

$$u = Xv,\ where\ \hat{v} = X^T u \tag{18}$$

$$u \leftarrow \frac{\hat{u}}{\|\hat{u}\|}, \text{where } \hat{u} = \mathcal{P}_{\mathcal{G}}(z,k) \text{ and } z = Xv \tag{19}$$

$$v \leftarrow \frac{\hat{v}}{\|v\|}, \text{where } \hat{v} = X^T u \tag{20}$$

Ultimately, the AWGE-ESPCA model can be expressed as equation (21):

$$\underset{\|u\|_2 \le 1, \|v\|_2 \le 1}{maximize}\ u^T Xv,\ s.t. \|u\|_{AW} \le s \tag{21}$$

Algorithm 1: The algorithm of AWGE-ESPCA

Require : $X \in \mathbb{R}^{m\times n}, \nu \in \mathbb{R}^{n\times 1}, parameter\ k, \omega, \rho,$

$edge\ set\ \mathcal{G}_{\mathcal{P}} = \{e_1, e_2, \cdots e_n\}$

$1: \mathcal{Z} = Xv$ #Randomize the initial v and compute Z

$2: \mathcal{Z} = \| \mathcal{Z} \|_{AW}$ #Adaptive Regularization for Noise Elimination

$3: Let\ norm_{\mathcal{G}_{\mathcal{P}}'}^{AW}(e') = \left(\| e_1' \|, \cdots, \| e_n' \|\right)^T$

4 : ***for*** *any weight of edge e in* $\mathcal{G}_{\mathcal{P}}$ ***do*** qq

$5: w_n' = \sqrt{wq_i z_i^2 + wq_j z_j^2}$ #*Generate* WGIN *network.*

$6: update\ \mathcal{G}_{\mathcal{P}n}' = w_n'$

7 : ***end for***

$8: I = supp\left(norm_{\mathcal{G}_{\mathcal{P}}}^{DM}(e'), (1+\omega)\times k\right)$ #*Extract* $(1+\omega)\times k$ *edges.*

$9: J_k = sample(I,k)$ #*Randomly select k edges from I.*

10 : ***if*** $\omega > 0$ ***then*** $\omega = \omega - \rho$ #*Reduce random rate*

$11: V_{\mathcal{G}_p'} = V\left(\mathcal{G}_p'\right)$
$12:$ ***for*** *any gene i in* $V_{\mathcal{G}_p'}$ ***do***
$13: \hat{u}_i = z_i$ *# Determine the sparse u*
$14:$ ***end for*** $14: u = \frac{\hat{u}}{\|\hat{u}\|}$
$15:$ ***return*** *u and* $\mathcal{P}_{\mathcal{G}_p}(z,k) = \hat{u}$

4. Results

In total, we conduct comprehensive comparisons between AWGE-ESPCA and multiple methods, including traditional Sparse PCA variants (PCA, SPCA, ESPCA, DM-ESPCA), advanced dimensionality reduction methods (t-SNE, UMAP, AEs, VAEs), and supervised methods (Lasso, Elastic Net). For supervised methods, given the small sample size (6 samples), we utilize all samples as training data with Lasso (α = 0.01) and Elastic Net ($\alpha = 0.01$, $l_1\_ratio = 0.5$). All Sparse PCA methods calculate the first two PCs.

The first experiment is a simulation experiment. We generate a batch of carefully designed simulation data. The final results are to prove that the existing methods can not perform feature selection well in the face of strong noise. The second experiment is conducted based on the *Hermetia illucens* dataset. The experiments are divided into two steps. First, the *Hermetia illucens* dataset is distinctly divided into two groups based on the FPKM((Fragments Per Kilobase of transcript per Million mapped reads, a normalized measurement of gene expression level) values of copper ions, one group is low concentrations, including Cu_0_FPKM, Cu_75_FPKM, Cu_150_FPKM and Cu_300_FPKM, the other is high concentrations, including Cu_600_FPKM and Cu_1200_FPKM. The second step is to construct experiments based on different models. The goal is to screen the gene probes and pathway information that may affect the growth and development of *Hermetia illucens*. The third experiment is built on the Drosophila dataset with the aim of further verifying whether AWGE-ESPCA can be extended to the analysis of other insect genomic data. The fourth experiment is an ablation experiment. On the basis of *Hermetia illucens* dataset, we conduct two independent ablation experiments, which independently remove the L_{AR} regularizer and Weighted Gene Interactive Network to validate the effect of the model. The fifth experiment is enrichment analysis, where we perform bio-enrichment analysis on the *Hermetia illucens* dataset to identify potential key pathway information.

To validate the performance of the model, we calculate five independent metrics, including: Heatmap (Top-50, Sparse PCA model using PC1), Sample Distribution Plot (based on PC1 and PC2), Number of Pathways (Top-500, Sparse PCA model using PC1), Target Gene Percentage (genes/total number of genes with experimental target function) and Box Plots of Gene Probe Expression (Top-100, Sparse PCA model using PC1). Among them, the heatmap and sample distribution map allow us to intuitively see whether the gene probes selected by each method can clearly distinguish the samples. The number of pathways and the distribution of target genes can be used to compare the biological feature selection capabilities of the models. The box plots of gene probe expression allow us to intuitively see the number of outliers and the range of average expression values contained in the gene probes selected by the model. This will validate the conjecture of this paper about data noise, that is the existing methods may select more gene probes containing larger expression values. In addition, since the T-SNE and UMAP models do not have feature selection capabilities, only a sample distribution plot comparison is performed for the T-SNE and UMAP models. The Lasso and Elastic

Net models do not have dimensionality reduction capabilities, so no Sample Distribution Plot is drawn.

4.1. Simulation study

In this section, we construct an example to illustrate how it addresses the limitations inherent in the existing Sparse PCA model. We generate a simulated gene expression matrix X and an interaction network defined by an edge set G.

4.1.1. Principal component (PC) loadings generation. Two principal component loadings are defined as equations (22) and (23):

$$\boldsymbol{u}_1 = \left[1, 0.86, 0.66, 0.9, , \left[\mathrm{rep}(0,8)\right]^T\right]^T, \tag{22}$$

$$\boldsymbol{u}_2 = \left[\left[\mathrm{rep}(0,4)\right]^T, 0.2, -0.55, -0.35, 0.17, \left[\mathrm{rep}(0,4)\right]^T\right]^T \tag{23}$$

where $rep(0,a)$ denotes a column-vector of size a, with each element being zero.

4.1.2. Principal components generation. Two principal components are generated using the formulas in equation (24):

$$\boldsymbol{v}_1 = \mathrm{rnorm}(100), \boldsymbol{v}_2 = \mathrm{rnorm}(100) \tag{24}$$

here, $rnorm(b)$ represents a column-vector of size b, with elements randomly sampled from a standard normal distribution.

4.1.3. Expression matrix construction. The expression matrix $X \in R^{100\times 12}$ is constructed for 100 genes across 12 samples. Among them, the first 8 gene probes are the gene probes contained in PC1 and PC2. The last four gene probes are noisy gene probes, which are not associated with any PC but have large expression values (200–300), denoted as $\gamma 0 = x^{100\times 4}$. We generate these four gene probes to validate whether the model can remove the noise information from the data better. Finally, the simulated data expression matrix is represented as equation (25):

$$\boldsymbol{X} = d_1\boldsymbol{u}_1\boldsymbol{v}_1^T + d_2\boldsymbol{u}_2\boldsymbol{v}_2^T + \gamma 0 + \gamma\epsilon \tag{25}$$

where $d_1\boldsymbol{u}_1\boldsymbol{v}_1^T$ represents contributions related to PC1, and $d_2\boldsymbol{u}_2\boldsymbol{v}_2^T$ corresponds to those from PC2. $\gamma\epsilon$ stands for random perturbation.

4.1.4. Expression matrix construction. We have constructed a genetic network $y = y_1 + y_2$, where y_1 and y_2 are as equations (26) and (27):

$$y_1 = \left\{(1,2),(1,3),(1,4),(2,3),(2,4),(3,4),(5,9),(6,10)\right\} \tag{26}$$

$$y_2 = \left\{(5,6),(5,7),(5,8),(6,7),(6,8),(7,8),(1,11),(3,12)\right\} \tag{27}$$

The experimental results strongly demonstrate the performance of AWGE-ESPCA model. Among the five methods, only the AWGE-ESPCA model successfully selects all the correct gene probes, while the remaining methods all select a large number of noisy gene probes. Especially the SPCA model, which selects all four noisy gene probes as PC1 and PC2 (Table 1 and Table A in S1 Text). The results of the simulation experiments confirm our conjecture that gene probes with excessive expression values significantly affect the results of the model in a high-noise environment.

4.2. Hermetia illucens experiment

The AWGE-ESPCA model achieves good results in the *Hermetia illucens* dataset. Heatmap analysis shows that the gene probes selected by AWGE-ESPCA model can better distinguish the two groups of data (Fig 3A–3C and Figs A and B in S1 Text). The heatmap generated by the AWGE-ESPCA model shows clear boundaries and significant color differences between different categories, which helps to make a clear distinction. Among other comparative models, DM-ESPCA performed best. However, the DM-ESPCA model heatmap contains a significant set of noise genes. Red indicates that the expression value of the gene probes is too large, but this group of gene probes are obviously not highly correlated with the sample distribution. In contrast, models including DM-ESPCA, SPCA, AEs, VAEs and Lasso are unable to distinguish between *Hermetia illucens* samples with low concentrations and high concentrations of Cu^{2+}-sressed well. This is a good proof of our conjecture that noise in the data can significantly affect the results of feature selection.

In Fig 4A–4C and Fig B in S1 Text, we observe that the AWGE-ESPCA model distinguishes the dataset with clear boundaries. Box Plots analysis (Fig 4E) further validates this advantage. In the analysis of Top-100 gene probes under Cu_0_FPKM condition, genes selected by AWGE-ESPCA show more compact distribution and fewer outliers, indicating higher inter-group differences and intra-group consistency (Fig C in S1 Text). All the unsupervised models that participated in the comparison performed poorly, particularly in the group of high-concentration samples. The probes in the two groups are far apart, making it difficult to determine whether they belong to the same group. In addition, Lasso and Elastic Net show relatively good performance in Box Plots analysis, with fewer outliers in their selected gene probes. Although Lasso and Elastic Net demonstrate relatively good control over outliers in the box plot analysis, their poor performance in gene probe selection is evident in the sample distribution plots. We attribute this to their ability to utilize grouping information through supervised learning to avoid potentially anomalous gene probes.

In pathway number analysis, AWGE-ESPCA demonstrate exceptional performance (Fig 4F and Table 2). Among the top 500 pathways with the largest loads, AWGE-ESPCA identified 353 pathways related to growth and development, with 70.4% of target genes associated with these pathways. After AWGE-ESPCA, ESPCA models follows with a performance effectiveness of 67.0%. The worst performer is the AEs model, which identify 277 pathways (with a

Table 1. The top two identified PC1 and PC2 loadings by ESPCA, DM-ESPCA and AWGE-ESPCA.

Method	ESPCA		DM-ESPCA		AWGE-ESPCA	
PC	PC1	PC2	PC1	PC2	PC1	PC2
Var1	−0.001	−0.006	0	0.002	0.596	0
Var2	0	0	0	0	0.508	0
Var3	0	0.028	−0.001	0.028	0.422	0
Var4	0	0	0	0	0.458	0
Var5	0	0	0	0	0	0.294
Var6	−0.002	−0.052	−0.002	−0.046	0	−0.933
Var7	0	0	0	0	0	−0.071
Var8	0	0	0	0	0	0.195
Var9	−0.568	0	−0.571	0	0	0
Var10	−0.587	−0.525	−0.591	−0.333	0	0
Var11	−0.577	0.531	0	0.885	0	0
Var12	0	0.662	−0.57	0.321	0	0

https://doi.org/10.1371/journal.pcbi.1012773.t001

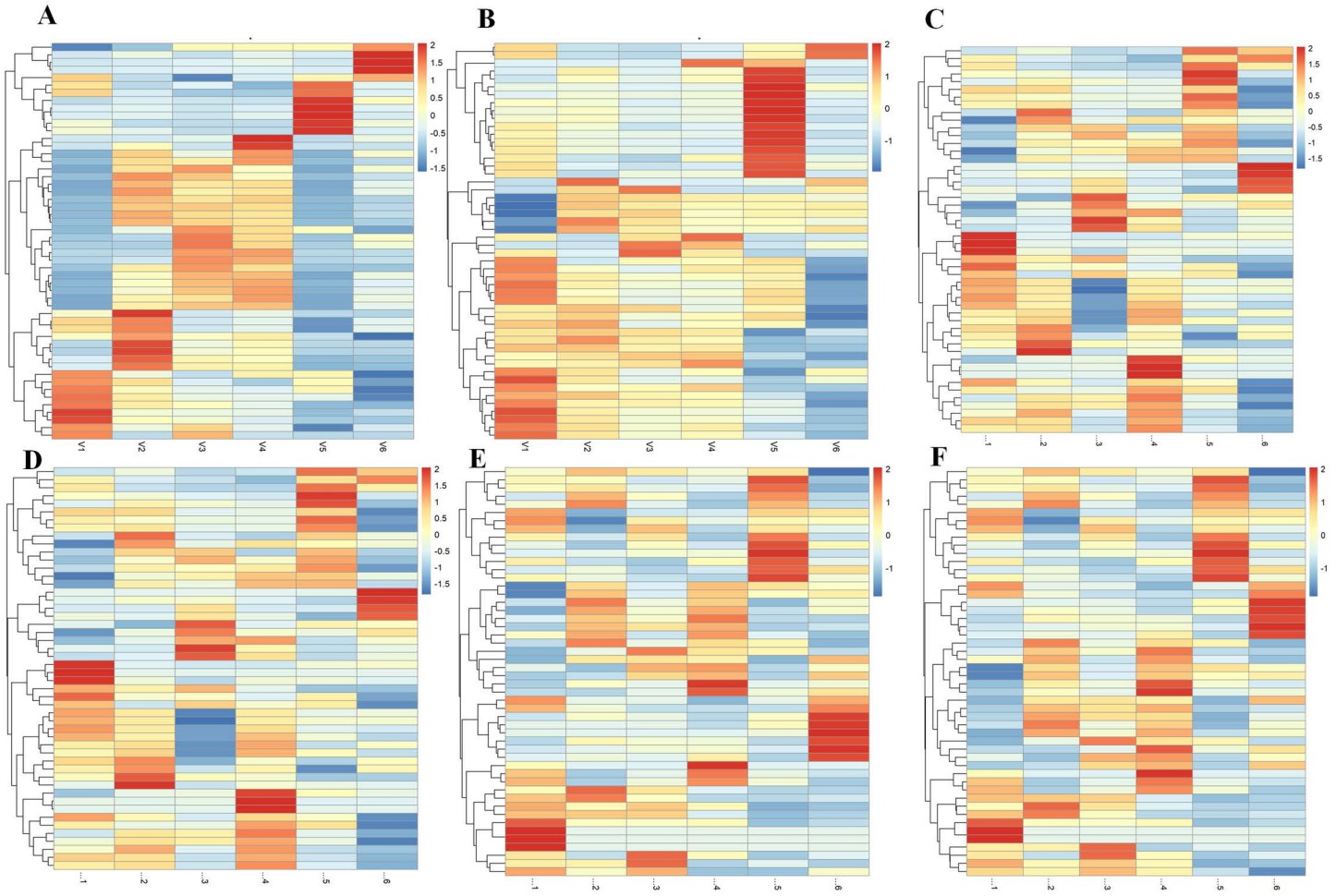

Fig 3. Heatmaps comparing different methods for sample classification. (A) the result of the AWGE-ESPCA model. (B) the result of the DM-ESPCA model. (C) the result of the AEs model. (D) the result of the VAEs model. (E) the result of the Lasso model. (F) the result of the Elastic Net model. The columns represent different categories, namely: Cu_0_FPKM, Cu_75_FPKM, Cu_150_FPKM. The rows are samples, and the colors in the heatmap represent the gene expression values.

https://doi.org/10.1371/journal.pcbi.1012773.g003

correlation of 45.4%). The supervised models also perform less than optimally, with Lasso identifying 169 pathways (with a correlation of 33.8%) and Elastic Net identifying 229 pathways (with a correlation of 45.8%).

In summary, the results of experiment highlight the potential of AWGE-ESPCA to provide more comprehensive insights into genomic responses to Cu^{2+}-stressed in *Hermetia illucens*, particularly in growth and development processes. Compared with other models, our model not only identifies more relevant pathways but also demonstrates stronger biological associations with research objectives through its adaptive noise removal and targeted gene selection capabilities.

4.3. Drosophilamelanogaster dataset

To verify the scalability of the AWGE-ESPCA model, we use the publicly available drosophila melanogaster dataset for verification.

The results are similar to the experimental results of *Hermetia illucens* dataset. AWGE-ESPCA model achieve the best experimental performance. Both heatmap analysis and sample

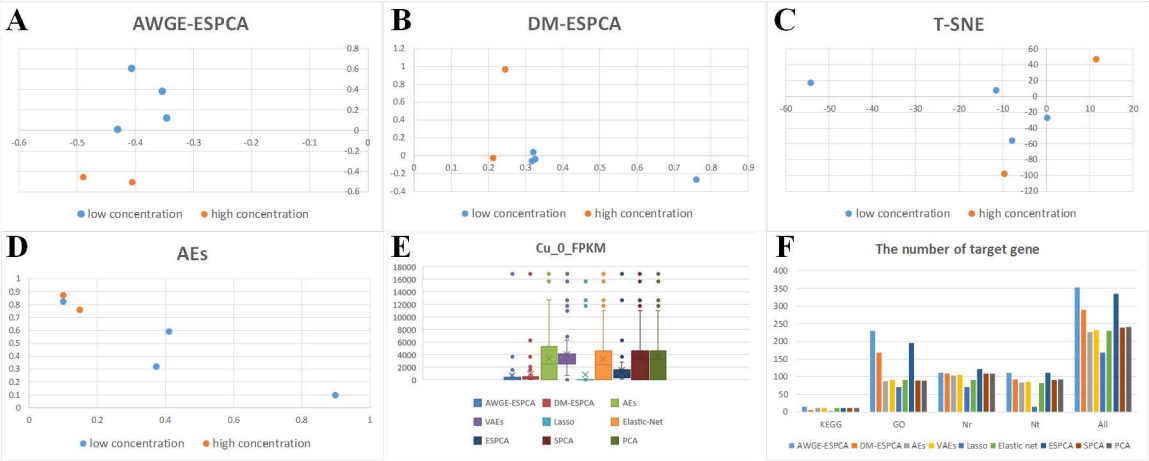

Fig 4. Sample distribution visualization and analysis results across different models. (A) the score plots of the AWGE-ESPCA model. (B) the score plots of the DM-ESPCA model. (C) the score plots of the T-SNE model. (D) the score plots of the AEs model. (E) boxplots comparing gene expression levels under Cu_0_FPKM condition across different models. (F) The number of target pathway genes identified by each model.

https://doi.org/10.1371/journal.pcbi.1012773.g004

Table 2. The proportion of target pathway genes for Hermetia illucens experiment.

PCA model	The proportion of target pathway genes
AWGE-ESPCA	70.4%
DM-ESPCA	57.8%
ESPCA	67.0%
SPCA	47.8%
PCA	48.0%
Lasso	33.8%
Elastic Net	45.8%
AEs	45.4%
VAEs	46.2%

https://doi.org/10.1371/journal.pcbi.1012773.t002

distribution plot show that only the AWGE-ESPCA model clearly distinguished between the two groups of drosophila melanogaster samples, and the rest of the models contained an anomalous t315 group of drosophila melanogaster samples (Figs 5 and 6, and Figs D and E in S1 Text). In contrast, other unsupervised models, including DM-ESPCA and AEs, cannot get better results.

Box Plots analysis shows that only the AWGE-ESPCA model contains the smallest data, and none of the compared models can avoid choosing gene probes with larger expression values (Fig 6E and 6F, and Fig F in S1 Text). In addition, the results of the supervised models are also less than ideal, which may be due to the influence of small samples with high dimensions.

In pathway gene identification (Table 3), AWGE-ESPCA achieve an identification rate of 27.27%, which is significantly better than other models, proving that AWGE-ESPCA is still the strongest in target screening.

To summarize, all the experiments show that the AWGE-ESPCA model is superior to the existing feature extraction models, which can effectively remove the noise from insect genomic data and screen out the key gene targets and pathways.

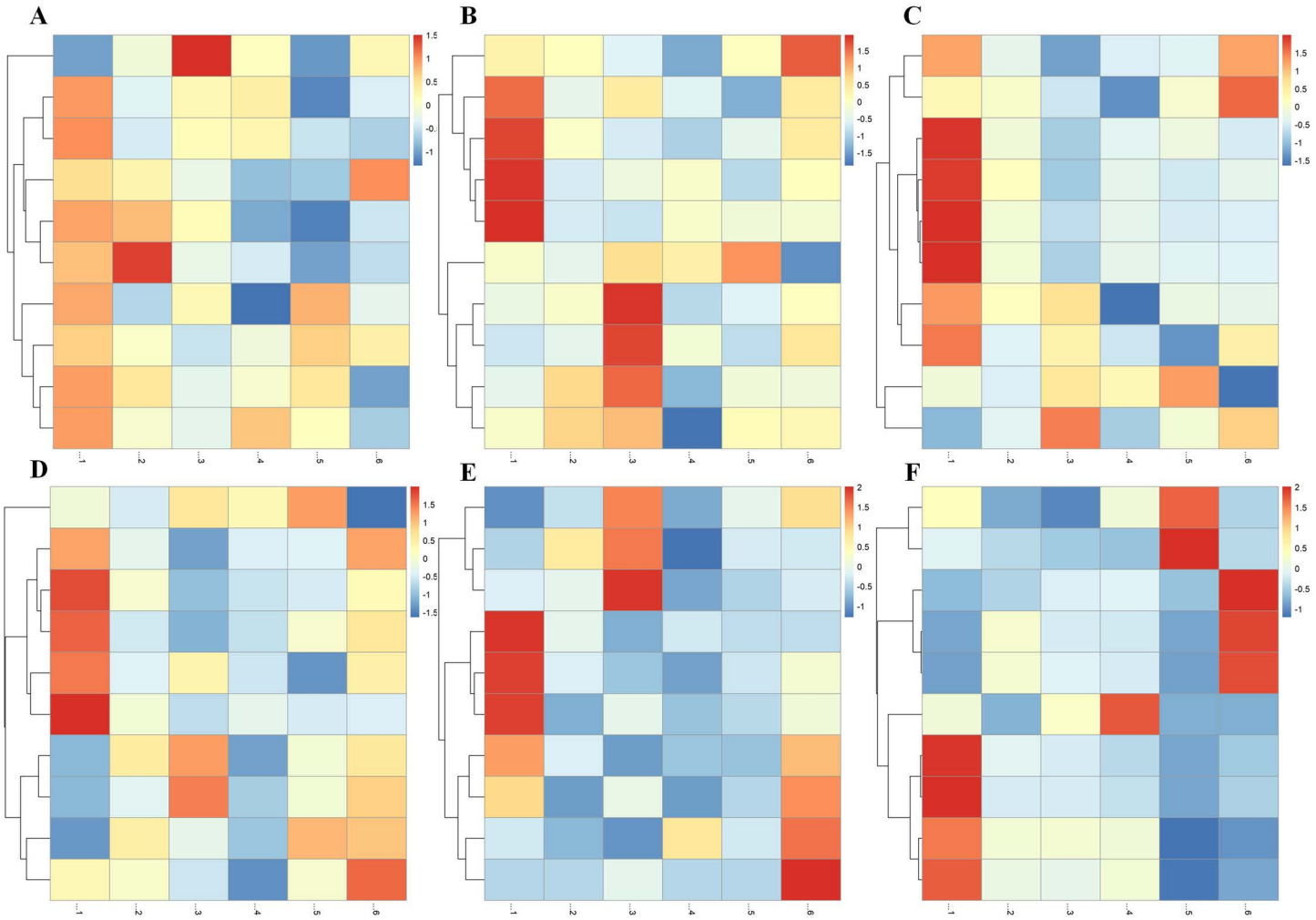

Fig 5. Heatmaps comparing different methods for sample classification. (A) the result of the AWGE-ESPCA model. (B) the result of the DM-ESPCA model. (C) the result of the AEs model. (D) the result of the VAEs model. (E) the result of the Lasso model. (F) the result of the Elastic Net model. The columns display two samples - P210 (P210_1, P210_2, P210_3) and T315 (T315_1, T315_2, T315_3). The rows display different genes.

https://doi.org/10.1371/journal.pcbi.1012773.g005

4.4. Ablation experiment

To further investigate the utility of the two core modules in the AWGE-ESPCA model, Adaptive Regularization for Noise Elimination and Weighted Gene Interaction Network. We conduct ablation experiments based on the *Hermetia illucens* dataset.

We make comparisons using four metrics: sample distribution plots, number of pathways, percentage of target genes, and box plots of the distribution of expression values at gene probes. According to Table 4, the experimental results show that after the removal of the Weighted Gene Interaction Network module, the proportion of target pathway genes significantly decreased from the original 69.1% to 57.8%. In Non-Regularization models, it is 68.8%. The experimental results strongly demonstrate the role of Weighted Gene Interaction Network module in recognizing target genes.

The sample distribution plots show that only the Non-Regularization model fails to correctly distinguish between the two groups of samples (Fig 7A–7C). This shows that

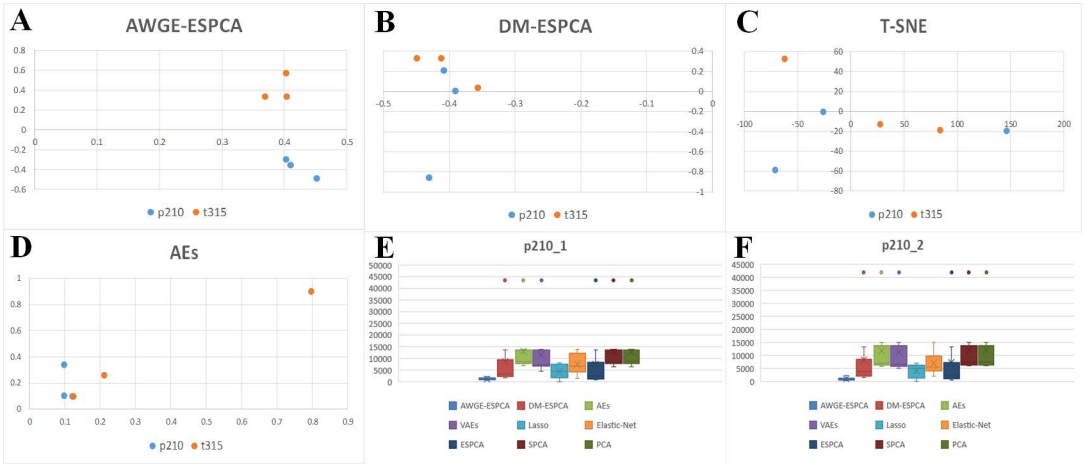

Fig 6. Principal component score plots and boxplots of various models. (A) The score plot of the AWGE-ESPCA model. (B) The score plot of the DM-ESPCA model. (C) The score plot of the T-SNE model. (D) The score plot of the AEs model. (E) The boxplot of the p210_1. (F) The boxplot of the p210_2.

https://doi.org/10.1371/journal.pcbi.1012773.g006

Table 3. The percentage of target pathway genes for Drosophilamelanogaster dataset.

PCA model	The proportion of target pathway genes
AWGE-ESPCA	27.27%
Elastic-Net	9.79%
Lasso	8.70%
AEs	22.8%
VAEs	22.8%
DM-ESPCA	13.64%
ESPCA	9.09%
SPCA	22.40%
PCA	22.40%

https://doi.org/10.1371/journal.pcbi.1012773.t003

Adaptive Regularization for Noise Elimination module can effectively enhance the model's ability to discriminate between samples. Meanwhile, the number of outliers genes in the Non-Regularization model increase significantly compared to AGWE-ESPCA (Fig 8). At the same time, the performance of the Non-Weighted model is significantly better than that of Non-Regularization. These experiments strongly demonstrate the ability of Adaptive Regularization for Noise Elimination module in eliminating genomic data noise.

4.5. Bioenrichment analysis

To explore the pathway information that plays a key role in the growth of *Hermetia illucens* under Cu^{2+}-stressed, we extract the PC1-Top 500 gene probes of the AWGE-ESPCA model in the *Hermetia illucens* dataset for bioenrichment analysis. We identify several key pathways including: GO:0007511, GO:0002181, GO:0006338 and GO:0032504 (Fig 9). Among them, GO:0002181 relates to protein synthesis through cytoplasmic translation, GO:0006338 is involved in chromatin structure modification, and GO:0032504 is associated with reproduction in multicellular organisms.

Among them, GO:0007511 is the pathway with the strongest correlation to the Top 500 gene probes set. Therefore, we perform a more in-depth analysis. Based on literature review

Table 4. The proportion of target pathway genes for Ablation experiment.

PCA model	The proportion of target pathway genes
AWGE-ESPCA	69.1%
Non-Weighted	57.8%
Non-Regularization	68.8%

https://doi.org/10.1371/journal.pcbi.1012773.t004

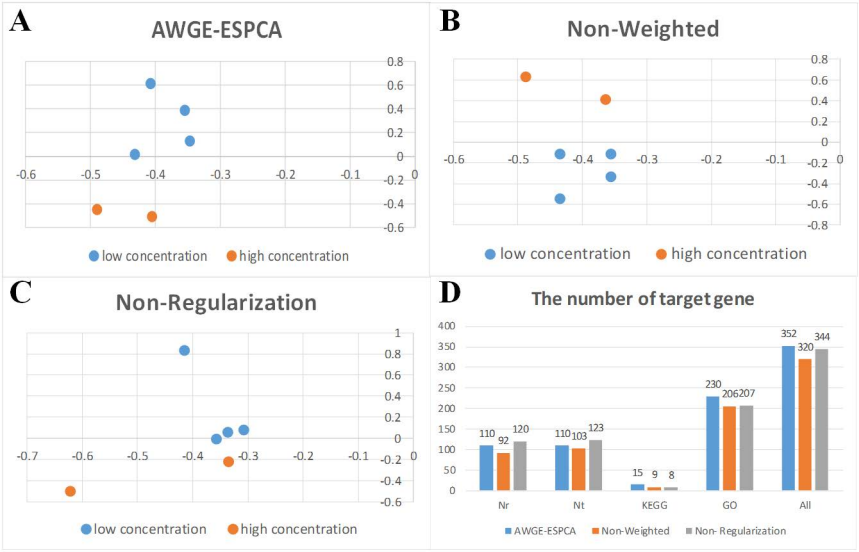

Fig 7. Principal Component Score Plot and the proportion of target pathway genes. (A) the result of the AWGE-ESPCA model. (B) the result of the Non-Regularization model. (C) the result of the Non-Weighted model. (D) the number of target pathway genes.

https://doi.org/10.1371/journal.pcbi.1012773.g007

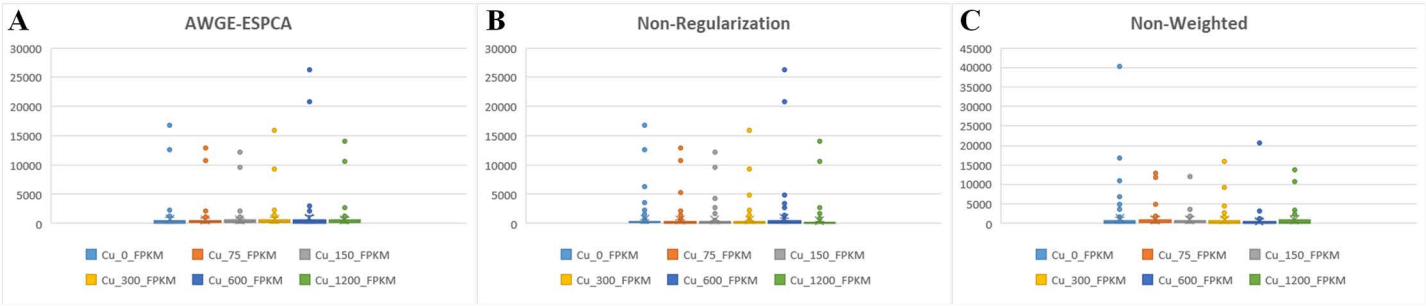

Fig 8. Boxplots. (A) the result of the AWGE-ESPCA model. (B) the result of the Non- Regularization model. (C) the result of the Non-Weighted model.

https://doi.org/10.1371/journal.pcbi.1012773.g008

and enrichment analysis, we identify FGF8, BMP2 and Notch1 gene probes as potential key sites affecting insect cardiac development.

Studies by Lewandoski, Moon, Harada and Falardeau et al. have shown that the FGF8 signaling pathway regulates the development of the midbrain and hindbrain mainly through the Otx2/Gbx2 transcription factor system [42,43], while also regulating nervous system development through the survival of GnRH neurons [45], and influencing the formation of limbs through expression in the AER region [44]. These findings suggest that Cu^{2+}-stressed may

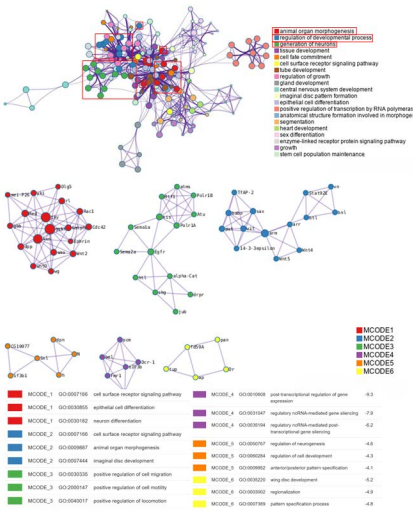

Fig 9. Bio-enrichment analysis picture. Bio-enrichment analysis revealing functional interactions across six MCODE modules, highlighting key pathways in cellular differentiation, neurogenesis, and morphogenesis with corresponding enrichment scores ($p < 0.05$).

https://doi.org/10.1371/journal.pcbi.1012773.g009

affect the normal development process of multiple organ systems in the *Hermetia illucens* by disrupting the FGF8 signaling pathway.

Studies by Rivera-Feliciano, Ma and Prall et al. have shown that BMP2 regulates the fate decision of cardiac progenitor cells through the Mad/Medea transcription factor system and initiates the cardiomyocyte differentiation program through the Punt/Tkv receptor [45–47]. These findings suggest that Cu^{2+}-stressed may affect the normal differentiation and development of cardiac progenitor cells by disrupting the BMP2 signaling pathway.

Studies by Boni, Urbanek and Metrich et al. have elucidated that Notch1 regulates the fate of cardiac progenitor cells through Delta/Serrate-Notch ligand-receptor interactions and regulates cardiomyocyte differentiation in a manner dependent on the Su(H) transcription factor [48–50]. These findings suggest that Cu^{2+}-stressed may affect the differentiation and development of heart cells by interfering with the Notch1 signaling pathway.

In summary, the studies of MacGrogan and de la Pompa et al. show that BMP2 expression in the heart is required to maintain Notch1 expression in endothelial cells, and that these two signaling pathways regulate heart development by synergistically inducing Snail1 expression [51,52]. Meanwhile, Notch1 signaling promotes FGF8 expression in the second heart field, which in turn can remedy EMT defects [53]. These findings suggest that Cu^{2+}-stressed may interfere with the heart development process of the *Hermetia illucens* by affecting the precise temporal regulatory network among these signaling pathways.

We believe that these pathways are the key pathway information for Cu^{2+}-stressed to affect the growth of *Hermetia illucens*. To summarize, we demonstrate the performance of the AWGE-ESPCA model through a simulation experiment and two real datasets. The ablation experiment proves the usefulness of the Adaptive Regularization for Noise Elimination and Weighted Gene Interaction Network module proposed in this paper. In addition, in the enrichment experiments, we identify these key pathways that may affect the growth of *Hermetia illucens*. We are reasonably confident that the AWGE-ESPCA model is superior to existing Sparse PCA models, and it can provide a new artificial intelligence tool for feature extraction from insect genomic data.

5. Discussion

In the process of studying the effects of Cu^{2+}-stressed on *Hermetia illucens*, we delve into several significant challenges faced in the genomic analysis of *Hermetia illucens*. The main issues include the scarcity of high-quality genomic data, noise in genomic data, and the inability of existing models to select pathway-enriched regions well. It is important to note that, unlike human gene data sets, the sample size of insect genomic data set is particularly small. Therefore, it is difficult to use supervised models (e.g., Lasso or Elastic Net) to build a classification model with sufficient robustness and generalization performance. The same is true for the choice of regularization models, and L_1 and L_2 regularizers are widely used in the genomics field for dimensionality reduction and feature selection with good results. However, this type of regularizer requires supervised operations. Therefore, it is difficult to obtain good enough performance in such small sample datasets. This is also verified in our experiments. We make progress in overcoming these challenges by developing and applying the AWGE-ESPCA model.

Firstly, we design experiments and construct a genomic dataset containing six samples of Cu^{2+} stressed *Hermetia illucens* growth. This dataset allows us to study how *Hermetia illucens* larvae are affected by Cu^{2+}-stressed during growth resulting in growth differences. This is essential to enhance the safety of *Hermetia illucens* as feed. Secondly, for the noisy characteristics of the *Hermetia illucens* genomic dataset, we propose a novel Adaptive Regularization for Noise Elimination and integrate it into the existing Sparse PCA framework. The L_{AR} regularizer can adaptively remove gene probes with excessive expression values in the *Hermetia illucens* genomic data. Our experimental analysis demonstrates that gene probes exhibiting unusually high expression values frequently lack discriminatory power between different sample groups. The L_{AR} regularizer proves to be more efficient and flexible than traditional manual data correction approaches in addressing this challenge. To validate whether the eliminated gene probes represent important biological signals, we conduct enrichment analysis on the potential noise gene probe set (1037 gene probes). The results show that 58.8% of gene probes are not associated with any known pathways, 35.2% are associated with only one pathway, and merely 6% are associated with multiple pathways. Furthermore, P-value analysis reveal that only 10.41% of these genes show statistical significance ($P < 0.05$), with an average P-value of 0.42, suggesting these eliminated probes are more likely to represent random fluctuations rather than biologically significant signals (Fig G in S1 Text). Thirdly, we propose the Weighted Gene Interaction Network. This enables the model to focus more on regions within the gene network that are rich in pathways, which are crucial in biological processes. In addition, we would like to discuss the running time, memory consumption and scalability of the model in particular. All our experiments are performed on a computer with an AMD 7950X CPU and 32GB RAM. We compare the running time and memory consumption of the AWGE-ESPCA model with the comparison model on the *Hermetia illucens* dataset (Table B in S1 Text). The experiment shows that researchers do not need expensive computing equipment to use the AWGE-ESPCA model (42 min/520 MB). We also test the AWGE-ESPCA model on a large genomic dataset (100 samples × 25,000 genes), which requires approximately 4.5 hours of processing time and 3.8 GB memory usage. This supplementary experiment shows that the AWGE-ESPCA model has good scalability. Theoretically, the AWGE-ESPCA model can be applied to the analysis of even larger and more complex human and other biological datasets.

Although the experimental results prove the superiority of AWGE-ESPCA model, this study still has the following two limitations. First, the sample size of the genomic dataset constructed in this study for Cu^{2+}-stressed *Hermetia illucens* growth is still small, which may affect the accuracy of the results. Second, the model only considers short-connected biological

pathway information, i.e., pairwise pathway relationships, when performing feature selection. However, it is clear that researchers will like the model to select long pathway information that is as complete as possible, which will facilitate further wet experimental design. In the future, we plan to further design experiments to supplement the genomic dataset of Cu^{2+}-stressed *Hermetia illucens* growth and continue to improve the AWGE-ESPCA model to enhance its feature selection capability. In summary, we propose an artificial intelligence feature extraction model specifically designed for *Hermetia illucens* genome analysis. Experiments confirm the superior performance of the AWGE-ESPCA model. In addition, we confirm that the AWGE-ESPCA model can be effectively extended to other insect genome analysis tasks. We believe that the AWGE-ESPCA model can help researchers identify potential biomarkers in insect genomes more efficiently.

Supporting information

S1 Text. Supplementary information. Fig A. Heatmaps of the ESPCA, SPCA, PCA model. (A) the result of the ESPCA model. (B) the result of the SPCA model. (C) the result of the PCA model. **Fig B. Principal component score plots of UMAP, VAEs, ESPCA, SPCA, PCA model.** (A) the score plots of the UMAP model. (B) the score plots of the VAEs model. (C) the score plots of the ESPCA model. (D) the score plots of the SPCA model. (E) the score plots of the PCA model. **Fig C. Boxplots of Cu_75_FPKM, Cu_150_FPKM, Cu_300_FPKM, Cu_600_FPKM, Cu_1200_FPKM.** (A) the boxplots of the Cu_75_FPKM. (B) the boxplots of the Cu_150_FPKM. (C) the boxplots of the Cu_300_FPKM. (D) the boxplots of the Cu_600_FPKM. (E) the boxplots of the Cu_1200_FPKM. **Fig D. Heatmaps of the ESPCA, SPCA, PCA model.** (A) the result of the ESPCA model. (B) the result of the SPCA model. (C) the result of the PCA model. **Fig E. Principal component score plots of UMAP, VAEs, ESPCA, SPCA, PCA models.** (A) the result of the UMAP model. (B) the result of the VAEs model. (C) the result of the ESPCA model. (D) the result of the SPCA model. (E) the result of the PCA model. **Fig F. Boxplots comparing gene expression levels between P210 and T315I samples.** (A) the boxplots for P210 replicate 3. (B) the boxplots for T315I replicate 1. (C) the boxplots for T315I replicate 2. (D) the boxplots for T315I replicate 3. **Fig G. Noise probes correlation analysis and pathway number analysis plots.** (A) Distribution of P_value for probes included in the noisy probes. (B) Distribution of the number of individual gene probes associated with known pathways in the noisy gene probes. **Table A. The top two identified PC1 and PC2 loadings by SPCA and PCA. Table B. Time and memory usage comparison across methods.**
(DOCX)

Author contributions

Conceptualization: Rui Miao, Hao-Yang Yu, Qiang Xia.

Data curation: Rui Miao, Hao-Yang Yu, Bing-Jie Zhong, Hong-Xia Sun.

Formal analysis: Rui Miao, Hao-Yang Yu, Bing-Jie Zhong.

Investigation: Hong-Xia Sun, Qiang Xia.

Methodology: Rui Miao, Hao-Yang Yu.

Software: Rui Miao, Hao-Yang Yu.

Supervision: Bing-Jie Zhong.

Validation: Bing-Jie Zhong.

Visualization: Rui Miao, Hao-Yang Yu, Bing-Jie Zhong.

Writing – original draft: Rui Miao, Hao-Yang Yu.

Writing – review & editing: Rui Miao, Bing-Jie Zhong, Qiang Xia.

References

1. Kaczor M, Bulak P, Proc-Pietrycha K, Kirichenko-Babko M, Bieganowski A. The variety of applications of *Hermetia illucens* in industrial and agricultural areas-review. Biology (Basel). 2022;12(1):25. https://doi.org/10.3390/biology12010025 PMID: 36671718
2. Triunfo M, Tafi E, Guarnieri A, Salvia R, Scieuzo C, Hahn T, et al. Characterization of chitin and chitosan derived from *Hermetia illucens*, a further step in a circular economy process. Sci Rep. 2022;12(1):6613. https://doi.org/10.1038/s41598-022-10423-5 PMID: 35459772
3. Zhan S, Fang G, Cai M, Kou Z, Xu J, Cao Y, et al. Genomic landscape and genetic manipulation of the black soldier fly *Hermetia illucens*, a natural waste recycler. Cell Res. 2020;30(1):50–60. https://doi.org/10.1038/s41422-019-0252-6 PMID: 31767972
4. Wang YS, Shelomi M. Review of black soldier fly (*Hermetia illucens*) as animal feed and human food. Foods. 2017;6(10).
5. Kawasaki K, Hashimoto Y, Hori A, Kawasaki T, Hirayasu H, Iwase S-I, et al. Evaluation of black soldier fly (*Hermetia illucens*) larvae and pre-pupae raised on household organic waste, as potential ingredients for poultry feed. Animals (Basel). 2019;9(3):98. https://doi.org/10.3390/ani9030098 PMID: 30893879
6. Wang X, et al. Effect of Cu^{2+}, Zn^{2+}, Cd^{2+} on the growth of *Hermetia illucens* larvae and accumulation in larvae and feces. J Environ Entomol. 2019;41(2):387–93.
7. Yin Y, Wang S, Li Y, Yao D, Zhang K, Kong X, et al. Antagonistic effect of the beneficial bacterium *Enterobacter hormaechei* against the heavy metal Cu^{2+} in housefly larvae. Ecotoxicol Environ Saf. 2024;272:116077. https://doi.org/10.1016/j.ecoenv.2024.116077 PMID: 38335578
8. Zhou M, Wu SJ, Tan XH, Sun QX, Li XC, Dong YW, et al. Growth performance and dynamic copper accumulation in tissues of black soldier fly (*Hermetia illucens*) larvae under copper exposure. J Insects Food Feed. 2023;9(12):1655–61. https://doi.org/10.1163/23524588-20230020
9. Zhuang P, Zou H, Shu W. Biotransfer of heavy metals along a soil-plant-insect-chicken food chain: field study. J Environ Sci (China). 2009;21(6):849–53. https://doi.org/10.1016/s1001-0742(08)62351-7 PMID: 19803093
10. Amrul NF, Kabir Ahmad I, Ahmad Basri NE, Suja F, Abdul Jalil NA, Azman NA. A review of organic waste treatment using black soldier fly (*Hermetia illucens*). Sustainability. 2022;14(8):4565. https://doi.org/10.3390/su14084565
11. Abd El-Hack M, Shafi M, Alghamdi W, Abdelnour S, Shehata A, Noreldin A, et al. Black soldier fly (*Hermetia illucens*) meal as a promising feed ingredient for poultry: a comprehensive review. Agriculture. 2020;10(8):339. https://doi.org/10.3390/agriculture10080339
12. Almeida C, Murta D, Nunes R, Baby AR, Fernandes A, Barros L, et al. Characterization of lipid extracts from the *Hermetia illucens* larvae and their bioactivities for potential use as pharmaceutical and cosmetic ingredients. Heliyon. 2022;8(5):e09455. https://doi.org/10.1016/j.heliyon.2022.e09455; PMID: 35637671
13. Birney E, Daniel Andrews T, Bevan P, Caccamo M, Chen Y, Clarke L, et al. An overview of Ensembl. Genome Res. 2004;14(5):925–8. https://doi.org/10.1101/gr.1860604; PMID: 15078858
14. Generalovic TN, McCarthy SA, Warren IA, Wood JMD, Torrance J, Sims Y, et al. A high-quality, chromosome-level genome assembly of the Black Soldier Fly (*Hermetia illucens L.*). G3 (Bethesda). 2021;11(5):jkab085. https://doi.org/10.1093/g3journal/jkab085 PMID: 33734373
15. Costagli S, Abenaim L, Rosini G, Conti B, Giovannoni R. De Novo Genome Assembly at chromosome-scale of *Hermetia illucens* (*Diptera Stratiomyidae*) via PacBio and Omni-C proximity ligation technology. Insects. 2024;15(2):133. https://doi.org/10.3390/insects15020133 PMID: 38392552
16. Cornet L, Baurain D. Contamination detection in genomic data: more is not enough. Genome Biol. 2022;23(1):60. https://doi.org/10.1186/s13059-022-02619-9 PMID: 35189924
17. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. Genome Biol. 2021;22(1):173. https://doi.org/10.1186/s13059-021-02388-x PMID: 34103055
18. Thompson JA, Koestler DC. Equivalent change enrichment analysis: assessing equivalent and inverse change in biological pathways between diverse experiments. BMC Genomics. 2020;21(1):180. https://doi.org/10.1186/s12864-020-6589-x PMID: 32093613
19. Cork JM, Purugganan MD. The evolution of molecular genetic pathways and networks. Bioessays. 2004;26(5):479–84. https://doi.org/10.1002/bies.20026 PMID: 15112228

20. Deng B, Wang G, Yuan Q, Zhu J, Xu C, Zhang X, et al. Enrichment and speciation changes of Cu and Cd in black soldier fly (Hermetia illucens) larval compost and their effects on larval growth performance. Sci Total Environ. 2022;845:157299. https://doi.org/10.1016/j.scitotenv.2022.157299 PMID: 35842144
21. Tuda M, Luna-Maldonado AI. Image-based insect species and gender classification by trained supervised machine learning algorithms. Ecol Infor. 2020;60:101135. https://doi.org/10.1016/j.ecoinf.2020.101135
22. Zhou J, Li X, Mitri HS. Comparative performance of six supervised learning methods for the development of models of hard rock pillar stability prediction. Nat Hazards. 2015;79(1):291–316. https://doi.org/10.1007/s11069-015-1842-3
23. Jawalkar AP, Swetcha P, Manasvi N, Sreekala P. Early prediction of heart disease with data analysis using supervised learning with stochastic gradient boosting. J Eng Appl Sci. 2023;70(1):122.
24. Kuno EJA. Sampling and analysis of insect populations. Ann Rev Entomol. 1991;36(1):285–304.
25. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–32. https://doi.org/10.1038/nrg3920 PMID: 25948244
26. Hastie T, Tibshirani R, Friedman JH. Unsupervised learning. 2009:485–585.
27. Zhu L, Zhang C, Zhang C, Zhang Z, Nie X, Zhou X, et al. Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semisupervised learning. Appl Soft Comput. 2019;83:105596. https://doi.org/10.1016/j.asoc.2019.105596
28. Dike HU, Zhou Y, Deveerasetty KK, Wu Q. Unsupervised learning based on artificial neural network: a review. 2018 IEEE International Conference on Cyborg and Bionic Systems (CBS). IEEE; 2018.
29. Aggarwal CC, Yu PS. Outlier detection for high dimensional data. Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data. 2001.
30. Ur Rehman A, Belhaouari SB. Unsupervised outlier detection in multidimensional data. J Big Data. 2021;8(1):80.
31. Wong KC, Li Y, Zhang ZJU. Unsupervised learning in genome informatics. arXiv. 2016:405–48.
32. Maćkiewicz A, Ratajczak WJC. Principal components analysis (PCA). Geosciences. 1993;19(3):303–42.
33. Roweis SJA. EM algorithms for PCA and SPCA. NeurIPS Proc. 1997;10.
34. Deolindo CS, Kunicki ACB, Brasil FL, Moioli RC. Limitations of principal component analysis as a method to detect neuronal assemblies. 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom). IEEE; 2014.
35. Elgamal T, Yabandeh M, Aboulnaga A, Mustafa W, Hefeeda M. sPCA: scalable principal component analysis for Big Data on distributed platforms. Proceedings of the ACM SIGMOD International Conference on Management of Data. 2015.
36. Li Z, Safo SE, Long QJB. Incorporating biological information in sparse principal component analysis with application to genomic data. BMC Bioinformatics. 2017;18(1):1–12.
37. Zhu L, Lei J, Devlin B, Roeder K. Testing high-dimensional covariance matrices, with application to detecting schizophrenia risk genes. Ann Appl Stat. 2017;11(3):1810.
38. Min W, Liu J, Zhang S. Edge-group sparse PCA for network-guided high dimensional data analysis. Bioinformatics. 2018;34(20):3479–87. https://doi.org/10.1093/bioinformatics/bty362 PMID: 29726900
39. Miao R, Dong X, Liu X-Y, Lo S-L, Mei X-Y, Dang Q, et al. Dynamic meta-data network sparse PCA for cancer subtype biomarker screening. Front Genet. 2022;13:869906. https://doi.org/10.3389/fgene.2022.869906 PMID: 35711917
40. Lipovetsky SJ. PCA and SVD with nonnegative loadings. Pattern Recogn. 2009;42(1):68–76.
41. Min W, Liu J, Zhang S. Edge-group sparse PCA for network-guided high dimensional data analysis. Bioinformatics. 2018;34(20):3479–87. https://doi.org/10.1093/bioinformatics/bty362 PMID: 29726900
42. Harada H, Sato T, Nakamura H. Fgf8 signaling for development of the midbrain and hindbrain. Dev Growth Differ. 2016;58(5):437–45. https://doi.org/10.1111/dgd.12293 PMID: 27273073
43. Lewandoski M, Meyers E, Martin G. Analysis of Fgf8 gene function in vertebrate development. Cold Spring Harbor Symposia on Quantitative Biology. Cold Spring Harbor Laboratory Press; 1997.
44. Moon AM, Capecchi MR. Fgf8 is required for outgrowth and patterning of the limbs. Nat Genet. 2000;26(4):455–9. https://doi.org/10.1038/82601 PMID: 11101845
45. Ma L, Lu M-F, Schwartz RJ, Martin JF. Bmp2 is essential for cardiac cushion epithelial-mesenchymal transition and myocardial patterning. 2005.

46. Prall OWJ, Menon MK, Solloway MJ, Watanabe Y, Zaffran S, Bajolle F, et al. An Nkx2-5/Bmp2/Smad1 negative feedback loop controls heart progenitor specification and proliferation. Cell. 2007;128(5):947–59. https://doi.org/10.1016/j.cell.2007.01.042 PMID: 17350578

47. Rivera-Feliciano J, Tabin CJ. Bmp2 instructs cardiac progenitors to form the heart-valve-inducing field. Dev Biol. 2006;295(2):580–8.

48. Boni A, Urbanek K, Nascimbene A, Hosoda T, Zheng H, Delucchi F, et al. Notch1 regulates the fate of cardiac progenitor cells. Proc Natl Acad Sci U S A. 2008;105(40):15529–34. https://doi.org/10.1073/pnas.0808357105 PMID: 18832173

49. Metrich M, Bezdek Pomey A, Berthonneche C, Sarre A, Nemir M, Pedrazzini T. Jagged1 intracellular domain-mediated inhibition of Notch1 signalling regulates cardiac homeostasis in the postnatal heart. Cardiovasc Res. 2015;108(1):74–86. https://doi.org/10.1093/cvr/cvv209 PMID: 26249804

50. Urbanek K, Cabral-da-Silva MC, Ide-Iwata N, Maestroni S, Delucchi F, Zheng H, et al. Inhibition of notch1-dependent cardiomyogenesis leads to a dilated myopathy in the neonatal heart. Circ Res. 2010;107(3):429–41. https://doi.org/10.1161/CIRCRESAHA.110.218487 PMID: 20558824

51. MacGrogan D, Luna-Zurita L, de la Pompa JL. Notch signaling in cardiac valve development and disease. Birth Defects Res A Clin Mol Teratol. 2011;91(6):449–59. https://doi.org/10.1002/bdra.20815 PMID: 21563298

52. MacGrogan D, Münch J, de la Pompa JL. Notch and interacting signalling pathways in cardiac development, disease, and regeneration. Nat Rev Cardiol. 2018;15(11):685–704. https://doi.org/10.1038/s41569-018-0100-2 PMID: 30287945

53. Luxán G, D'Amato G, MacGrogan D, de la Pompa JL. Endocardial notch signaling in cardiac development and disease. Circ Res. 2016;118(1):e1–18. https://doi.org/10.1161/CIRCRESAHA.115.305350 PMID: 26635389