Automated staging of zebrafish embryos with deep learning

Rebecca A. Jones[1,2], Matthew J. Renshaw[3], Danelle Devenport[1], David J. Barry[3*]

[1] Department of Molecular Biology, Princeton University, Princeton, NJ, 08544, USA

[2] Developmental Biology Laboratory, The Francis Crick Institute, London, UK

[3] Crick Advanced Light Microscopy (CALM), The Francis Crick Institute, London, UK

* david.barry@crick.ac.uk

Abstract

The zebrafish (*Danio rerio*), is an important biomedical model organism used in many disciplines. The phenomenon of developmental delay in zebrafish embryos has been widely reported as part of a mutant or treatment-induced phenotype. However, the detection and quantification of these delays is often achieved through manual observation with reference to staging guides, which is both time-consuming and subjective. We recently reported a machine learning-based classifier, capable of quantifying the developmental delay between two populations of zebrafish embryos. Here, we build on that work by introducing a deep learning-based model (KimmelNet) that has been trained to predict the age (hours post fertilisation) of populations of zebrafish embryos. We show that when KimmelNet is tested on 2D brightfield images of zebrafish embryos, the predictions generated agree closely with those expected from established approaches to staging. Random sampling of the test data demonstrate that KimmelNet can be used to detect developmental delay between two populations with high confidence based on as few as 100 images of each population. Finally, we show that KimmelNet generalises to previously unseen data, with limited transfer learning improving this performance significantly. With the ability to analyse tens of thousands of standard brightfield microscopy images on a timescale of minutes, we envisage that KimmelNet will be a valuable resource for the developmental biology community. Furthermore, the approach we have used could easily be adapted to generate models for other organisms.

1 Introduction

The zebrafish (*Danio rerio*) is a commonly-used model organism in the field of developmental biology, with their transparent embryos making them particularly conducive to microscopy observation. Many developmental studies require accurate staging of zebrafish embryos and larvae. Although the timing of fertilisation can be estimated to within approximately 30 minutes, the hours post fertilisation (hpf) at the standard temperature of 28.5°C provides only an approximation of the actual developmental stage, because other factors, like population density and water quality, can affect maturation (Singleman and Holtzman, 2014). Even controlling for such factors, embryos within a clutch may develop at different rates (Parichy et al., 2009). Researchers therefore use both hpf and staging guides that are based on morphological criteria to stage individual embryos. These morphological features include the number of somites and the appearance of landmark structures such as the embryonic shield, tail bud and eye primordium (Kimmel et al., 1995; Westerfield, 2000).

Staging of embryos is of particular importance because many studies report "developmental delay" as part of a genetic or drug-induced phenotype (Akthar et al., 2019; Byrnes et al., 2018; Elabd et al., 2019; Farooq et al., 2019; Giraldez et al., 2005; Jia et al., 2020; Li et al., 2017). Zebrafish have also emerged as important models in which to study the effects of environmental toxins, with many of these treatments also resulting in developmental delay (Aksakal and Sisman, 2020; Li et al., 2020; Mesquita et al., 2017). Such delays are difficult to quantify without manually staging large numbers of embryos, which is subjective and time-consuming (Jeanray et al., 2015; Teixidó et al., 2018). A more standardized and automated approach to assessing developmental delay is therefore desirable, in order to avoid the substantial time cost and subjectivity associated with manual staging.

The use of image analysis has become increasingly popular in the life sciences in recent

years, facilitating the automated quantification of microscopy images in an unbiased fashion (Meijering et al., 2016). But, designing an image analysis algorithm to specifically detect the wide range of morphological features on which staging guides depend would be a challenging endeavour. However, the staging of embryos based on microscopy images could also be considered an image classification task, something to which machine learning is well-suited. Machine learning approaches, where a computer program uses algorithms and statistical models to continuously learn and improve pattern prediction, is already used widely in biological studies (Greener et al., 2022).

Machine learning has found some use in the analysis of zebrafish embryos, but mainly for phenotype classification, rather than assessing developmental delay. For example, using a support vector machine, Liu et al. (2012) achieved a classification accuracy of approximately 97% in differentiating between hatched, unhatched and dead zebrafish embryos. Jeanray et al. (2015) used a supervised machine learning approach to place brightfield images of 3 day old zebrafish larvae in one of eleven phentotypic classes. But their method involved significant image pre-processing and their classifier failed to accurately classify certain phenotypes.

More recently, convolutional neural networks (CNNs), have grown in popularity in life science research in general, and they are particularly well-suited to the analysis of image data (Greener et al., 2022). There are a number of examples in the literature of CNNs being used for zebfrafish embryo phenotype classification. Ishaq et al. (2017) used a CNN-based approach to predict the probability of deformation in multi-fish microwell plates. While few images were needed for training (< 100) and an F score of 0.915 was achieved, their network could only differentiate between two different classes. Tyagi et al. (2018) tested multiple CNN architectures to classify zebrafish embryo phenotypes, but achieved a relatively low accuracy (83.55%) when attempting to classify a large number of phenotypes (11). Using the image data from Jeanray et al. (2015), Shang et al. (2020) achieved a higher

accuracy using a two-tier CNN-based approach, but again, the number of images used for training (529) and testing (341) across ten classes was limited.

In terms of developmental staging, Pond et al. (2021) developed a convolutional neural network (CNN) to stage zebrafish tail-buds at four discrete developmental stages just 1.5 hours apart. Although they trained their network with a small number of images (< 100) and achieved 100% test accuracy in some cases, the number of test images per class was in several cases as low as 2.

We recently published a machine-learning based method for the staging of zebrafish embryos (Jones et al., 2022). While this method successfully discriminated between different populations of zebrafish developing at different rates, the developmental profiles it produced were distinctly non-linear. This was largely due to the fact that the underlying classifier was trained to recognise two distinct classes corresponding to two different stages of development (embryos at 4.5hpf and 17.5hpf). As such, it performed poorly at accurately staging embryos in between those specific time-points, producing a developmental profile with a distinctive 'S' shape.

Here we present a more robust classifier, based on a convolutional neural network, that provides more accurate predictions of developmental stage across a much wider range of timepoints (4.5 to 51 hpf). Named KimmelNet and using simple brightfield microscopy images as training data (with no manual annotation required), the developmental profiles produced by our CNN closely agree with those predicted by established approaches (Kimmel et al., 1995). We demonstrate that KimmelNet accurately quantifies developmental delay between two populations of zebrafish embryos. We also demonstrate that KimmelNet can succesfully distinguish between populations in previously "unseen" data (that is, completely distinct from the training data) and this performance can be improved significantly with modest levels of transfer learning. We envisage that our classifier will become a valuable resource for the zebrafish community.

2 Results & Discussion

2.1 Network Architecture

The architecture of KimmelNet is based on that described by Ishaq et al. (2017), which is in turn a simplified version of AlexNet (Krizhevsky et al., 2012). To reduce the number of trainable parameters (and change the nature of the network from classification to regression), we simplified the architecture further still (Fig. 1). The architecture consists of repeating units of 3 layers, consisting of 3 x 3 2D convolution, a Rectified Linear Unit (ReLU) and 2 x 2 max pooling. The output consists of a single fully connected layer, a dropout layer and a final fully connected (dense) layer consisting of a single output neuron. The total number of repeating convolution units, the number of convolutional filters in each and the dropout rate in the second-to-last layer were all determined using hyperparameter tuning, implemented using KerasTuner (Chollet et al., 2015).

2.2 Training KimmelNet

KimmelNet was trained on 23,370 individual images of zebrafish embryos grown at 28.5°C, ranging in age from 4.5 to 51.75 hours post-fertilisation (hpf; Fig. 2A, Table 1). We trained KimmelNet for 500 epochs as, for a greater number of iterations, the validation loss began to plateau (while the training loss continued to decrease slightly), indicative of over-fitting. After 500 epochs, training loss was approximately 8.5 (Fig. 2B). Training took approximately 2.5 hours on the High Performance Computing (HPC) Cluster at the Francis Crick Institute (see Methods for hardware specifications).

2.3 KimmelNet detects developmental delay with high confidence

We evaluated KimmelNet on two test datasets, one of zebrafish embryos grown at 28.5°C and one at 25.0°C (Datasets 'A' and 'B' in Table 1). Testing KimmelNet on Dataset A took approximately 160 seconds on the Crick's HPC hardware (see Methods), equating to just under 7 milliseconds per image.

To evaluate the accuracy of predicted hpf values, we fit an equation of the form $y = mx$ to predictions produced by KimmelNet. We used an equation of this form in order to be consistent with that used by Kimmel et al. (1995):

$$H_T = \frac{h}{(0.055T - 0.57)} \tag{1}$$

where H_T corresponds to hours of development at temperature T and h denotes the number of hours required to reach the equivalent developmental stage at 28.5°C.

Fitting an equation of the form $y = mx$ to Dataset A predictions resulted in a line with a slope equal to 0.974 (R^2 = 0.951) - a slope of 1.0 would indicate a perfect 1-to-1 correspondence between true hpf labels and predicted values (Fig. 3A).Saliency maps to illustrate pixel attribution, generated using an implementation of DeconvNet (Zeiler and Fergus, 2013), indicated that network predictions were driven by morphological features of the embryos, rather than any irrelevant features of the image background (Fig. 4).

For Dataset B (Table 1), Equation 1 predicts that embryos will develop at approximately 80.5% of the rate of those grown at 28.5°C, equivalent to a straight line with slope 0.805. A straight line fit to the predictions produced by KimmelNet reveal a lower rate of development (slope = 0.725, $R^2 = 0.949$; Fig. 3A). However, the straight line corresponding to Equation 1 falls within the range of predictions produced by KimmelNet (Fig. 3A).

To determine how KimmelNet would perform on smaller test datasets, we took random samples of the data points shown in Figure 3A and fit lines to these points. Performing this

sampling 10,000 times for either 100 or 200 random data points produced the range of line fits shown in Figure 3B. From approximately 2 hpf, there is no overlap between the confidence intervals for the two test datasets. This demonstrates that KimmelNet can reliably distinguish between two populations of zebrafish embryos developing at different rates given just 100 images of each.

Plotting the residuals to the straight line fits showed that the mean prediction errors for both test datasets were distributed approximately uniformly about zero (Fig. 3C & D). For Dataset A, the mean prediction error was 0.364 ± 0.019 h (mean $\pm$ s.e.m.) and the standard deviation was 2.824 h. For Dataset B, the mean prediction error was 0.302 ± 0.017 h and the standard deviation was 2.136 h.

It is possible that the use of a more sophisticated model, or modification of the approach to model training, could result in a narrower distribution of prediction errors. However, it is also possible that these errors result from the approach used for annotating the training data. Each image in the training dataset is annotated with its associated hpf. However, it is known that different embryos within the same clutch can develop at different rates (Parichy et al., 2009). It is therefore very likely that images in our training data corresponding to x hpf depict embryos at slightly different developmental stages. However, the aim of this study was not to develop a tool to accurately stage individual embryos. The intention of this study is to demonstrate that, even in the absence of time-consuming manual annotation of training data, it is possible to differentiate between populations developing at different rates using deep learning.

2.4 KimmelNet can generalise to previously unseen data

In order for a deep-learning model to useful to the broader community, it needs to generalise to previously unseen data. We challenged KimmelNet with two new datasets, acquired at a different site to the training data (‘C’ and ‘D’ in Table 1, Figure 5A). Even though the

predicted hpfs produced by KimmelNet contained larger errors (Fig. 5 D and E) than those shown in Figure 3, it was still possible to distinguish between two populations incubated at different temperatures (Fig. 5C). However, trendlines fit to the predictions in Figure 5B were a poor fit ($R^2 = 0.64$ for Dataset C and $R^2 = 0.53$ for Dataset B) relative to Figure 3A

2.5 KimmelNet's predictions can be refined with transfer learning

We investigated whether the results shown in Figure 5 B-E could be improved by way of transfer learning, to fine-tune the model to new data. We ran a variety of different transfer learning regimes (Table 2), experimenting with the number of training images used and the proportion of KimmelNet's tunable parameters that required updating to improve results.

We found that the number of training images used had a significant impact on the outcome of the transfer learning, while the proportion of parameters in the model that were updated had a comparatively small influence (Table 2, Fig. 6, 7). For example, using 12 wells from the original 96 well plate and retraining 23% of KimmelNet's parameters resulted in a reduction in the mean prediction error from 1.30 $\pm$ 5.52 h (Table 1, Fig. 5D) to 0.90 $\pm$ 3.85 h (Table 2, Fig. 7D).

Overall, these results demonstrate that KimmelNet can be used "as-is" without any further training on new data. However, in the event that subtle differences between populations cannot be elucidated, a minimal amount of transfer learning can significantly improve the predictive power of the model.

3 Conclusion

The developing zebrafish embryo is used in many different types of studies and accurate staging is essential. When comparing an experimental group of embryos with a control group, ensuring the embryos have reached the same developmental stage allows for

meaningful comparisons to be made. KimmelNet enables rapid, unbiased assessment of thousands of images with minimal time commitment. We anticipate that our classifier will be a useful tool for the zebrafish community. There is also potential for KimmelNet to be useful in other developmental contexts, by retraining with images of other organisms.

4 Acknowledgements

We are grateful for the support provided by the Princeton Core Imaging Facility, particularly Gary S. Laevsky, and Triveni Menon of Rebecca D. Burdine's lab in Princeton for help with provision of zebrafish embryos.

5 Conflict of Interest

The authors declare that they have no conflict of interest.

6 Materials & Methods

6.1 Image Data Acquisition

All images used for training and testing are summarised in Table 1. The details of acquisition of the images comprising Datasets 'A' and 'B' can be found in Jones et al. (2022). Datasets 'C' and 'D' were acquired as follows.

6.1.1 Zebrafish Husbandry

Zebrafish were maintained according to the protocols outlined by Princeton University Institutional Animal Care and Use Committee (Approval number- 1915). Wild type embryos were collected from the PWT strain generated in Burdine lab (Schottenfeld et al.,

2007). Male and female adult fish separated by barriers were set up for breeding the night before the experiment. In the morning upon lights on, embryos laid within 30 minutes of opening the barriers were collected and treated as a single clutch. Embryos were raised at 28°C in 1x E3 embryo medium without methylene blue until plating for microscopy.

6.1.2 Live Imaging

Zebrafish embryos were maintained at 28°C until 2-3 hpf when they were transferred into a U-bottomed 96-well plate in E3 medium. The plate was covered with an FEP membrane (1 mil Teflon FEP film, American Durafilm) to prevent condensation and allow for gas exchange. Brightfield images of embryos individually seeded in the 96-well plate were acquired every 15 minutes starting at 2.5-3.5 hpf for approximately 43 hours on a Nikon Ti-E inverted microscope with a Hamamatsu BT-Fusion sCMOS camera, using a 4X/0.13 Plan Fluor objective. Sample temperature was maintained at either 25.0 or 28.5°C using a Tokaii Hit stage top incubator. The microscope was controlled with Nikon NIS- Elements Software. Images were manually checked after capture to ensure embryo health.

6.2 Data Partitioning and Augmentation

Dataset 'A' (Table 1) was split randomly into training and test data on a well-by-well basis. Certain wells were manually excluded if it was apparent that the embryos died or were significantly out of view - the complete list of excluded wells is explicitly listed in the code (see github.com/djpbarry/KimmelNET/blob/main/train_model.py).

Six data augmentation layers were used for training purposes to prevent over-fitting to the training data (Fig. 1). These consisted of randomly applied histogram equalisation, randomly applied image saturation, randomly added Gaussian noise, random horizontal and vertical image flipping, random translation in the x-direction (max 20% of image width) and random zoom (maximum 30%). Random translation in the y-direction was avoided due to

the aspect ratio of the images (width $>$ height) - any artifacts introduced by translation in the x-direction would be removed by the centre crop layer, but this would not be the case for translation in the y-direction.

6.3 Network Training

KimmelNet was trained on a single GPU node on the Francis Crick Institute's High Performance Computing (HPC) platform (Crick data Analysis and Management Platform - CAMP). Each node consists of 4 NVIDIA Tesla V100 SXM2 graphics cards. KimmelNet was trained on all 4 graphics cards simultaneously using the Tensorflow's distribute library. Training was performed using the Adam optimiser (Kingma and Ba, 2014) with a learning rate of 0.0005 and mean squared error as the loss function.

6.4 Bootstrapping Test Predictions

To test the correspondence between predictions made by KimmelNet and those made by the equation published by Kimmel et al. (1995), test predictions were fit to a straight line through the origin of the form $y = mx$. To calculate confidence intervals associated with these fits, a series of 10,000 line fits were performed, each with a different random sample of the test data, consisting of either 100 or 200 data points. The maximum and minimum values of m produced by these line fits were then interpreted as a confidence range for the line fit to the complete data set.

6.5 Transfer Learning

Transfer learning was performed on the Francis Crick Institute's HPC platform, using an individual graphics card. Training was performed for 1,200 epochs using the Adam optimiser (Kingma and Ba, 2014) with a learning rate of 0.0001 and mean squared error as

the loss function. Note that the learning rate is five times lower than that used for the initial training, to prevent the model converging too quickly to a local minimum. Data was partitioned into training and test datasets (Table 2) by randomly selecting wells for training.

6.6 Statistical Analysis

Calculation of mean and standard deviation (sd) were performed using mean and std functions in Numpy (Harris et al., 2020). The standard error of the mean (sem) was calculated using the sem function in the SciPy stats module (Virtanen et al., 2020).

6.7 Code Implementation

KimmelNet was implemented using Python 3.9 (www.python.org) and Tensorflow 2.9 (www.tensorflow.org). All plots in this manuscript were produced using Matplotlib 3.5 (Hunter, 2007) and all curve-fitting performed with SciPy 1.8 (Virtanen et al., 2020). Training and testing were performed with NVIDIA CUDA 11.2 (developer.nvidia.com/cuda-zone) and cuDNN 8.1 (developer.nvidia.com/cudnn).

7 Funding

This work was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (FC001157, FC001999), the UK Medical Research Council (FC001157, FC001999) and the Wellcome Trust (FC001157, FC001999).

8 Data Availability

All image data used in this study is available to download from the BioImage Archive (accession numbers S-BIAD531 and S-BIAD840).

9 Software Availability

All code for building, training and testing KimmelNet is publicly available online (github.com/djpbarry/KimmelNET). A Binder implementation is also available to test KimmelNet on a subset of our test data (mybinder.org/v2/gh/djpbarry/KimmelNET/main?labpath=zebrafish_age_estimator.ipynb).

References

F. I. Aksakal and T. Sisman. Developmental toxicity induced by $Cu(OH)_2$ nanopesticide in zebrafish embryos. *Environmental Toxicology*, 35(12):1289–1298, 2020. doi: https://doi.org/10.1002/tox.22993.

I. S. T. Akthar, P. T. Pichiah, S. Arunachalam, and S. Raja. Adriamycin inhibits embryonic development in zebrafish through downregulation of Kruppel-like factor4. *Journal of Biochemical and Molecular Toxicology*, 33(1):e22235, 2019. doi: https://doi.org/10.1002/jbt.22235.

J. Byrnes, R. Ganetzky, R. Lightfoot, M. Tzeng, E. Nakamaru-Ogiso, C. Seiler, and M. J. Falk. Pharmacologic modeling of primary mitochondrial respiratory chain dysfunction in zebrafish. *Neurochemistry International*, 117:23–34, 2018. doi: https://doi.org/10.1016/j.neuint.2017.07.008.

F. Chollet et al. Keras. `https://keras.io`, 2015.

S. Elabd, N. A. Jabeen, V. Gerber, R. Peravali, J.-C. Bourdon, S. Kancherla, D. Vallone, and C. Blattner. Delay in development and behavioural abnormalities in the absence of p53 in zebrafish. *PLOS ONE*, 14(7):1–18, 07 2019. doi: 10.1371/journal.pone.0220069.

M. Farooq, A. Sharma, Z. Almarhoon, A. Al-Dhfyan, A. El-Faham, N. A. Taha, M. A. Wadaan, B. G. de la Torre, and F. Albericio. Design and synthesis of mono-and di-pyrazolyl-s-triazine derivatives, their anticancer profile in human cancer cell lines, and in vivo toxicity in zebrafish embryos. *Bioorganic Chemistry*, 87:457–464, 2019. doi: https://doi.org/10.1016/j.bioorg.2019.03.063.

A. J. Giraldez, R. M. Cinalli, M. E. Glasner, A. J. Enright, J. M. Thomson, S. Baskerville, S. M. Hammond, D. P. Bartel, and A. F. Schier. MicroRNAs Regulate Brain Morphogenesis in Zebrafish. *Science*, 308(5723):833–838, 2005. doi: 10.1126/science.1109020.

J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones. A guide to machine learning for biologists. *Nature Reviews Molecular Cell Biology*, 23:40–55, 2022.

C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020. doi: 10.1038/s41586-020-2649-2.

M. Hartley, G. J. Kleywegt, A. Patwardhan, U. Sarkans, J. R. Swedlow, and A. Brazma. The bioimage archive – building a home for life-sciences microscopy data. *Journal of Molecular Biology*, 434(11):167505, 2022. doi: https://doi.org/10.1016/j.jmb.2022.167505.

J. D. Hunter. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007. doi: 10.1109/MCSE.2007.55.

O. Ishaq, S. K. Sadanandan, and C. Wählby. Deep Fish: Deep Learning–Based

Classification of Zebrafish Deformation for High-Throughput Screening. *SLAS Discovery*, 22(1):102–107, 2017. doi: https://doi.org/10.1177/1087057116667894.

N. Jeanray, R. Marée, B. Pruvot, O. Stern, P. Geurts, L. Wehenkel, and M. Muller. Phenotype Classification of Zebrafish Embryos by Supervised Learning. *PLOS ONE*, 10 (1):1–20, 01 2015. doi: 10.1371/journal.pone.0116989.

S. Jia, X. Wu, Y. Wu, X. Cui, B. Tao, Z. Zhu, and W. Hu. Multiple Developmental Defects in ¡i¿sox11a¡/i¿ Mutant Zebrafish with Features of Coffin-Siris Syndrome. *Int J Biol Sci*, 16:3039–3049, 2020. doi: 10.7150/ijbs.47510.

R. Jones, M. Renshaw, D. Barry, and J. Smith. Automated staging of zebrafish embryos using machine learning [version 1; peer review: awaiting peer review]. *Wellcome Open Research*, 7(275), 2022. doi: 10.12688/wellcomeopenres.18313.1.

C. B. Kimmel, W. W. Ballard, S. R. Kimmel, B. Ullmann, and T. F. Schilling. Stages of embryonic development of the zebrafish. *Developmental Dynamics*, 203(3):253–310, 1995. doi: https://doi.org/10.1002/aja.1002030302.

D. P. Kingma and J. Ba. Adam: A Method for Stochastic Optimization, 2014.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. P. Weinberger, C. Burges, L. Bottou, and K.Q., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

S.-Z. Li, W. Liu, Z. Li, W.-H. Li, Y. Wang, L. Zhou, and J.-F. Gui. greb1 regulates convergent extension movement and pituitary development in zebrafish. *Gene*, 627: 176–187, 2017. doi: https://doi.org/10.1016/j.gene.2017.06.016.

W. Li, M. Yuan, Y. Wu, and X. Liu. Bixafen exposure induces developmental toxicity in zebrafish (Danio rerio) embryos. *Environmental Research*, 189:109923, 2020. doi: https://doi.org/10.1016/j.envres.2020.109923.

R. Liu, S. Lin, R. Rallo, Y. Zhao, R. Damoiseaux, T. Xia, S. Lin, A. Nel, and Y. Cohen. Automated Phenotype Recognition for Zebrafish Embryo Based In Vivo High Throughput Toxicity Screening of Engineered Nano-Materials. *PLOS ONE*, 7(4):1–10, 04 2012. doi: 10.1371/journal.pone.0035014.

E. Meijering, A. E. Carpenter, H. Peng, F. A. Hamprecht, and J.-C. Olivo-Marin. Imagining the future of bioimage analysis. *Nature Biotechnology*, 34:1250–1255, 2016.

B. Mesquita, I. Lopes, S. Silva, M. J. Bessa, M. Starykevich, J. Carneiro, T. L. P. Galvão, M. G. S. Ferreira, J. Tedim, J. P. Teixeira, and S. Fraga. Gold nanorods induce early embryonic developmental delay and lethality in zebrafish (Danio rerio). *Journal of Toxicology and Environmental Health, Part A*, 80(13-15):672–687, 2017. doi: 10.1080/15287394.2017.1331597.

D. M. Parichy, M. R. Elizondo, M. G. Mills, T. N. Gordon, and R. E. Engeszer. Normal table of postembryonic zebrafish development: Staging by externally visible anatomy of the living fish. *Developmental Dynamics*, 238(12):2975–3015, 2009. doi: https://doi.org/10.1002/dvdy.22113.

A. J. R. Pond, S. Hwang, B. Verd, and B. Steventon. A deep learning approach for staging embryonic tissue isolates with small data. *PLOS ONE*, 16(1):1–17, 01 2021. doi: 10.1371/journal.pone.0244151.

J. Schottenfeld, J. Sullivan-Brown, and R. D. Burdine. Zebrafish curly up encodes a Pkd2 ortholog that restricts left-side-specific expression of southpaw. *Development*, 134(8): 1605–1615, 04 2007. doi: 10.1242/dev.02827.

S. Shang, S. Lin, and F. Cong. Zebrafish Larvae Phenotype Classification from Bright-field Microscopic Images Using a Two-Tier Deep-Learning Pipeline. *Applied Sciences*, 10(4), 2020. doi: 10.3390/app10041247.

C. Singleman and N. G. Holtzman. Growth and maturation in the zebrafish, Danio rerio: a staging tool for teaching and research. *Zebrafish*, 11(4):396–406, 2014.

E. Teixidó, T. R. Kießling, E. Krupp, C. Quevedo, A. Muriana, and S. Scholz. Automated Morphological Feature Assessment for Zebrafish Embryo Developmental Toxicity Screens. *Toxicological Sciences*, 167(2):438–449, 10 2018. doi: 10.1093/toxsci/kfy250.

G. Tyagi, N. Patel, and I. Sethi. A Fine-Tuned Convolution Neural Network Based Approach For Phenotype Classification Of Zebrafish Embryo. *Procedia Computer Science*, 126:1138–1144, 2018. doi: https://doi.org/10.1016/j.procs.2018.08.051.

P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020. doi: 10.1038/s41592-019-0686-2.

M. Westerfield. *The Zebrafish Book: A Guide for the Laboratory Use of Zebrafish (Danio Rerio)*. University of Oregon Press, 2000.

M. D. Zeiler and R. Fergus. Visualizing and Understanding Convolutional Networks, 2013.

Figure Legends

Fig 1. Illustration of the KimmelNet architecture. Augmentation layers active for training only are shown in blue.

Fig 2. KimmelNet converged to a training loss of approximately 8.5 after 500 epochs A. Example images of zebrafish embryos at various indicated stages of development - labels indicate hours post-fertilisation (hpf). Images were drawn from randomly selected wells in two different multi-well plates (datasets 'A' and 'B' in Table 1). B. Plot of training and validation loss during training of KimmelNet with approximately 23,000 images similar to those shown in 'A'.

Fig 3. KimmelNet successfully detects different developmental in two populations of zebrafish embryos A. Predictions produced by KimmelNet for Datasets 'A' and 'B' (Table 1) closely agree with those predicted by Equation 1. Each data point represents a single image. The solid lines represent the line of best fit (through the origin) to KimmelNet's predictions, while the dotted lines indicate the expected rate of development for the indicated temperatures based on Equation 1. B. Confidence intervals associated with lines of best fit illustrate that, even with just a small number of images, KimmelNet can discriminate between two different populations. The inner and outer confidence intervals indicate the range of lines of best fit determined for 10,000 randomly chosen sets of 100 (outer) or 200 (inner) data points from 'A'. C and D. Distribution of prediction errors produced by KimmelNet relative to lines of best fit (sold lines in 'A') and Equation 1 (dotted lines in 'A').

Fig 4. Saliency maps illustrate that KimmelNet responds to image pixels relevant to the zebrafish embryos Image background contributes little to predictions.

Fig 5. KimmelNet can distinguish populations based on previously unseen data. A. Predictions produced by KimmelNet for Datasets ‘C’ and ‘D’ (Table 1). Each data point represents a single image. The solid lines represent the line of best fit (through the origin) to KimmelNet’s predictions, while the dotted lines indicate the expected rate of development for the indicated temperatures based on Equation 1. B. Confidence intervals associated with lines of best fit illustrate that, even with just a small number of images, KimmelNet can discriminate between two different populations. The inner and outer confidence intervals indicate the range of lines of best fit determined for 10,000 randomly chosen sets of 100 (outer) or 200 (inner) data points from ’A’. C and D. Distribution of prediction errors produced by KimmelNet relative to lines of best fit (sold lines in ’A’) and Equation 1 (dotted lines in ’A’).

Fig 6. Transfer learning results in more linear predictions by KimmelNet. Labels ‘A’ - ‘I’ correspond to Datasets in Table 2. The number of images used for transfer learning increases when moving from bottom to top. The number of network parameters retrained increases when moving from left to right.

Fig 7. Transfer learning reduces the magnitude of errors produced by KimmelNet. Labels ‘A’ - ‘I’ correspond to Datasets in Table 2. The number of images used for transfer learning increases when moving from bottom to top. The number of network parameters retrained increases when moving from left to right.

Table 1. Overview of datasets used for training and testing. Datasets A and B are the same as those used in a previously published study (Jones et al., 2022). All data is available to download from the BioImage Archive (Hartley et al., 2022).

Dataset	BioImage Archive Accession No.	Imaging Location	Incubation Temp. (°C)	Number of Plates	Number of Images	% Images Used for Training	% Images Used for Testing	Prediction Errors (mean $\pm$ sd)	Slope of Line Fit to Predictions	R^2 of Line Fit to Predictions
A	S-BIAD531	Crick	28.5	3	50,730	46	54	0.36 ± 2.82	0.974	0.951
B	S-BIAD531	Crick	25.0	1	18,050	0	100	0.30 ± 2.14	0.725	0.949
C	S-BIAD840	Princeton	28.5	1	16,512	0	100	1.30 ± 5.52	0.923	0.644
D	S-BIAD840	Princeton	25.0	1	16,896	0	100	1.18 ± 5.82	0.773	0.533

Table 2. Partitioning of dataset 'C' in Table 1 for transfer learning and subsequent testing.

Dataset	% Images Used for Transfer Learning	% Model Parameters Retrained	Test Prediction Errors (mean $\pm$ sd)	Slope of Line Fit to Test Predictions	R^2 of Line Fit to Test Predictions
A*	38.5	23	0.64 ± 4.16	0.938	0.847
B*	38.5	41	0.60 ± 3.83	0.955	0.874
C*	38.5	68	0.48 ± 3.84	0.957	0.881
D*	12.5	23	0.90 ± 3.85	0.930	0.843
E*	12.5	41	0.72 ± 3.69	0.925	0.866
F*	12.5	68	0.65 ± 3.85	0.932	0.864
G*	2	23	1.47 ± 5.33	1.007	0.700
H*	2	41	1.42 ± 5.42	1.004	0.702
I*	2	68	1.43 ± 5.53	1.015	0.699

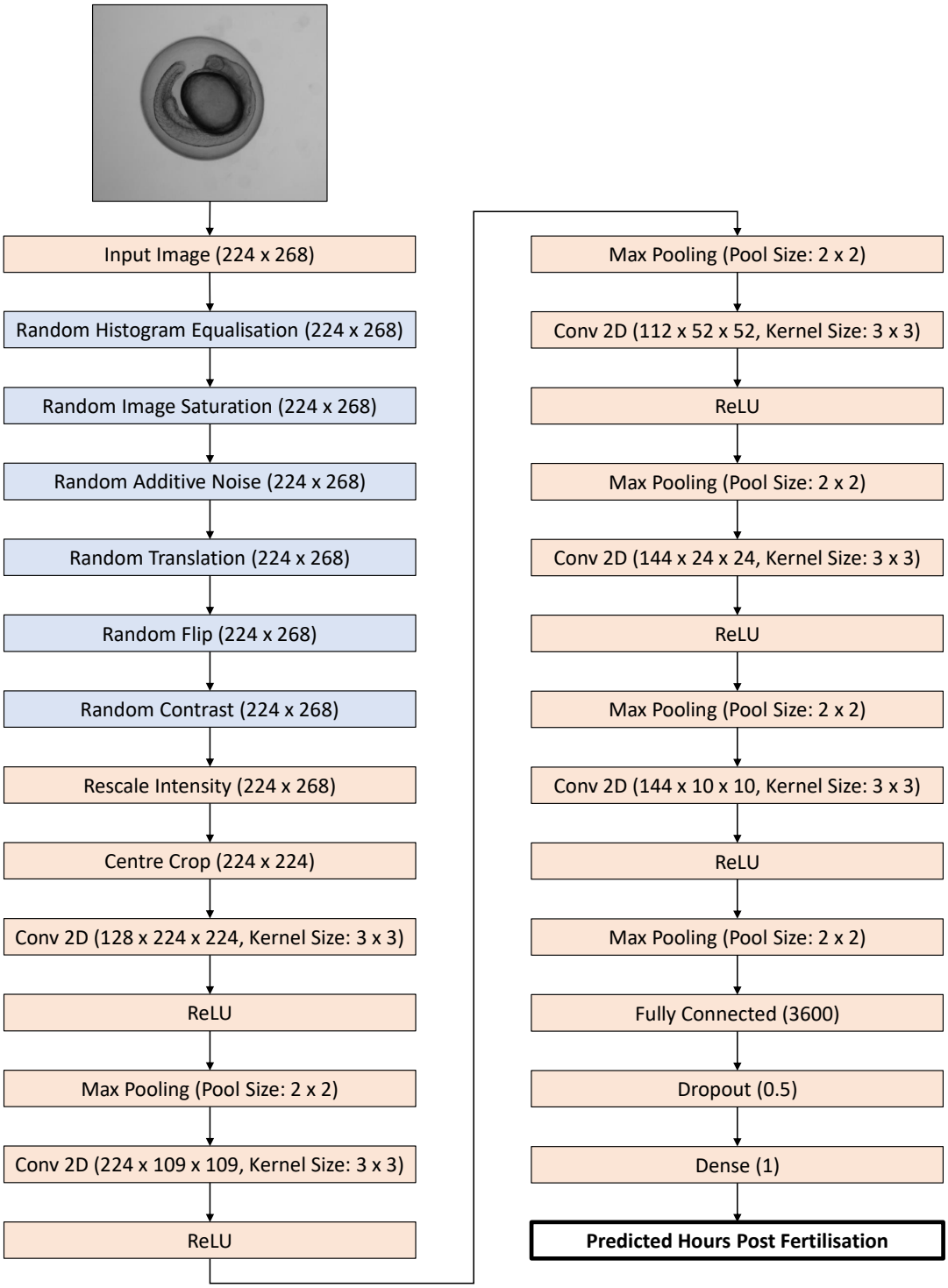
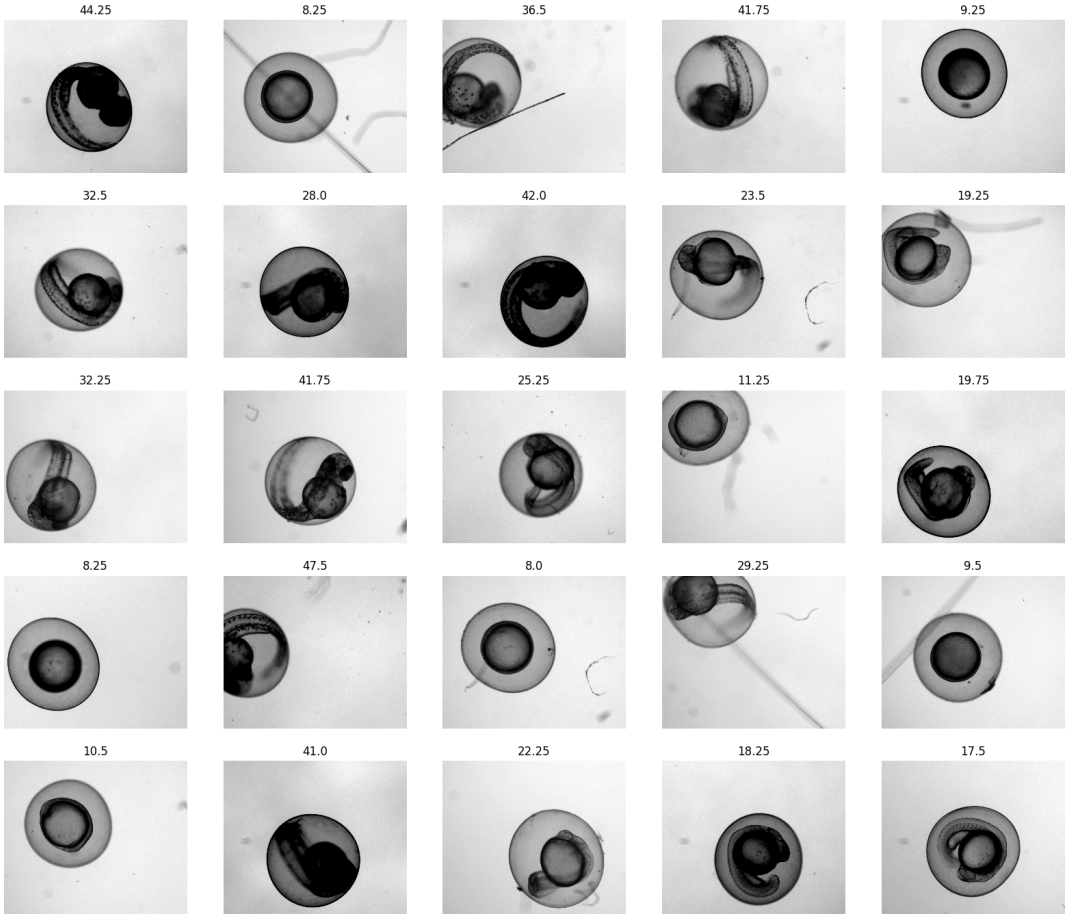

Input Image (224 x 268)
Random Histogram Equalisation (224 x 268)
Random Image Saturation (224 x 268)
Random Additive Noise (224 x 268)
Random Translation (224 x 268)
Random Flip (224 x 268)
Random Contrast (224 x 268)
Rescale Intensity (224 x 268)
Centre Crop (224 x 224)
Conv 2D (128 x 224 x 224, Kernel Size: 3 x 3)
ReLU
Max Pooling (Pool Size: 2 x 2)
Conv 2D (224 x 109 x 109, Kernel Size: 3 x 3)
ReLU
Max Pooling (Pool Size: 2 x 2)
Conv 2D (112 x 52 x 52, Kernel Size: 3 x 3)
ReLU
Max Pooling (Pool Size: 2 x 2)
Conv 2D (144 x 24 x 24, Kernel Size: 3 x 3)
ReLU
Max Pooling (Pool Size: 2 x 2)
Conv 2D (144 x 10 x 10, Kernel Size: 3 x 3)
ReLU
Max Pooling (Pool Size: 2 x 2)
Fully Connected (3600)
Dropout (0.5)
Dense (1)
Predicted Hours Post Fertilisation

Figure 1

A.


44.25
8.25
36.5
41.75
9.25
32.5
28.0
42.0
23.5
19.25
32.25
41.75
25.25
11.25
19.75
8.25
47.5
8.0
29.25
9.5
10.5
41.0
22.25
18.25
17.5

B.

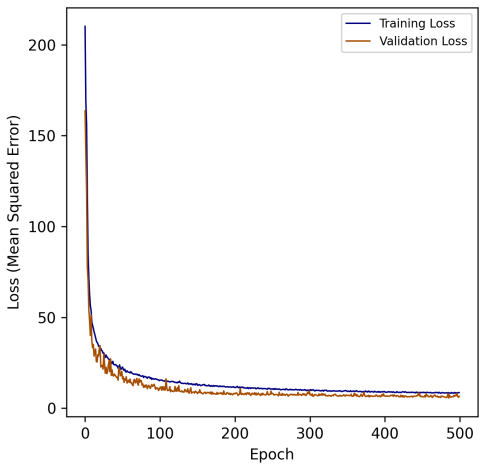
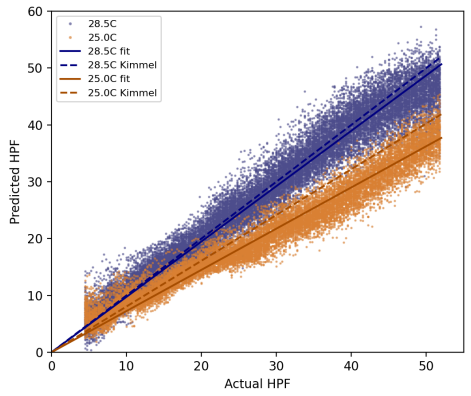
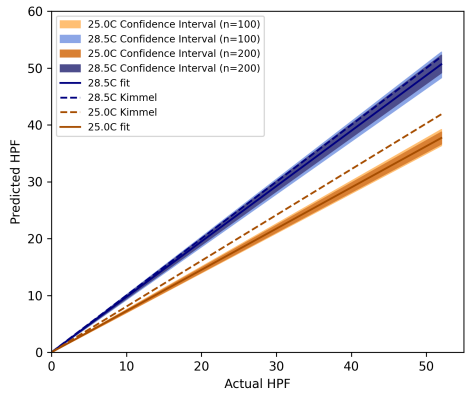

Training Loss
Validation Loss
Loss (Mean Squared Error)
Epoch
0
50
100
150
200
0
100
200
300
400
500

Figure 2

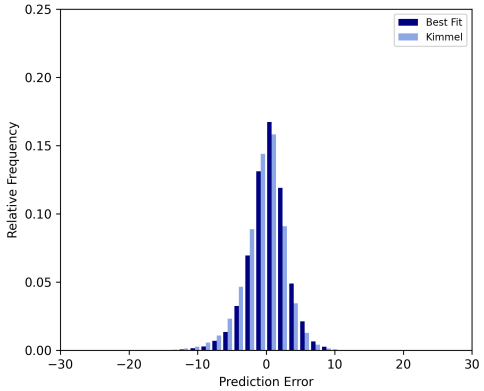
A.



B.



C.



D.

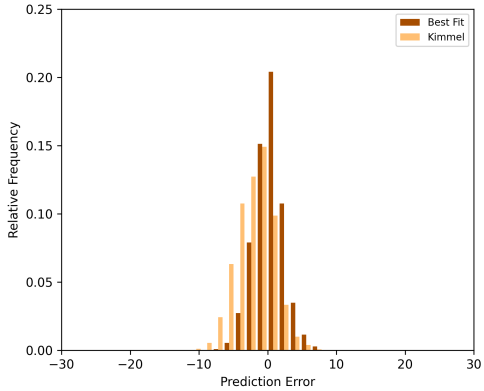


Figure 3

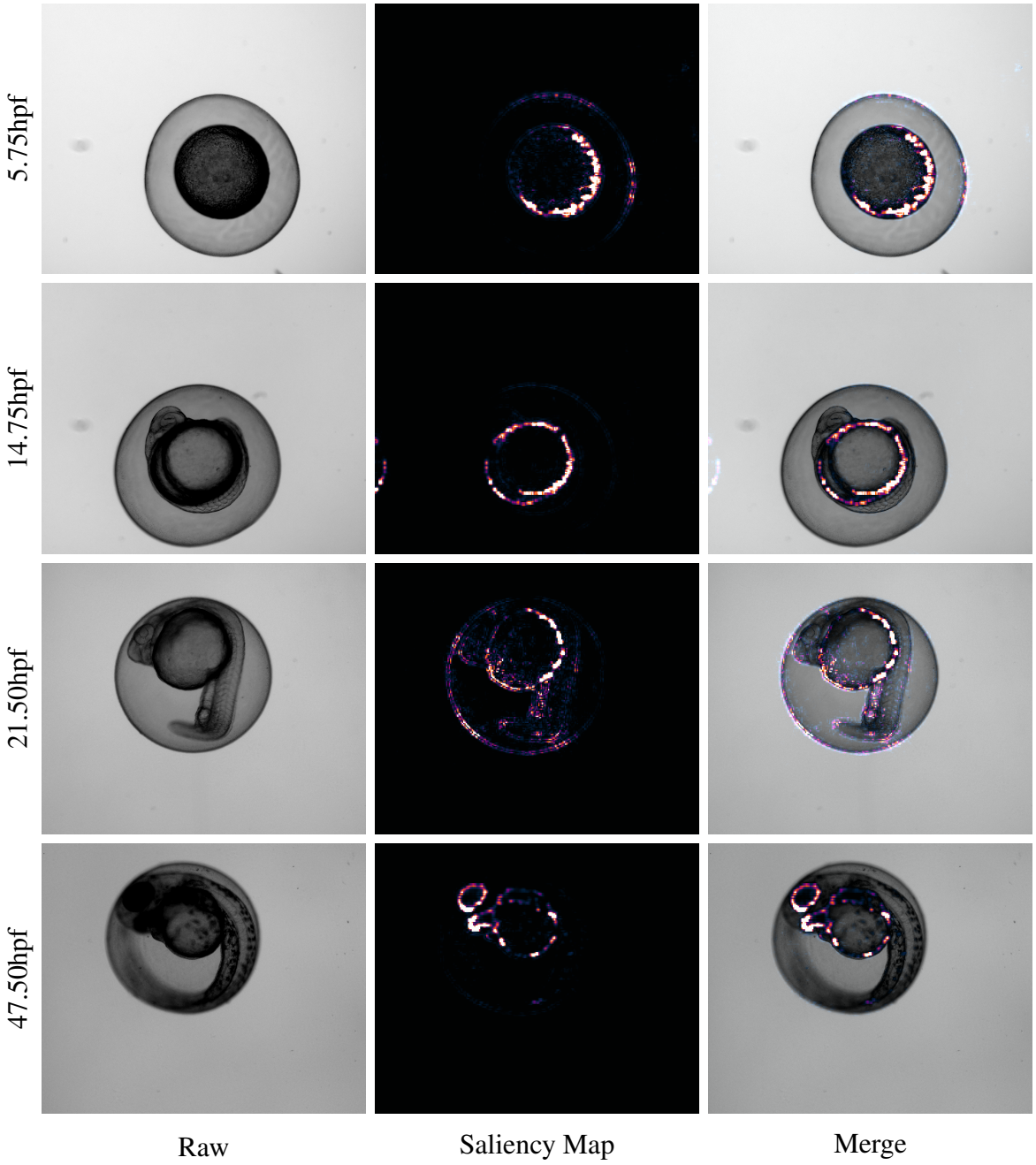
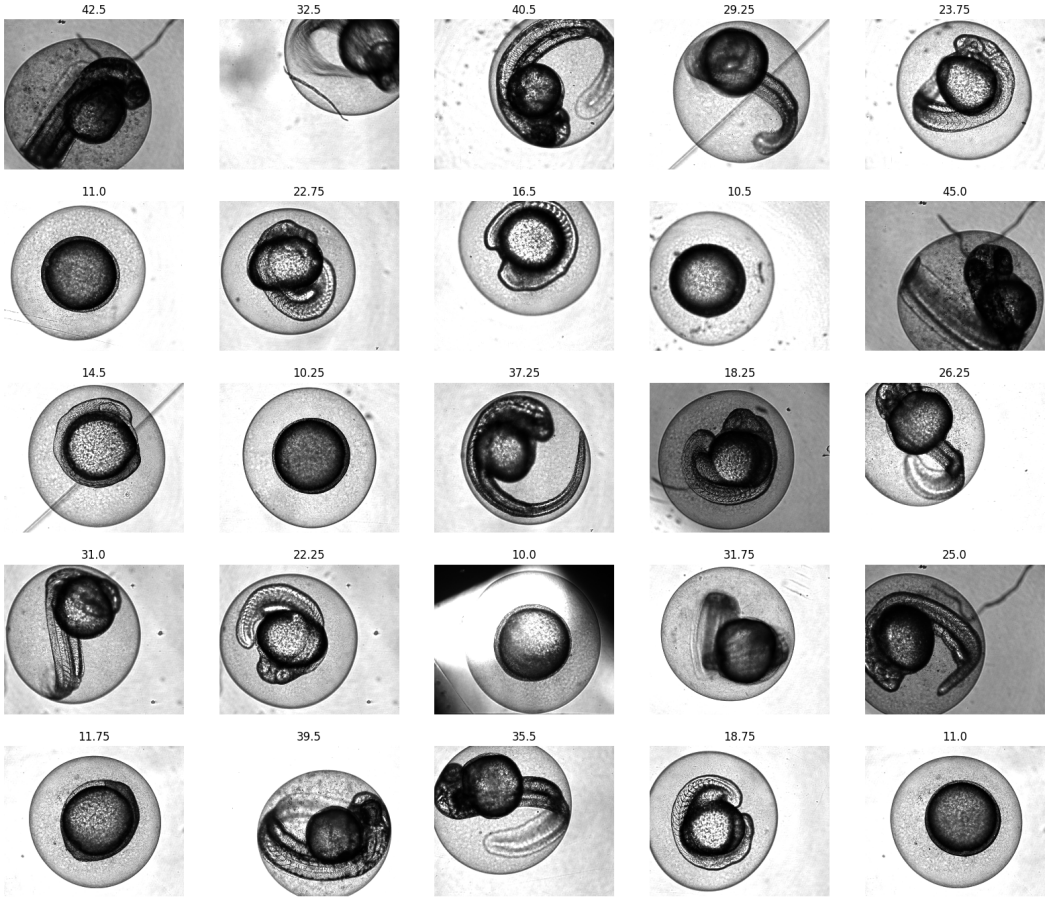


Figure 4

A.



B. C. D. E.

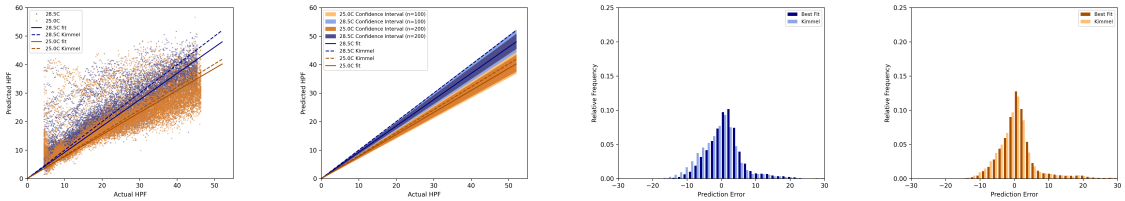


Figure 5

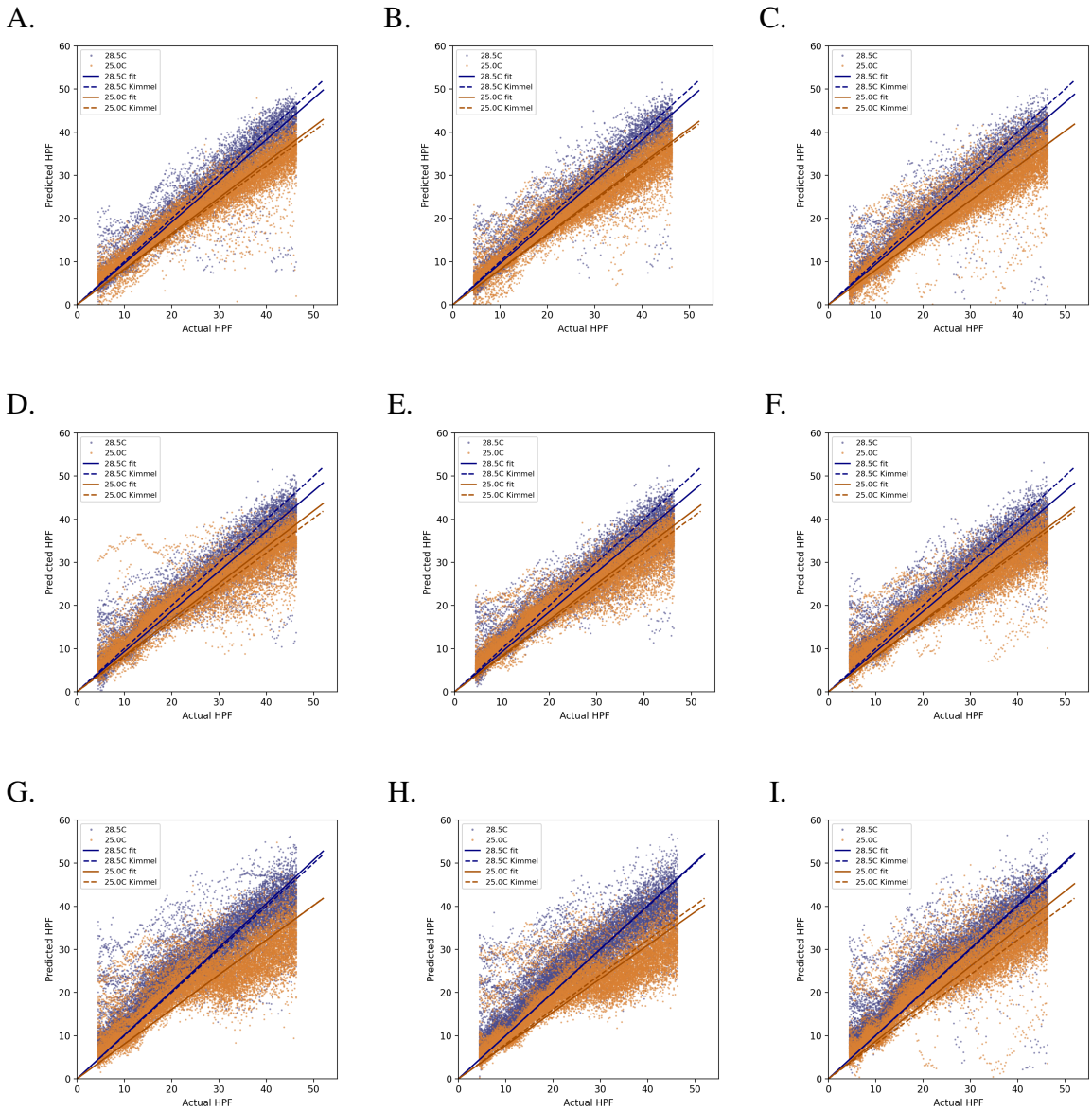
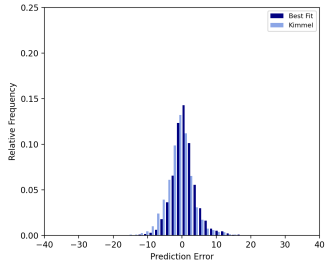
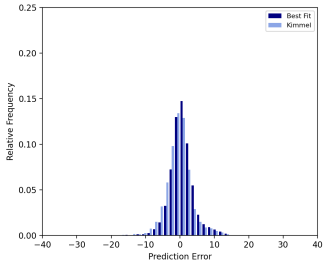

A.
B.
C.
D.
E.
F.
G.
H.
I.
28.5C
25.0C
28.5C fit
28.5C Kimmel
25.0C fit
25.0C Kimmel
Predicted HPF
Actual HPF

Figure 6

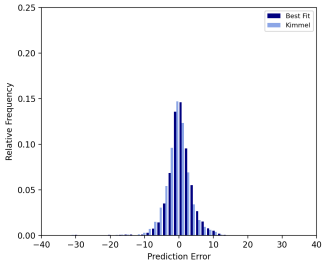
A.



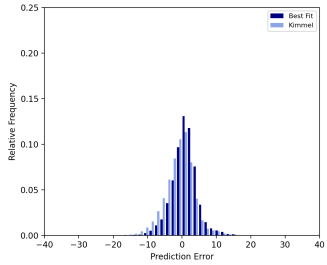
B.



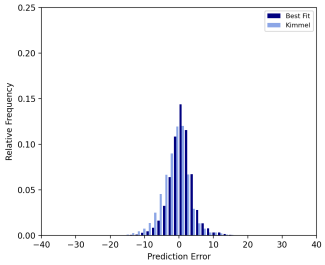
C.



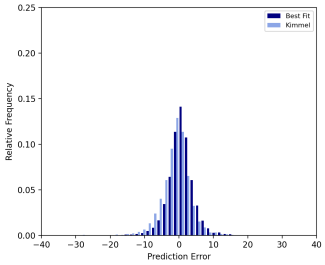
D.



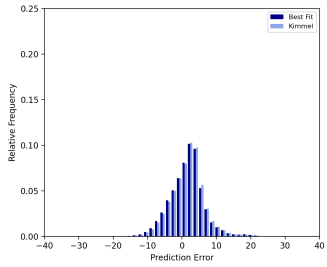
E.



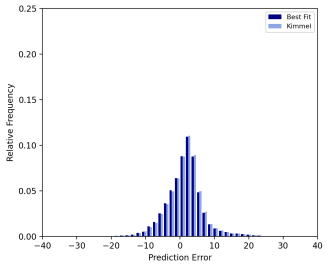
F.



G.



H.



I.

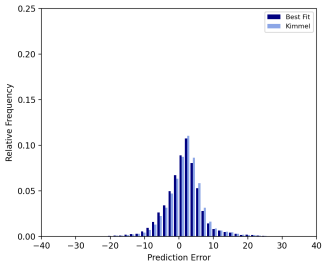


Figure 7