FINAL SCRIPT: APPROVED FOR FILMING

Submission ID #: 68553
Scriptwriter Name: Sulakshana Karkala
Project Page Link: https://review.jove.com/account/file-uploader?src=20908713

Title: Constructing and Visualizing Models using Mime-Based Machine-Learning Framework

Authors and Affiliations:
Hongwei Liu[1–4*], Wei Zhang[1–4*], Yihao Zhang[1–4*], Xuejun Li[1–4], Siyi Wanggou[1–4]

[1]Department of Neurosurgery, Xiangya Hospital, Central South University
[2]National Clinical Research Center for Geriatric Disorders, Xiangya Hospital, Central South University
[3]Hunan International Scientific and Technological Cooperation Base of Brain Tumor Research, Xiangya Hospital, Central South University
[4]Furong Laboratory

***These authors contributed equally to this work**

Corresponding Authors:

Xuejun Li	(lxjneuro@csu.edu.cn)
Siyi Wanggou	(siyi.wanggou@gmail.com)

Email Addresses for All Authors:

Hongwei Liu	(l_magnificence@126.com)
Wei Zhang	(2204170416@csu.edu.cn)
Yihao Zhang	(2204170403@csu.edu.cn)
Xuejun Li	(lxjneuro@csu.edu.cn)
Siyi Wanggou	(siyi.wanggou@gmail.com)

Author Questionnaire

1. We have marked your project as author-provided footage, meaning you film the video yourself and provide JoVE with the footage to edit. JoVE will not send the videographer. Please confirm that this is correct.

√ Correct

2. Microscopy: Does your protocol require the use of a dissecting or stereomicroscope for performing a complex dissection, microinjection technique, or something similar? **No**

3. Software: Does the part of your protocol being filmed include step-by-step descriptions of software usage? **Yes, all done**

4. Proposed filming date: To help JoVE process and publish your video in a timely manner, please indicate the proposed date that your group will film here: **07/07/2025**

When you are ready to submit your video files, please contact our China Location Producer, Yuan Yue.

Current Protocol Length

Number of Steps:	17
Number of Shots:	20

Introduction

1.1. **Siyi Wanggou:** High-throughput sequencing technologies significantly impacts our understanding of biology and cancer heterogeneity. However, with numerous high-throughput sequencing data, it is difficult to rapidly screen and identify disease associated genes or biomarkers.

1.1.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera. *Suggested B.roll:4.1.1*

What technologies are currently used to advance research in your field?

1.2. **Siyi Wanggou:** Numerous machine learning frameworks exist, yet none offer integrated comparison for informed decision-making. To address this gap, we developed Mime—a unified platform for evaluating model strengths and weaknesses.

1.2.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera.

What significant findings have you established in your field?

1.3. **Hongwei Liu:** Mime offers four functions: optimal prognosis modeling, binary response prediction, core prognostic feature identification, and model performance visualization—leveraging 7–10 machine learning algorithms for integrated, interpretable analysis.

1.3.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera.

How will your findings advance research in your field?

1.4. **Wei Zhang:** Researchers often struggled with choosing predictive algorithms and managing ML environments. Mime, an open-source R package, simplifies model setup, parameter selection, and deployment, enabling users to analyze their own data easily.

1.4.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera.

What research questions will your laboratory focus on in the future?

1.5. **Xuejun Li:** Mime marks a milestone in applying AI to biomedicine, aiming to integrate machine learning across single-cell sequencing layers to uncover intra-tumoral heterogeneity within inter-tumoral diversity.

1.5.1. INTERVIEW: Named talent says the statement above in an interview-style shot, looking slightly off-camera.

Protocol

2. Installing and Preparing Transcriptional Cohorts for Prognostic Modeling Using the Mime R Package

Demonstrator: Hongwei Liu

2.1. To begin, open the GitHub *(Gitt-Hub)* website on a desktop computer **[1].** Install the development version of Mime from GitHub using the devtools *(Devv-tools)* package in R **[2]**.

2.1.1. WIDE: Talent lauching GitHUb on computer.

2.1.2. SCREEN: 68553_screenshot_1.mp4 00:00-00:17

2.2. Prepare multiple cohorts containing transcriptional sequencing data with survival or clinical response information. Use the example datasets Example.cohort *(Example-Co-Hort)* and Example.ici *(Example-I-C-I)*, which are accessible from the Mime GitHub repository **[1]**.

2.2.1. SCREEN: 68553_screenshot_2.mp4 00:00-00:13

2.3. The Example.cohort contains two glioma datasets with randomly selected 100 samples from the TCGA *(T-C-G-A)* and CGGA *(C-G-G-A)* database, respectively **[1]**.

2.3.1. SCREEN: 68553_screenshot_3.mp4 00:00-00:05

2.4. Include multiple datasets to construct predictive models for prognosis in Example.cohort. Verify that the dataset format includes the sample ID in the first column, survival time and status in the second and third columns, and log-transformed gene expression levels in the remaining columns **[1]**. Confirm that Dataset1 *(Data-set-One)* is used for training and other datasets for validation **[2]**.

2.4.1. SCREEN: 68553_screenshot_3.mp4 00:10-00:15

2.4.2. SCREEN: 68553_screenshot_3.mp4 00:06-00:09

2.5. Next, load the Example.ici dataset and confirm the format includes sample ID *(I-D)* in the first column, therapeutic response in the second column, and log-transformed gene expression levels in remaining columns **[1-TXT]**. ~~Use the given format for the Example.ici **[2]**, the training dataset for training and the others for validation **[3].**~~
NOTE: Shots 2.5.2-2.5.3 converted into on-screen text

2.5.1. SCREEN: 68553_screenshot_3.mp4 00:16-00:31
TXT: Use given format for example.ici, training dataset for training and others for validation

~~2.5.2. SCREEN: 68553_screenshot_3.mp4 00:19-00:25~~

~~2.5.3. SCREEN: 68553_screenshot_3.mp4 00:26-00:31~~

2.6. Prepare the gene list using the gene set associated with Wnt/β-catenin *(W-N-T-Beta-Cat-A-nin)* signaling in R from the genelist *(gene-list)* file**[1-TXT]**.

2.6.1. SCREEN: 68553_screenshot_3.mp4 00:32-00:35
TXT: Use provided format for genelist

3. Construction of Predictive Models for Prognosis and Response

Demonstrator: Hongwei Liu

3.1. Use the function ML.Dev.Prog.Sig() *(M-L-Devv-Prog-Sigg)* and the given codes to construct predictive models for prognosis based on Example.cohort and the genelist **[1]**.

3.1.1. SCREEN: 68553_screenshot_3.mp4 00:36-00:43, 00:52-00:55,01:30-01:34, 04:50-04:52,07:20-07:30
AND
TEXT ON PLAIN BACKGROUND:

```
library(Mime1)
load("./Example.cohort.Rdata")
load("./genelist.Rdata")
res <- ML.Dev.Prog.Sig(train_data = list_train_vali_Data$Dataset1,
list_train_vali_Data = list_train_vali_Data,
unicox.filter.for.candi = T,
  unicox_p_cutoff = 0.05,
  candidate_genes = genelist,
mode = 'all',nodesize =5,seed = 5201314 )
```

Video Editor: Please play both shots side by side

3.2. Then use the function cindex_dis_all() *(C-index-this-All)* to plot the C-index *(C-Index)* of each model and identify the optimal model **[1]**.

3.2.1. SCREEN: 68553_screenshot_3.mp4 09:52-09:58

3.3. Calculate the survival curves of patients using the according to risk score using a specific model among different datasets and process that in Mime, using the given codes **[1]**.

3.3.1. SCREEN: 68553_screenshot_3.mp4 10:05-10:12

AND

TEXT ON PLAIN BACKGROUND:

```
for (i in c(1:2)) {
print(survplot[[i]]<-rs_sur(res,
model_name = "StepCox[forward] + plsRcox",
dataset = names(list_train_vali_Data)[i],
median.line = "hv",
cutoff = 0.5,
conf.int = T,
  xlab="Day",pval.coord=c(1000,0.9)))
}
aplot::plot_list(gglist=survplot,ncol=2)
```

Video Editor: Please play both shots side by side

3.4. Calculate time-dependent AUC *(A-U-C)* for the predictive models using the function cal_AUC_ml_res() *(Cal-A-U-C-M-L-Res)* and the given codes **[1]**. Now, plot the time-dependent AUC for each model using the function auc_dis_all() *(A-U-C-This-All)* and the given codes **[2]**.

3.4.1. SCREEN: 68553_screenshot_3.mp4. 10:13-10:21

AND

TEXT ON PLAIN BACKGROUND:

```
all.auc.1y <- cal_AUC_ml_res(res.by.ML.Dev.Prog.Sig =res,
train_data = list_train_vali_Data[["Dataset1"]],
inputmatrix.list =list_train_vali_Data,
mode = 'all',AUC_time = 1,
auc_cal_method="KM")
```

Video Editor: Please play both shots side by side

3.4.2. SCREEN: 68553_screenshot_3.mp4 10:22-10:26

AND

TEXT ON PLAIN BCAKGROUND:

```
auc_dis_all(all.auc.1y,
    dataset = names(list_train_vali_Data),
    validate_set=names(list_train_vali_Data)[-1],
    order= names(list_train_vali_Data),
    width = 0.35,
    year=1)
```

Video Editor: Please play both shots side by side

3.5. Process the time-dependent ROC curve of a specific model among different datasets in Mime using the function roc_vis() *(Rock-Viz)* and the given codes **[1]**.

3.5.1. SCREEN: 68553_screenshot_3.mp4 10:33-10:36

AND

TEXT ON PLAIN BACKGROUND:

```
roc_vis(all.auc.1y,
model_name = "StepCox[forward] + plsRcox",
dataset = names(list_train_vali_Data),
order= names(list_train_vali_Data),
anno_position=c(0.65,0.55),
year=1)
```

Video Editor: Please play both shots side by side

3.6. To construct predictive models for therapeutic response, use the function ML.Dev.Pred.Category.Sig() *(M-L-Devv-Pred-Catergory-Sigg)* based on the Example.ici dataset and the genelist **[1]**.

3.6.1. SCREEN: 68553_screenshot_3.mp4 10:38-10:45, 12:23-12:36.

AND

TEXT ON PLAIN BACKGROUND:

```
load("./Example.ici.Rdata")
load("./genelist.Rdata")
res.ici <- ML.Dev.Pred.Category.Sig(
train_data = list_train_vali_Data$training,
list_train_vali_Data = list_train_vali_Data,
candidate_genes = genelist,
methods = c('nb','svmRadialWeights','rf',
'kknn','adaboost','LogitBoost',
'cancerclass'),
seed = 5201314,
cores_for_parallel = 60
)
```

Video Editor: Please play both shots side by side

3.7. Visualize AUC for each response model using auc_vis_category_all() *(Auck-Viz-Category-All)* **[1]**.

3.7.1. SCREEN: 68553_screenshot_3.mp4 12:38-12:46

AND

TEXT ON PLAIN BACKGROUND:

```
auc_vis_category_all(res.ici,dataset = c("training","validation"),
order= c("training","validation"))
```

Video Editor: Please play both shots side by side

3.8. Then generate the ROC curves for each model using roc_vis_category() *(Rock-Viz-Category)* **[1]**.

3.8.1. SCREEN: 68553_screenshot_3.mp4 12:47-12:56

AND

TEXT ON PLAIN BACKGROUND:

```
plot_list<-list()
methods <-
c('nb','svmRadialWeights','rf','kknn','adaboost','LogitBoost','cancerclass')
for (i in methods) {
plot_list[[i]]<-roc_vis_category(res.ici,model_name = i,
dataset = c("training","validation"),
order= c("training","validation"),
anno_position=c(0.4,0.25))

}
aplot::plot_list(gglist=plot_list,ncol=3)
```

Video Editor: Please play both shots side by side

3.9. For the core feature selection, identify core genes associated with prognosis using ML.Corefeature.Prog.Screen() *(M-L-Core-Feature-Prog-Screen)* based on the Example.cohort and genelist **[1]**.

3.9.1. SCREEN: 13:09-13:21, 18:01-18:20, 44:06-44:11.
Video Editor: Please speed up if needed
AND
TEXT ON PLAIN BACKGROUND:

```
load("./Example.cohort.Rdata")
load("./genelist.Rdata")
res.feature.all <- ML.Corefeature.Prog.Screen(
InputMatrix = list_train_vali_Data$Dataset1,
candidate_genes = genelist,
mode = "all",nodesize =5,seed = 5201314 )
```

Video Editor: Please play both shots side by side

3.10. Plot the rank of genes filtered by different methods using core_feature_rank() *(Core-Feature-Rank)* to highlight frequently identified core genes **[1]**.

3.10.1. SCREEN: 68553_screenshot_3.mp4 44:12-44:25
AND
TEXT ON PLAIN BACKGROUND:

```
core_feature_rank(res.feature.all, top=20)
```

Video Editor: Please play both shots side by side

Results

4. Results

4.1. Among the 117 prognostic models constructed by Mime, the StepCox[forward] + plsRcox *(Step-Cocks-Forward-P-L-S-R-Cocks)* combined model showed the highest concordance index across all cohorts **[1]**. Patients with high-risk scores had significantly worse outcomes in all cohorts **[2].** The 1-year area under the curve predicted by SPCOM *(S-P-Com)* ranked highest among all models, with the highest mean AUC value across cohorts **[3]**.

4.1.1. LAB MEDIA: Figure 1A. *Video editor: Highlight the row labeled "StepCox[forward] + plsRcox"*

4.1.2. LAB MEDIA: Figure 1B. *Video editor: Highlight the red survival curves*

4.1.3. LAB MEDIA: Figure 1C and D. *Video editor: Highlight the row labeled "StepCox[forward] + plsRcox" in 1 C*

4.2. Among the 7 therapeutic response prediction models, the svmRadialWeights *(S-V-M-Radial-Weights)* model achieved the highest performance with an area under the curve of 0.81 in the training dataset **[1]** and 0.68 in the validation dataset **[2]**.

4.2.1. LAB MEDIA: Figure 2A. *Video editor: Highlight blue curve in graph for "AUC predicted by svmRadialWeights"*

4.2.2. LAB MEDIA: Figure 2B (middle top panel). *Video editor: Highlight orange curve in graph for "AUC predicted by svmRadialWeights"*

4.3. Core feature selection identified PSEN2 *(P-S-E-N-Two)*, WNT5B *(W-N-T-Five-B)*, and SKP2 *(S-K-P-Two)* as the top-ranked genes based on their recurrence across different algorithms **[1]**.

4.3.1. LAB MEDIA: Figure 3B. *Video editor: Please highlight the top three horizontal bars labeled PSEN2, WNT5B, and SKP2*

Pronunciation Guide:
1. glioma
Pronunciation link: https://www.merriam-webster.com/dictionary/glioma
IPA: /gliˈoʊmə/
Phonetic Spelling: glee-OH-muh

2. prognostic
Pronunciation link: https://www.merriam-webster.com/dictionary/prognostic
IPA: /prɑgˈnɑstɪk/
Phonetic Spelling: prog-NAH-stik

3. svmRadialWeights
(This is a technical R function name, pronounced as separate components.)
No confirmed link found
IPA: /ˌɛs viː ˌɛm ˈreɪdiəl weɪts/
Phonetic Spelling: ess-vee-em RAY-di-uhl wayts

4. plsRcox
(Another R package/function name.)
No confirmed link found
IPA: /piː ɛl ɛs ˌɑr ˈkoʊks/
Phonetic Spelling: pee-el-ess ar KOX

5. β-catenin
Pronunciation link: https://forvo.com/word/%CE%B2-catenin/
IPA: /ˌbeɪtəˈkætənɪn/
Phonetic Spelling: BAY-tuh KAT-uh-nin

6. concordance
Pronunciation link: https://dictionary.cambridge.org/us/pronunciation/english/concordance
IPA: /kənˈkɔːrdəns/
Phonetic Spelling: kun-KOR-dens

7. AUC
Pronunciation link: https://www.howtopronounce.com/auc
IPA: /eɪ siː juː/
Phonetic Spelling: ay-see-you

8. PSEN2
(Gene: presenilin-2)
Pronunciation link: https://pronounceonline.com/word/presenilin/
IPA: /prɛˈsɛnəlɪn/-tuː/
Phonetic Spelling: pre-SEN-uh-lin too

9. WNT5B

(Gene name)
Pronunciation link: https://www.pronouncekiwi.com/WNT5B%20%28gene%29
IPA: /dʌbəl.ju ɛn ti faɪv biː/
Phonetic Spelling: double-you en-tee five bee

---

10. SKP2
(Gene: S-phase kinase-associated protein 2)
Pronunciation link: https://www.pronouncekiwi.com/SKP2%20%28gene%29
IPA: /ɛs keɪ piː tuː/
Phonetic Spelling: ess-kay-pee two