Bioinformatics Advances, 2024, **00**, vbae145
https://doi.org/10.1093/bioadv/vbae145
Advance Access Publication Date: 7 October 2024
Application Note

OXFORD

Systems biology

DynProfiler: a Python package for comprehensive analysis and interpretation of signaling dynamics leveraged by deep learning techniques

Masato Tsutsui[1,2] **and Mariko Okada** [1,*]

[1]Institute for Protein Research, Osaka University, Suita 565-0871, Japan
[2]Biological/Pharmacological Research Laboratories, JT Central Pharmaceutical Research Institute, Takatsuki 569-1125, Japan

*Corresponding author. Institute for Protein Research, Osaka University, 3-2, Yamadaoka, Suita, Osaka 565-0871, Japan. Email: mokada@protein.osaka-u.ac.jp

Associate Editor: Sofia Forslund

Abstract

Summary: Signaling dynamics encode important features and regulatory mechanisms of biological systems, and recent studies have reported the use of simulated signaling dynamics with mechanistic modeling as biomarkers for human diseases. Given the success of deep learning techniques, it is expected that they can extract informative patterns from simulation results more effectively than traditional approaches involving manual feature selection, which can be used for subsequent analyses, such as patient stratification and survival prediction. Here, we propose DynProfiler, which utilizes the entire signaling dynamics, including intermediate variables, as input and leverages deep learning techniques to extract informative features without requiring any labels. Furthermore, DynProfiler incorporates a modern explainable AI solution to provide quantitative time-dependent importance scores for each dynamics. Using simulated dynamics of patients with breast cancer as an example, we demonstrate DynProfiler's ability to extract high-quality features that can predict mortality risk and identify important dynamics, highlighting upregulated phosphorylated GSK3β as a biomarker for poor prognosis. Overall, this tool can be useful for clinical application, as well as for elucidating biological system dynamics.

Availability and implementation: The DynProfiler Python library is available in GitHub at https://github.com/okadalabipr/DynProfiler.

1 Introduction

Mechanistic modeling using ordinary differential equations (ODE) has been used extensively to uncover network regulation and molecular mechanisms in biological systems, particularly in intracellular signaling systems, which often exhibit nonlinear activation dynamics. Conversely, such dynamics can be considered an intrinsic feature of the signaling system of interest, as different outputs are generated depending on the network structure, initial values of the genes and proteins, and their interaction parameters or rate constants (Kholodenko 2006). Based on this idea, signaling dynamics have been proposed as biomarkers for human diseases, such as cancer, and are being utilized in patient-specific models or *digital twins* (Fey *et al.* 2015). Indeed, we previously reported that patient-specific models of the ErbB receptor signaling network could be developed using the gene expression levels of each patient as initial values of the network and common parameters estimated from protein activities of cultured cell lines, and that classification of the resulting dynamics enabled stratification of triple-negative breast cancer (TNBC) patients more accurately than previously reported classification by transcriptome signatures (Imoto *et al.* 2022).

There is a growing trend toward larger mathematical models, which lead to more high-dimensional outputs, highlighting the need for computational approaches to handle such complex data. Previous studies have attempted to extract various features from these complex outputs, including traditional features related to the shapes of the dynamics, such as maxima and area under the curve (AUC), and novel features like *dynamics fingerprints*, which reflect dominant reaction paths from the simulated parameter sets using efficient algorithms (Ortega *et al.* 2024). Additionally, it has been reported that the physiological state can be captured from signaling dynamics by referencing a previously defined comprehensive flux database, which can be utilized to understand the pathophysiology of patients through, for example, predicting drug responses (Khalilimeybodi *et al.* 2024). Furthermore, these dynamic features are optimally selected and weighted through machine learning, enabling a more accurate prediction of ligand-specific outcomes of cancer cells (Hass *et al.* 2017). More recently, improvements in predictive accuracy can be expected via deep learning approaches that capture all possible dynamic features, rather than relying on manually curated ones, although these methods typically hinge on supervised learning and require labeled data (Jacques *et al.* 2021). However, in the context of biology, such labels are frequently absent or only partially accessible, highlighting the need to develop unsupervised learning methods to capture signaling dynamics.

Self-supervised learning, particularly non-contrastive learning, is an unsupervised, label-free representation learning method that has undergone significant advancements,

Received: May 31, 2024; **Revised:** August 25, 2024; **Editorial Decision:** September 18, 2024; **Accepted:** October 3, 2024

especially in the field of computer vision. This approach trains the model to bring representations of two similar inputs closer together in the latent space. Although these methods can extract more effective features from label-less time-series data, few implementations have explored the use of self-supervised learning against signaling dynamics for biological applications. Therefore, we developed DynProfiler, a software package that utilizes 1-dimensional convolutional neural networks (1DCNN) and a self-supervised learning scheme similar to that of SimSham (Chen and He 2020) for encoding time-dependent multidimensional signaling dynamics.

In this study, we examined the use of DynProfiler for clustering simulated patient-specific dynamics of patients with breast cancer. Furthermore, to tackle the "black-box" nature of deep learning approaches that obscure model interpretation, we implemented an explainable AI approach, specifically DeepLift (Shrikumar *et al.* 2017), in DynProfiler to analyze the impact of each signaling dynamic for a specific cluster.

2 Methods

2.1 Example dataset

We obtained preprocessed gene expression datasets of patients with breast cancer from The Cancer Genome Atlas (TCGA) and constructed patient-specific ODE models of the ErbB receptor signaling network with the expression data of 38 genes, 319 rate equations, 228 species, and 648 parameters, as reported previously (Imoto *et al.* 2022). The signaling dynamics under two ligand conditions—epidermal growth factor (EGF) and heregulin, which activate ErbB1 and ErbB3/4, respectively—were simulated, and 121 timestamps were generated by numerically solving the ODE. After post-processing, including the aggregation of multiple variables involved in the phosphorylation of the same protein, we obtained 448 variables with 121 timestamps derived from 369 patients with breast cancer. In conducting the survival analysis, cases with insufficient follow-up or zero survival time were excluded, resulting in a sample of 368 patients for analysis.

2.2 Encoder architecture

To encode signaling dynamics into the latent space, we constructed 1DCNN encoder, which consisted of six 1DCNN blocks that each contained a 1DCNN layer, a pooling layer, batch normalization, and an activation layer. Owing to the convolutions and pooling layers, the input dynamics were compressed in the time dimension and eventually became reduced to a single point within the latent space. To compare the model architectures, we built long short-term memory (LSTM) networks and gated recurrent unit (GRU) networks, which are commonly used for time-series data. For each recurrent neural network (RNN)-based architecture, a bidirectional layer was stacked three times based on prior validations. A bidirectional structure was employed because we did not need to ensure temporal causality. All implementations are available at https://github.com/okadalabipr/DynProfiler.

2.3 Self-supervised learning

The adopted learning scheme closely followed the methodology described in an earlier study (Chen and He 2020). Briefly, the signaling dynamics of the time series and its augmentation were processed using the same encoder. The encoder was trained to correlate the two latent vectors because they inherently contained the same information regarding the signaling pathway. However, to suppress projection to a fixed point, one of the two encoded dynamics was passed through an additional linear layer, and only one branch of the bifurcated computational graph was updated during optimization. More detailed training procedures are described in the Supplementary data.

3 Results

Unlike traditional methods that require manual feature selection based on signaling dynamics, DynProfiler captures arbitrary time-series signaling dynamics without the need for labels, which is achieved using deep learning techniques like self-supervised learning with a 1DCNN encoder (Fig. 1). In the subsequent sections, we describe the qualitative and quantitative evaluations of the resulting latent vectors (Section 3.1) and the interpretation of deep learning predictions to determine the important dynamics for a specific task using the obtained vectors (Section 3.2). The signaling dynamics used as an example were the simulated patient-specific dynamics of 369 patients with breast cancer, which were retrieved from TCGA as reported previously (Imoto *et al.* 2022).

3.1 Embedding signaling dynamics using self-supervised learning

An overview of self-supervised learning is presented in Fig. 1B. The inputs were the entire signaling dynamics obtained from previous simulations, and the augmented data were randomly sampled based on statistics calculated from multiple simulations (Supplementary data). These were processed using an encoder with learnable parameters and projected into the latent space. The encoder was trained to correlate the latent vectors from two similar dynamics. Training performance was monitored using cosine similarity and achieved significantly high scores (greater than 0.98) after approximately 100 iterations (Supplementary Fig. S1A). In this non-contrastive self-supervised learning approach, preventing *collapse*—when all inputs converge to a single fixed point—was crucial. Thus, we examined 2D visualizations of the projections to verify that this phenomenon did not occur (Fig. 2A, left), and we observed well-clustered results according to clinical subtypes, even though these labels were not explicitly used to train the model (Fig. 2A, right).

Next, we calculated the mortality risk scores for patients using the coefficients of Cox regression analysis (Supplementary data) and utilized these scores to predict the survival states at 3, 5, and 7 years. The results showed an area under the receiver operating characteristic curve (AUROC) score > 0.7 for all time periods (Fig. 2B). Using manually selected features, such as maximum amplitudes of the signaling dynamics (see Supplementary data), resulted in lower prediction scores for 5 and 7 years, whereas the 1DCNN encoder retained higher prediction scores, even at the 7-year point (Fig. 2C). These results indicate that more informative features can be obtained using the deep learning approach compared to manual feature selection. As DynProfiler accepts 1DCNN, as well as LSTM, GRU, and other encoders that are suitable for time-series inputs,

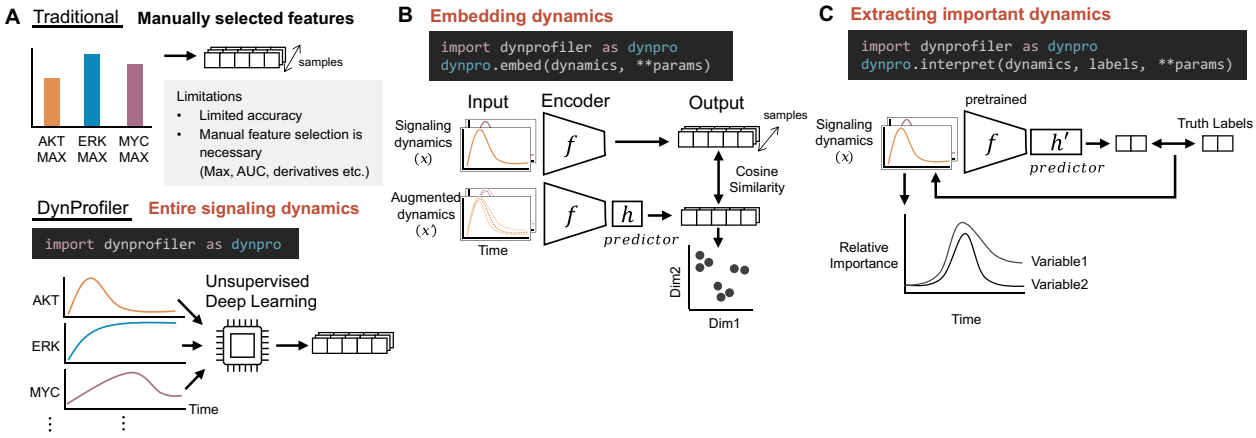

Figure 1. Overview of DynProfiler systems. (A) Conceptual diagrams of traditional signaling dynamics analysis (left) and analysis with DynProfiler (right). (B) Training diagram for embedding dynamics in DynProfiler. Encoder f is a deep learning model such as a 1DCNN; h is an additional layer to make the architecture asymmetric and a linear layer in this case. (C) Conceptual diagram of the DynProfiler calculating time-dependent contributions to the classification task.

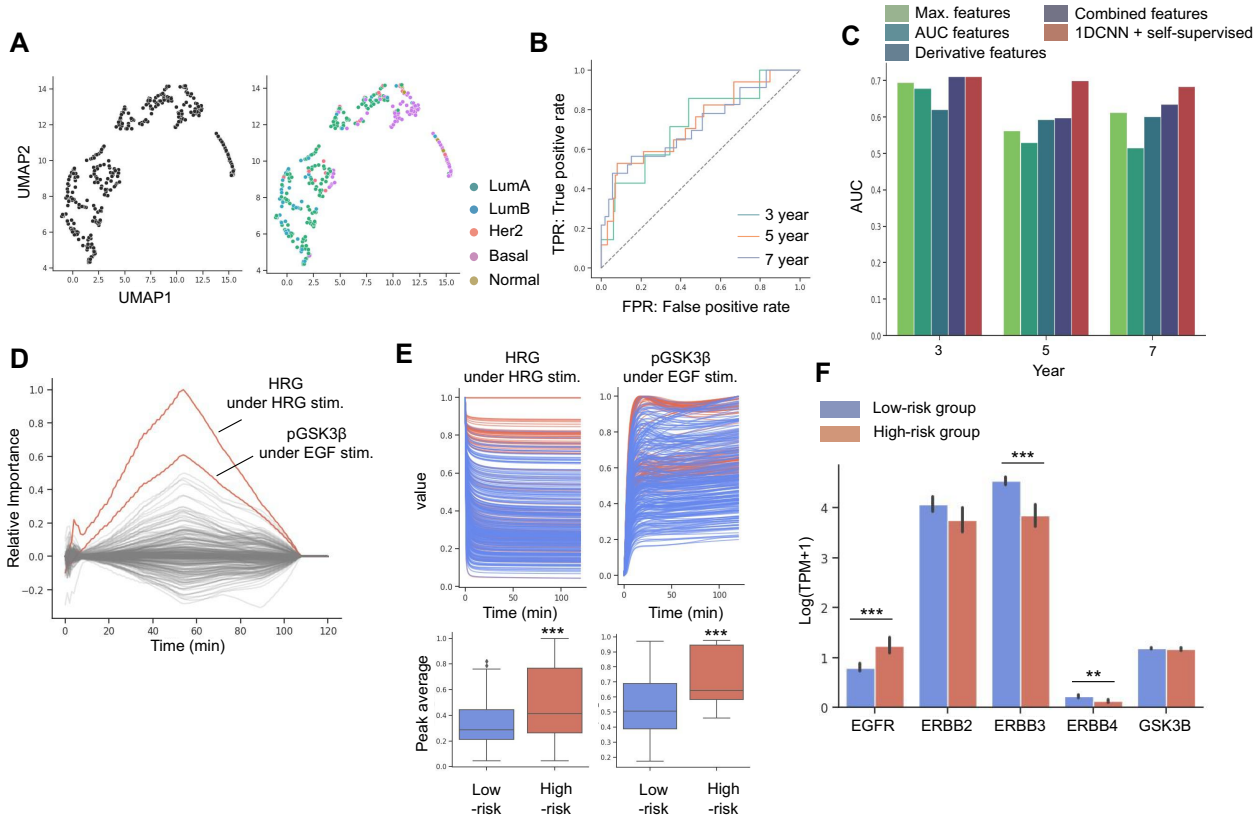

Figure 2. Detection of prognostic features for breast cancer patients by DynProfiler. (A) A 2D visualization of vectors embedded through self-supervised learning of patient-specific mathematical simulations from 369 breast cancer patients (left); the same graph colored according to patient mutations (right). (B) ROC analysis for survival prediction at 3, 5, and 7 years. (C) Comparison based on AUC between manually selected features and DynProfiler features. (D) Time-dependent relative importance of features using DeepLift. (E) Actual simulation results (upper). Blue represents the low-risk group and red represents the high-risk group. The average at time points exceeding a certain importance threshold (bottom). (F) Gene expression levels used for building a patient-specific mathematical model. ***$P < .001$, **$P < .01$ Welch's *t*-test adjusted by Bonfferoni.

we performed a comparison with other encoders and found that LSTM and GRU could acquire relatively effective features for survival prediction, although the 1DCNN still performed the best (Table 1).

To compare the embedding of signaling dynamics using the variational autoencoder (VAE), which is a well-known unsupervised learning method, with the self-supervised learning method, we implemented the VAE for each encoder and conducted the same survival analysis. Particularly for 1DCNN, it was necessary to design the deconvolution layers appropriately for the VAE decoder to match the shape of the data and achieve autoencoding. Comparisons among the encoders regarding VAE showed a similar trend to self-supervised learning, with 1DCNN providing the best survival prediction accuracy, followed by LSTM. However, when comparing self-supervised learning and the VAE using a

Table 1. Comparison of encoders for survival analysis.

Encoder	Training Strategy	C-index	3-year	5-year	7-year
Manual Selection (Max Features)	–	–	0.6755	0.5508	0.6021
1DCNN	Self-supervised	0.6838 (0.0212)	0.7110 (0.0982)	0.7000 (0.0297)	**0.6843** (0.0371)
LSTM	Self-supervised	0.6613 (0.0219)	0.6962 (0.0695)	0.6554 (0.0294)	0.6756 (0.0303)
GRU	Self-supervised	0.6810 (0.0226)	0.7064 (0.0663)	0.6808 (0.0353)	0.6744 (0.0339)
RNN	Self-supervised	0.5855 (0.0508)	0.4849 (0.1468)	0.6447 (0.0806)	0.6127 (0.0664)
1DCNN	VAE	**0.6878** (0.0057)	0.8220 (0.0322)	**0.7150** (0.0146)	0.6710 (0.0156)
LSTM	VAE	0.6828 (0.0135)	**0.9030** (0.0117)	0.6884 (0.0173)	0.6403 (0.0296)
GRU	VAE	0.6479 (0.0412)	0.8338 (0.0458)	0.6568 (0.0420)	0.5902 (0.0417)
RNN	VAE	0.6311 (0.0745)	0.7817 (0.1193)	0.6233 (0.0791)	0.6295 (0.0747)

The scores for deep learning encoders are shown as the average, followed by the standard deviation in parentheses across ten training runs with different random seeds. The highest score for each evaluation metric is highlighted in bold.

1DCNN encoder, there was little difference between them (Table 1). We further discuss the differences between the two methods in the Discussion section.

3.2 Interpreting deep learning predictions to extract important dynamics

The mathematical model of our interest involves interpretable biomolecules, such as the phosphorylation states of proteins. In the previous section, the signaling dynamics generated by the mathematical model were captured and classified using deep learning techniques. However, the application of deep learning often hinders an understanding of the features important for classification owing to its "black box" nature. Therefore, DynProfiler employs a modern technique to analyze the relationship between the inputs and predictions based on the gradients of the parameters of deep neural networks (Shrikumar *et al.* 2017; Fig. 2D). As an example of a specific prediction task, we classified breast cancer patients into high- and low-mortality risk groups, where all signaling dynamics were taken as the input, and a binary class of high or low risk was produced as the output. DynProfiler internally performs this classification task and provides time-dependent importance scores for all biomolecules. The temporal change in free heregulin value after heregulin stimulation was the most important predictor of mortality risk in patients with breast cancer (Fig. 2D and E), indicating that the gene expression levels of ErbB3 and ErbB4, which are heregulin receptors, were significantly lower in the high-risk group (Fig. 2F). This result supports the findings of our previous report, where competitively predominant downstream signaling through EGF receptors against lower ErbB3 and ErbB4 levels leads to a poor prognosis (Imoto *et al.* 2022). Furthermore, enhanced phosphorylated GSK3β dynamics under EGF stimulation was found to be the second most prominent change (Fig. 2D and E). Interestingly, however, the GSK3β gene expression levels did not differ between the two risk groups (Fig. 2F), suggesting that *in silico* signaling dynamics provided more informative features than gene expression for the stratification of cancers.

4 Discussion

Capturing biological dynamics is important for elucidating the molecular mechanisms that control cell fate (Kholodenko 2006). Our previous study showed that the peak maxima of *in silico* kinase activity and steady-state values, as well as the timing of the peaks, are crucial for stratifying patients with colon cancer (Imoto *et al.* 2022). This highlights the importance of capturing all the temporal information generated by biological systems, rather than focusing on a dynamic feature at a single point. In this study, we demonstrated that employing deep learning techniques to comprehensively encode the entire signaling dynamics, including intermediate variables in the signaling network, can enrich information for better prognosis prediction. In addition, our analysis showed that increased phosphorylation of GSK3β may be a potential biomarker for poor prognosis in breast cancer. Papers have shown that the upregulated phosphorylation of GSK3β is associated with colony formation, invasion, and migration in patients with TNBC (Jian *et al.* 2022), supporting our observation. In our analysis, the phosphorylated GSK3β levels were elevated in the high-risk group of patients with TNBC (Supplementary Fig. S3), suggesting that phosphorylated GSK3β can serve as a biomarker even in this population.

As shown in this study, DynProfiler can effectively extract dynamics information without manual feature selection. Encoding dynamics with 1DCNN, LSTM, GRU, and vanilla RNN were found to be relatively effective; however, the 1DCNN encoder performed particularly well in predicting the 7-year survival of patients with breast cancer, possibly because it captures more local patterns. In contrast, LSTM is better at capturing more global features. Considering that signaling dynamics often exhibit nonlinear and sometimes steep time-course patterns, a 1DCNN may be more advantageous as an encoder.

Furthermore, while self-supervised learning, which only requires encoders, has shown good performance, we found that a VAE with properly designed decoders also demonstrates high performance. However, it is worth noting that the VAE requires the learning weights of both the encoder and decoder to be precisely balanced, and in some cases, the decoder may dominate the overall learning process,

potentially reducing embedding quality. We investigated a particular case in which we increased the scale of the data, and this caused a decline in the survival prediction score of the VAE but not of the self-supervised method using a 1DCNN as an encoder (Supplementary data). Therefore, we concluded that self-supervised learning with a 1DCNN encoder is a more generalizable approach for this task.

Nonetheless, considering the comparable performance of the VAE and the success of other recent generative models, such as diffusion models, their application to signaling dynamics is worth exploring. In particular, Normalize Flow and Conditional Flow Matching, which perform continuous distribution transformations over the simulation time course, might be well-suited for time-series data such as signaling dynamics.

Taken together, these results indicate that mechanistically explainable mathematical models can be used to stratify human diseases and understand molecular mechanisms, and that combining them with deep learning techniques will be more effective in extracting important patterns from the dynamics. We argue that this strategy also has the potential for clinical applications, as well as the ability to elucidate biological systems.

Acknowledgements

We thank Mr Kiwamu Arakane and Dr Hiroaki Imoto for their discussions regarding the study.

Author contributions

Masato Tsutsui (Conceptualization [lead], Methodology [lead], Software [lead], Writing—original draft [lead]) and Mariko Okada (Funding acquisition [lead], Supervision [lead], Writing—review & editing [lead])

Supplementary data

Supplementary data are available at *Bioinformatics Advances* online.

Conflict of interest

The authors have no conflicts of financial or non-financial interest to declare relevant to the content of this article.

Funding

This work was supported by the Japan Society for the Promotion of Science KAKENHI [Grant Number 18H04031], the Japan Science and Technology Agency CREST [Program Number JPMJCR21N3], and the Uehara Memorial Foundation for MO.

References

Chen X, He K. Exploring Simple Siamese Representation Learning. https://doi.org/10.48550/arXiv.2011.10566, 2020.

Fey D, Halasz M, Dreidax D *et al.* Signaling pathway models as biomarkers: patient-specific simulations of JNK activity predict the survival of neuroblastoma patients. *Sci Signal* 2015;**8**:ra130.

Hass H, Masson K, Wohlgemuth S *et al.* Predicting ligand-dependent tumors from multi-dimensional signaling features. *NPJ Syst Biol Appl* 2017;**3**:27–15.

Imoto H, Yamashiro S, Okada M. A text-based computational framework for patient-specific modeling for classification of cancers. *iScience* 2022;**25**:103944.

Jacques M-A, Dobrzyński M, Gagliardi PA *et al.* CODEX, a neural network approach to explore signaling dynamics landscapes. *Mol Syst Biol* 2021;**17**:e10026.

Jian Y, Kong L, Xu H *et al.* Protein phosphatase 1 regulatory inhibitor subunit 14C promotes triple-negative breast cancer progression via sustaining inactive glycogen synthase kinase 3 beta. *Clin Transl Med* 2022;**12**:e725.

Khalilimeybodi A, Saucerman JJ, Rangamani P *et al.* Modeling cardiomyocyte signaling and metabolism predicts genotype-to-phenotype mechanisms in hypertrophic cardiomyopathy. *Comput Biol Med* 2024;**175**:108499.

Kholodenko BN. Cell-signalling dynamics in time and space. *Nat Rev Mol Cell Biol* 2006;**7**:165–76.

Ortega OO, Ozen M, Wilson BA *et al.* Signal execution modes emerge in biochemical reaction networks calibrated to experimental data. *iScience* 2024;**27**:109989.

Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *Proceedings of the 34th International Conference on Machine Learning—Volume 70*, ICML'17. JMLR.org, Sydney, NSW, Australia, 2017, pp. 3145–3153.

Bioinformatics Advances, 2024, 00, 1–5
https://doi.org/10.1093/bioadv/vbae145
Application Note