Dimensionality Reduction and Prediction of the Protein Macromolecule Dissolution Profile

Varun Kumar Ojha, Konrad Jackowski, Václav Snášel, and Ajith Abraham

IT4Innovations, VŠB Technical University of Ostrava, Ostrava, Czech Republic
{varun.kumar.ojha,vaclav.snasel}@vsb.cz, konrad.jackowski@pwr.wroc.pl,
ajith.abraham@ieee.org

Abstract. A suitable regression model for predicting the dissolution profile of Poly (lactic-co-glycolic acid) (PLGA) micro- and nanoparticles can play a significant role in pharmaceutical/medical applications. The rate of dissolution of proteins is influenced by several factors and taking all such influencing factors into account, we have a dataset in hand with three hundred input features. Therefore, a primary approach before identifying a regression model is to reduce the dimensionality of the dataset at hand. On the one hand, we have adopted Backward Elimination Feature selection techniques for an exhaustive analysis of the predictability of each combination of features. On the other hand, several linear and non-linear feature extraction methods are used in order to extract a new set of features out of the available dataset. A comprehensive experimental analysis for the selection or extraction of features and identification of corresponding prediction model is offered. The designed experiment and prediction models offers substantially better performance over the earlier proposed prediction models in literature for the said problem.

Keywords: Dimension reduction, Feature selection, Feature extraction, Regression, PLGA.

1 Introduction

Predicting dissolution profile of Poly (lactic-co-glycolic acid) (PLGA) micro and nanoparticles is a complex problem as there are several potential factors influencing dissolution of PLGA protein particles [1]. Collecting all such influencing factors leads to three hundred input features in dataset. Therefore, primary approach one may adopt is the reduction of dimensionality of dataset. Dimensionality reduction techniques transform a high dimension dataset to a low dimension datasets thereby, improving models computational speed, predictability and generalization ability. Dimensionality reduction may be categorised in two paradigms, feature selection and feature extraction. The former is useful when a dataset is available with high dimension and fewer cases (samples), while feature extraction is useful when a dataset has an extremely large dimension and high redundancy. In present problem, we shall explore both feature selection and feature extraction techniques to find out best possible solution. To figure

P. Krömer et al. (eds.), *Proceedings of the Fifth Intern. Conf. on Innov. in Bio-Inspired Comput. and Appl. IBICA 2014,* Advances in Intelligent Systems and Computing 303,
DOI: 10.1007/978-3-319-08156-4_30, © Springer International Publishing Switzerland 2014

out relationship between obtained input variables (features) and output variable, several regression models are employed. We shall analyse prediction models to obtain a suitable prediction model for the said problem.

In present scope of the study, we focus on the PLGA nano- or microspheres dissolution properties and drug release. Szlkeket et al. [2] and Fredenberg et al. [3] described that the drug release from the PLGA matrix is mainly governed by two mechanisms, diffusion and degradation/erosion. Several factors influencing the diffusion and degradation rate of PLGA described by Kang et al. [4, 5], Blanco and Alonso [6] and Mainardes et al. [7] are pore diameters, matrix active pharmaceutical ingredient (API) interactions, API - API interactions, and formulation composition. Szlkeket et al. [2] offered predictive model to describe underlying relationship of those influencing factors on the drug release profile, where they focus on the feature selection, artificial neural network and genetic programming to obtain a suitable predicting model for the said purpose. In past, several mathematical models including Monte Carlo and cellular automata microscopic models were proposed by Zygourakis and Markenscoff [8] and Gpferich [9]. A partial differential equations model was proposed by Siepmann et al [10] to address the influence of underlying PLGA properties on the drug release rate or protein dissolution.

We shall discuss the PLGA drug release problem and dataset collection mechanisms in section 2.1. In section 2.2, we shall discus the computational tools available for dimensionality reduction and prediction. A comprehensive discussion on the experimental setup is offered in section 3. Finally, we shall conclude our discussion in section 4.

2 Methodology

2.1 Problem Description

Poly (lactic-co-glycolic acid) (PLGA) micro- and nanoparticles could play significant role in medical application and toxicity evaluation of PLGA-based multiparticulate dosage form. Poly (lactic-co-glycolic acid) (PLGA) microparticles are important diluents in the formulation of drugs in the dosage form. Apart from playing a role as a filler, PLGA as an excipient, and alongside pharmaceutical active ingredients (APIs) plays crucial role in various ways. It helps dissolution of the drugs, thus increases absorbability and solubility of drugs. It helps in pharmaceutical manufacturing process by improving APIs powder's flowability and nonstickiness. Nonetheless, it helps in vitro stability such as prevention of denaturation over expected shelf life.

Present study is performed on the dataset offered by Szlkeket et al. [2] in their article "Heuristic modeling of macromolecule release from PLGA microspheres". Dataset collected from various literature by Szlkeket et al. [2] has three hundred input features divided into four groups, namely protein descriptor, plasticizer, formulation characteristics, and emulsifier. Formulation characteristics group contains features such as PLGA inherent viscosity, PLGA molecular weight, lactide-to-glycolide ratio, inner and outer phase Polyvinyl alcohol (PVA) concentration, PVA molecular weight, inner phase volume, encapsulation rate, mean particle size, and PLGA concentration and experimental condition (dissolution pH, number of dissolution additives, dissolution additive concentration and production method, and dissolution time). Feature groups protein descriptor, plasticizer and emulsifier contains 85, 98 and 101 features respectively. The regression model sought to predict dissolution percentage or solubility of PLGA which depends on the features mentioned above. In order to avoid overfitting, collected data are preprocessed by adding noise to them. Dataset is then normalized in the range [-1.0, 1.0].

2.2 Dimensionality Reduction Tools

Feature Selection (Backward Elimination). Feature selection techniques enable us to choose from the set of input features we have in our hand. Especially, feature selection become significant step towards development of a predication model where it requires expensive (both in time and cost) experimental examination. Backward Feature Elimination Filter provided in open-source platform KNIME[1] is used for feature elimination. The basic methodology and principle behind backward elimination filter is to start from the maximum number feature in hand (in this case it starts with three hundred features) and to search all possible combinations of the features in order to eliminate (marked) poorest feature in terms of its predictability in the set of all features. Moreover, the feature with the worst performance in terms of error as obtained by the regression model used is eliminated. In subsequent iteration the operation repeated for the remaining features and so on.

Feature Extraction. When it is affordable to generate test features easily, feature extraction technique may be useful to employ for dimensionality reduction. A regression model with reduced input dimension may performs as good as it can with a complete set of features. [11]. Therefore, feature extraction for dimensionality helps is reducing computational overhead which may incurred due to use of complete input dimension.

Principle Component Analysis (PCA): PCA is linear dimensionality reduction technique which transforms correlated data into uncorrelated data in the reduced dimension by the means of finding a linear basis of reduced dimensionality for the data, in which the amount of variance in the data is maximal.

Factor Analysis: Linear dimension reduction technique, Factor Analysis, as opposed to PCA, finds whether a number of features of interest are linearly related to small/reduced number of newly defined features called factors. In other words, it discovers reduced number of relatively independent features through the means of mapping correlated features to small set of features known as factors.

Independent Component Analysis (ICA): Similar to FA, ICA proposed by Hyvarinen et al. [12, 13] is a linear dimension reduction technique that transforms

[1] KNIME - Professional Open-Source Software of KNIME.com AG.

multidimensional feature vector into components that are statistically as independent as possible.

Kernel PCA (kPCA): Kernel PCA, a non-linear technique of dimension reduction, is an extension of PCA using kernel methods. Kernel PCA computes the principal eigenvectors of the kernel matrix, rather than those of the covariance matrix. The reformulation of PCA in kernel space is straightforward, since a kernel matrix is similar to the inner product of the data points in the high dimensional space that is constructed using the kernel function. Typically, Gaussian, Tangent hyperbolic, Polynomial, etc. functions are used as kernel.

Multidimensional Scaling (MDS): MDS is a non-linear dimension reduction technique, maps the high dimensional data representation to a low-dimensional representation while retaining the pairwise distances between the data points as much as possible.

2.3 Prediction Models

Regression/Prediction model tries to figure out the relationship between independent variable (input variables X) and dependent variables (output variable y). Moreover, it tries to find unknown parameters (β) such that error (2) is minimized given that dependent variable y, independent variable X and predicted output $\widehat{y}$

$$y = f(X, \beta) \tag{1}$$

Let $e_i = (\widehat{y_i} - y_i)$ be the difference between the values of the true value of the dependent variable y_i and predicted value $\widehat{y_i}$. Therefore, sum of square error ξ over data samples of size n.

$$\xi = \sum_{i=1}^{n} e_i^2 \tag{2}$$

Linear Regression (LReg). Linear regression is the simplest predictive model where p independent variables ($|X| = n \times p$), dependent variable y_i with noise ε_i may be written as (3).

$$y_i = \beta_1 x_i 1 + \beta_2 x_i 2 + \ldots + \beta_p x_i p + \varepsilon_i = \mathbf{x}_i^T . \beta + \varepsilon_i \tag{3}$$

where ε_i is called noise or error variable.

Gaussian Process Regression (GPreg). Rasmussen [14, 15]. A Gaussian process is fully specified by its mean function $m(x)$ and covariance function $k(x, x')$. This is a natural generalization of the Gaussian distribution whose mean m and covariance k is a vector and matrix respectively. The Gaussian distribution is over vectors, whereas the Gaussian process is over functions f. We may write.

$$f \sim \mathcal{GP}(m, k) \tag{4}$$

Multilayer Perceptron (MLP). Multilayer perceptron (MLP) is a feedforward neural network having one or more hidden layers in between input and output layers [16, 17]. A neuron in an MLP first computes linear weighted combination of real valued inputs and then limits its amplitude using an non-linear activation function. In present case, MLP is trained using Backpropagation algorithm propounded by Rumelhart et a. [18] and Resilient propagator (RProp) developed by Riedmiller et al. [19].

Sequential Minimal Optimization Regression (SMOReg). Sequential minimal optimization (SMO), an algorithm for the training of Support Vector Regression (SVR), proposed by Smola and Schlkopf [20, 21, 22] was an extension of the SMO algorithm proposed by Platt [23] for SVM classifier. SVR attempts to minimize the generalization error bound so as to achieve generalized performance. The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a non-linear function.

3 Experimental Setup and Results

Experimental setup for the dimensionality reduction and identification of corresponding regression model for the prediction of protein molecules is as follows. The experiment is conducted using MATLAB[2], KNIME and WEKA[3]. As mentioned in section 2, dataset obtained for the PLGA dissolution profile has three hundred features, therefore the primary objective is to reduce the dimension of the dataset. Feature selection and feature extraction discussed in section 2.2 are used for the dimensionality reduction. Subsequent to dimension reduction, predication models are employed and assessed using 10 cross-validation (10cv) sets prepared after dimension reduction. Selection of prediction model is based on the average and variance computed over a set of 10 Root Mean Square Errors (RMSE) obtained as result of 10cv experiment. A pictorial illustration of the experiment is shown in figure 1.

3.1 Experimental Results of Feature Selection Technique

After cleaning and preprocessing dataset, it goes under backward elimination treatment, where we have a set of prediction models such as GP regression with RBF kernel, LReg, three-layer MLP with fifty neurons at the hidden layer, learning rate 0.3, momentum rate 0.2 and SMOReg with polynomial kernel, epsilon value 0.001 and tolerance label 0.001. As a result of backward elimination process, each of the regression model ends with a list containing all combinations of the features starting from a single selected feature to two hundred ninety-nine

[2] MATLAB is trademark of MathWorks, Inc.

[3] WEKA - Data Mining Software in Java developed by machine learning group at the University of Waikato, Free Software Foundation, Inc.

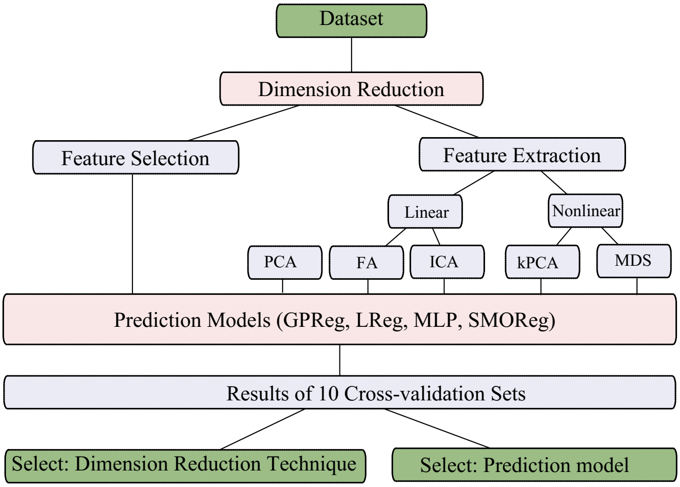

Dataset
Dimension Reduction
Feature Selection
Feature Extraction
Linear
Nonlinear
PCA
FA
ICA
kPCA
MDS
Prediction Models (GPReg, LReg, MLP, SMOReg)
Results of 10 Cross-validation Sets
Select: Dimension Reduction Technique
Select: Prediction model

Fig. 1. Experimental setup for the identification of the dimensionality reduction and selection prediction model for the prediction of PLGA dissolution profile

features and their corresponding mean sum square error. Therefore, the combination which offer least value of mean square error is termed as an optimal set of features for the corresponding regression model. For example, the optimal set of features obtained for regression models GPReg, LReg, MLP and SMOReg are 18, 32, 31 and 30 with mean square error (result of normalized dataset) 0.143, 0.156, 0.121, and 0.153 respectively. From 10cv experimental result presented in Table 1 and Figure 2, it is evident that considering the entire features SMOReg performs better in terms of Mean of RMSEs and Variance followed by GPReg and MLP. Whereas, in case of optimal features selection, GPReg performs better than the rest of the regression models, it also performs better than the performance of SMOReg model which performs best while considering all features. MLP is only next to GPReg in terms of RMSE when it comes to section of 10 features or optimal features. Examining Figure 2, it is evident that GPReg performs better in terms of both average RMSE and variance (VAR). Whereas, performance of SMOreg is only next to GPReg in terms of average RMSE. On the other hand MLP performs slightly poorer than SMOReg and LReg in terms of average RMSE. We may therefore conclude that GPReg offers best solution to the current problem. GPReg offers 17 and 10 selected features. However, the difference between average RMSE is insignificant. Therefore, we may conclude that the optimal reduced feature for the present problem may be considered as 10. The result of backward elimination filter 10 features are as follows. From the protein descriptor group, we have Aliphatic ring count, van der Waals volume and quaternary structure, from the formulation characteristics group, we have PLGA viscosity, PVA concentration inner phase, Mean particle size, and PLGA to Placticizer, from the Plasticizer group, we have pH7-msdon and from the Emulsifier group, we have Wiener index and dissolution time in days.

Table 1. Experimental results for 10cv datasets prepared with distinct random partitions of the complete dataset using feature selection technique (Identification of regression model) *Note. Mean and variance (VAR) is computed on 10 RMSE obtained.*

Regression Model	Reduced Number of Features									
	1		5		10		Optimal		300	
	Mean	VAR	Mean	VAR	Mean	VAR	Mean	VAR	Mean	VAR
GPReg	27.474	10.942	17.107	3.989	**15.322**	**3.782**	15.709	3.162	16.812	3.551
LReg	26.613	3.232	23.447	3.702	19.979	3.402	17.847	1.634	17.074	2.738
MLP	28.329	7.428	23.113	10.007	20.997	11.365	17.820	8.095	18.571	21.063
SMOReg	26.970	3.307	23.381	2.729	19.526	3.757	17.885	3.321	16.529	2.554

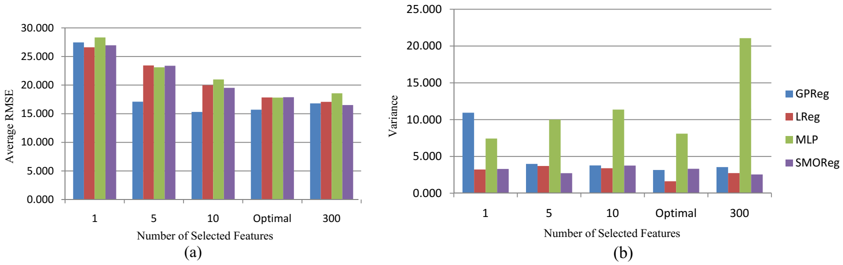


Fig. 2. Experimental results of feature selection, comparison between the regression models. (a) comparison using average RMSE (b) comparison using variance.

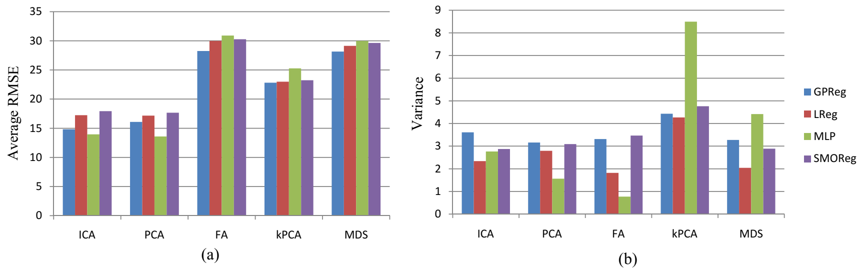
Nevertheless, it is worth mentioning that the best result presented by Szlkeket al. [2] is root mean square error (RMSE) of 15.4 considering 11 selected features using MLP and 17 features with RMSE of 14.3 using MLP. The process of the presented feature selection was able to find the most significant features influencing drug release rate. It may be observed that features vectors from the all four mentioned feature groups are among the selected features. Therefore, a general theory may be drawn about how features dominate PLGA drug release rate.

3.2 Experimental Results of Feature Extraction Technique

Unlike feature selection, feature extraction finds new set of reduced feature by computing linear or non-linear combinations of features from the available dataset. As described in section 2.2, various feature extraction techniques may be used for the said purpose. A comprehensive result is presented in Table 2 illustrating performance of feature extraction methods and regression models. Dimensionality reduction tools offered by van der Maaten et al. [11] are used for the feature extraction. Linear dimensionality reduction methods, PCA and FA and non-linear dimensionality reduction methods such as kPCA and MDS are used to reduce dimension of dataset from 300 to 50 , 30 , 20, 10 and 5. Whereas, ICA is used to reduced dimension of dataset from 300 to 50. Results obtained using ICA are as follows. Mean RMSE and variance corresponding to GPReg,

Table 2. Experimental results for 10cv datasets prepared with distinct random partitions of the complete dataset using feature extraction techniques *Note. Mean and variance (VAR) is computed on 10 RMSE obtained.*

Feature Extraction Method		Regression Model	Reduced Dimension									
			1		5		10		20		30	
			Mean	VAR	Mean	VAR	Mean	VAR	Mean	VAR	Mean	VAR
Linear Method	PCA	GPReg	28.88	1.62	27.22	3.00	24.80	3.85	19.82	2.49	16.08	3.16
		LReg	29.55	1.74	29.22	1.70	27.73	2.21	23.93	1.63	17.17	2.79
		MLP	30.36	3.36	29.77	6.37	26.58	3.98	19.89	2.27	**13.59**	**1.56**
		SMOReg	30.14	3.17	29.78	3.62	27.95	2.67	24.31	1.89	17.66	3.09
	FA	GPReg	29.23	1.77	28.56	2.67	28.31	3.34	28.30	3.42	28.26	3.31
		LReg	29.97	1.77	29.97	1.77	29.97	1.77	29.97	1.77	29.98	1.82
		MLP	30.64	2.02	30.50	1.91	31.01	1.83	30.93	2.30	30.91	0.77
		SMOReg	30.28	3.45	30.28	3.45	30.26	3.37	30.29	3.44	30.28	3.46
Non-linear Method	Kernel PCA	GPReg	28.60	1.68	27.08	2.12	24.96	1.96	24.32	2.17	22.81	4.43
		LReg	29.31	1.52	28.05	1.78	25.35	2.05	25.17	2.23	22.98	4.27
		MLP	29.81	3.57	29.65	7.94	27.07	4.09	25.97	5.52	25.27	8.49
		SMOReg	29.43	1.41	28.68	1.65	25.90	1.70	25.79	2.00	23.24	4.76
	MDS	GPReg	28.91	2.17	28.73	2.47	28.41	3.16	28.24	3.17	28.16	3.27
		LReg	29.56	1.86	29.21	2.08	29.19	2.08	29.11	1.92	29.14	2.04
		MLP	30.42	3.71	29.38	4.11	29.93	3.10	30.01	4.53	29.98	4.42
		SMOReg	29.98	2.62	29.64	2.55	29.64	2.76	29.66	2.85	29.65	2.89



(a)

(b)

Fig. 3. Experimental results of feature extraction with reduced dimension 30, comparison between the regression models. (a) comparison using average RMSE (b) comparison using variance.

LReg, MLP and SMOReg are 14.83, 17.23, **13.94**, and 17.92 and 3.61, 2.34, **2.77**, and 2.87 respectively. It may be observed from Table 2 that lower dimensions offers less significance improvement to results in terms of RMSE. However, if we compare the best result (result of reduced dimension to 50) of PCA (RMSE **13.59** corresponding MLP) and ICA (RMSE **13.94** corresponding to MLP) with the result with all features (RMSE **16.812** corresponding to GPReg), it is evident that reduction in dimension significantly improves the performance of the prediction model. Examining Figure 3, a RMSE and variance (VAR) comparison between chosen regression model applied on dataset reduced to dimension 50 by feature extraction techniques ICA, PCA, FA, kPCA and MDS, we may conclude that feature extraction using PCA performs best, both in terms of RMSE and VAR when regression model MLP is used, whereas, feature extraction using ICA performers only next to PCA when MLP is used, when it comes to GPReg, ICA has an edge over PCA result.

4 Conclusion

The challenge of predicting a protein molecules dissolution profile is due to the large number of input features available where each of the input features may potentially be an influencing factor affecting dissolution of proteins. Therefore, predicting the rate of dissolution is a complex problem. Hence, on the one hand we have adopted feature selection technique, which lets us select most influencing features among the available features without worsen performance. On the other hand we have features extraction techniques which let us consider the entire available feature, but provide a reduced set of new features which performs better than when considering all the features together. In order to identify regression models, we have analysed the performance of GPReg, LReg, MLP and SMOReg. As a result of comprehensive evaluation of the aforementioned experiments, we may conclude that GPReg performs best when it comes to feature selection where it select 10 features and offer lowest average RMSE and VAR. We may observe from the experiment of feature extraction that PCA used to reduce dimension to 50 offered best result using MLP with lowest average RMSE and VAR. From the aforementioned experiment and results, a general model for understanding PLGA drug release rate may be obtained for various medical and pharmaceutical applications.

Acknowledgement. This work was supported by the IPROCOM Marie Curie initial training network, funded through the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement No. 316555.

This paper has been elaborated in the framework of the project Opportunity for young researchers, reg. no. CZ.1.07/2.3.00/30.0016, supported by Operational Programme Education for Competitiveness and co-financed by the European Social Fund and the state budget of the Czech Republic.

References

[1] Astete, C.E., Sabliov, C.M.: Synthesis and characterization of plga nanoparticles. Journal of Biomaterials Science, Polymer Edition 17(3), 247–289 (2006)

[2] Szlkek, J., Paclawski, A., Lau, R., Jachowicz, R., Mendyk, A.: Heuristic modeling of macromolecule release from plga microspheres. International Journal of Nanomedicine 8, 4601 (2013)

[3] Fredenberg, S., Wahlgren, M., Reslow, M., Axelsson, A.: The mechanisms of drug release in poly (lactic-co-glycolic acid)-based drug delivery systems–a review. International Journal of Pharmaceutics 415(1), 34–52 (2011)

[4] Kang, J., Schwendeman, S.P.: Pore closing and opening in biodegradable polymers and their effect on the controlled release of proteins. Molecular Pharmaceutics 4(1), 104–118 (2007)

[5] Kang, J., Lambert, O., Ausborn, M., Schwendeman, S.P.: Stability of proteins encapsulated in injectable and biodegradable poly (lactide-co-glycolide)-glucose millicylinders. International Journal of Pharmaceutics 357(1), 235–243 (2008)

[6] Blanco, M., Alonso, M.: Development and characterization of protein-loaded poly (lactide-co-glycolide) nanospheres. European Journal of Pharmaceutics and Biopharmaceutics 43(3), 287–294 (1997)

[7] Mainardes, R.M., Evangelista, R.C.: Plga nanoparticles containing praziquantel: effect of formulation variables on size distribution. International Journal of Pharmaceutics 290(1), 137–144 (2005)

[8] Zygourakis, K., Markenscoff, P.A.: Computer-aided design of bioerodible devices with optimal release characteristics: a cellular automata approach. Biomaterials 17(2), 125–135 (1996)

[9] Gopferich, A.: Mechanisms of polymer degradation and erosion. Biomaterials 17(2), 103–114 (1996)

[10] Siepmann, J., Faisant, N., Benoit, J.P.: A new mathematical model quantifying drug release from bioerodible microparticles using monte carlo simulations. Pharmaceutical Research 19(12), 1885–1893 (2002)

[11] van der Maaten, L.J., Postma, E.O., van den Herik, H.J.: Dimensionality reduction: A comparative review. Journal of Machine Learning Research 10(1-41), 66–71 (2009)

[12] Hyvarinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks 13(4), 411–430 (2000)

[13] Hyvarinen, A.: Fast and robust fixed-point algorithms for independent component analysis. IEEE Transactions on Neural Networks 10(3), 626–634 (1999)

[14] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press (2005)

[15] Rasmussen, C.E., Nickisch, H.: Gaussian processes for machine learning (gpml) toolbox. The Journal of Machine Learning Research 9999, 3011–3015 (2010)

[16] Haykin, S.: Neural Networks: A Comprehensive Foundation, 1st edn. Prentice Hall PTR, Upper Saddle River (1994)

[17] Werbos, P.J.: Beyond regression: New tools for prediction and analysis in the behavioral sciences (1975)

[18] Rumelhart, D.E., McClelland, J.L.: Parallel distributed processing: explorations in the microstructure of cognition. foundations, vol. 1 (1986)

[19] Riedmiller, M., Braun, H.: A direct adaptive method for faster backpropagation learning: The rprop algorithm. In: IEEE International Conference on Neural Networks, 1993, pp. 586–591. IEEE (1993)

[20] Smola, A.J., Scholkopf, B.: Learning with kernels. Citeseer (1998)

[21] Smola, A.J., Schollkopf, B.: A tutorial on support vector regression. Statistics and Computing 14(3), 199–222 (2004)

[22] Scholkopf, B., Burges, C.J., Smola, A.J.: Advances in kernel methods: support vector learning. MIT Press (1999)

[23] Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in Large Margin Classifiers 10(3), 61–74 (1999)