Supplementary Materials

Develop machine learning based regression predictive models for engineering protein solubility

Prediction of protein solubility

1 Database

Table S1. Summary of 26 proteins excluded

Name of gene	Reasons for removing
yedM	There are multiply stop codons in the middle of the sequence.
rffT	Two hits were found for the gene name in NCBI.
spr	Two hits were found for the gene name in NCBI.
ade	Six hits were found for the gene name in NCBI.
tiaE	This gene was not from MG1655.
yiaI	This gene was not from MG1655.
gapC	There are multiply stop codons in the middle of the sequence.
ilvG	There are multiply stop codons in the middle of the sequence.
ilvG	There are multiply stop codons in the middle of the sequence.
insO	Two hits were found for the gene name in NCBI.
phnE	There are multiply stop codons in the middle of the sequence.
phnE	There are multiply stop codons in the middle of the sequence.
yagP	Two hits were found for the gene name in NCBI.
ybeM	There are multiply stop codons in the middle of the sequence.
ycgH	There are multiply stop codons in the middle of the sequence.
ydaY	There are multiply stop codons in the middle of the sequence.
yjiQ	There are multiply stop codons in the middle of the sequence.
yjgX	There are two proteins with same name and different solubility in the database.
yjgX	There are two proteins with same name and different solubility in the database.
gatR	There are two proteins with same name and different solubility in the database.
gatR	There are two proteins with same name and different solubility in the database.
ybfH	Amino acid X representing undetermined amino acid) appears in the sequence.
yhdW	Amino acid X representing undetermined amino acid) appears in the sequence.
ymgH	Amino acid X representing undetermined amino acid) appears in the sequence.
yohG	Amino acid X representing undetermined amino acid) appears in the sequence.
kdpF	The sequence is too short to be converted into numerical values by descriptors.

2 SVM tuning

Table S2. Performance of SVM for different kernels

Kernel	Accuracy	R^2
linear	0.6684	0.2240
sigmoid	0.4528	0.0011
radial basis	0.7500	0.4064
polynomial	0.6339	0.0732

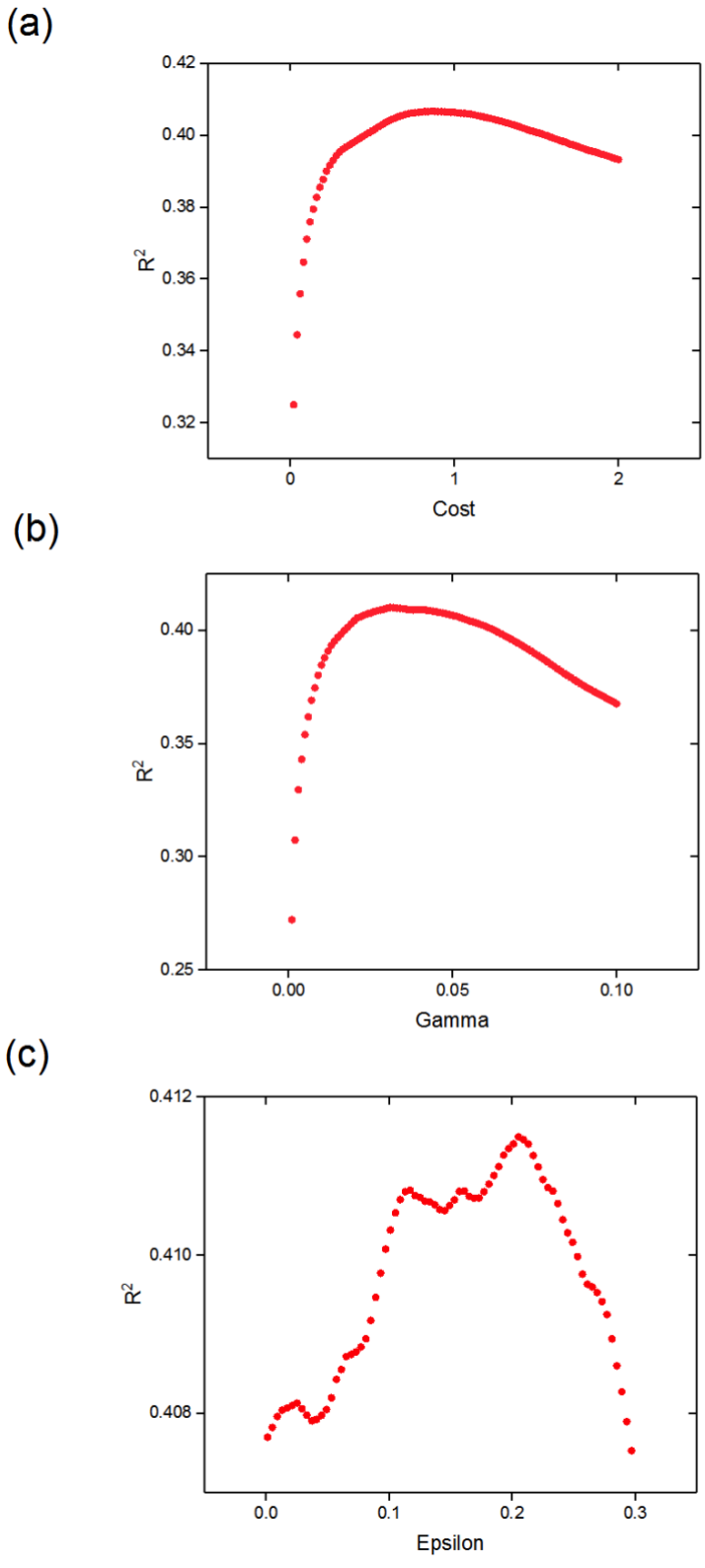

Fig. S1. Plot between R^2 of SVM and cost, gamma and epsilon respectively.

3 Data augmentation

Table S3. Performance of SVM based on data generated from different data augmentation algorithms

GAN version	R^2
No GAN	0.4092
GANs	0.4015
CGAN	0.4064
WGAN	0.3787
WCGAN	0.4003

Table S4. Performance of SVM based on GANs for 5000 iterations

GANs version	R^2
No GANs-1	0.4093
GANs-1	0.4044
No GANs-2	0.4200
GANs-2	0.4240
No GANs-3	0.4447
GANs-3	0.4462

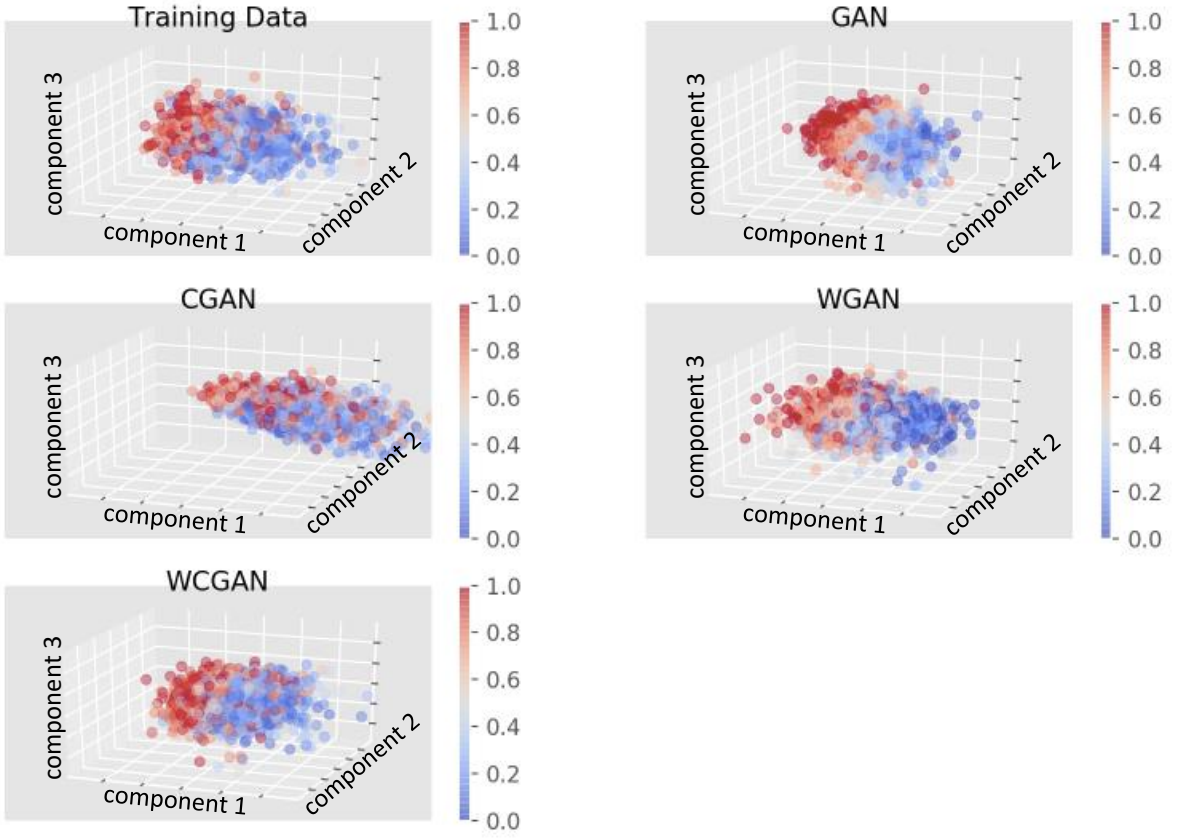

Fig. S2. Original data and generated data by different versions of GANs for 500 iterations.

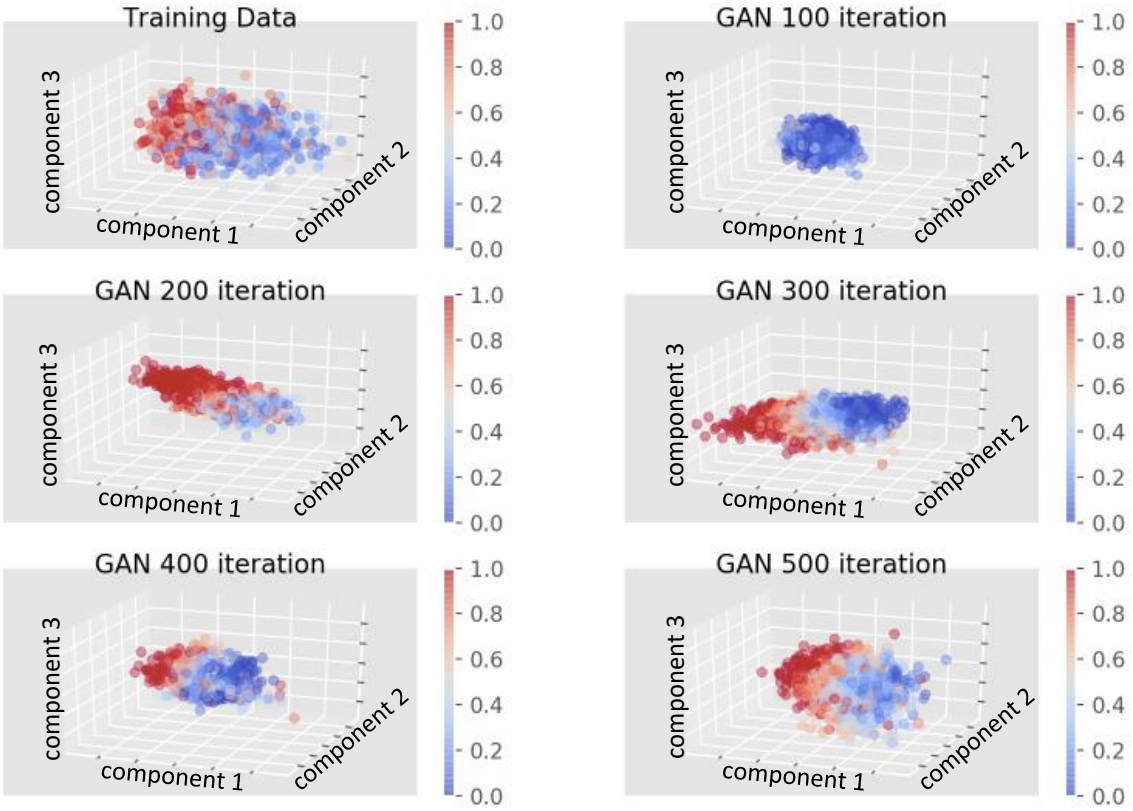

Fig. S3. Original data and generated data by GANs model in different steps for 500 iterations