

Contents

1	Introduction	1
1.1	What is Machine Learning?	1
1.2	Basic Machine Learning	2
1.3	Why Machine Learning?	4

2	Learning and Training	6
2.1	Learning	6
2.2	Training	7

3	Structure of a Machine Learning Model	8
3.1	Layers and Nodes	8
3.2	Connection Weights	9
3.3	Activation Functions	9
3.4	The Loss Function	13
3.5	Backpropagation	14
3.5.1	Which Weights Should Be Updated?	14
3.5.2	A Quick Way to Train?	16
3.5.3	The Problem With the Weights	16
3.6	Epochs	19
3.7	Batches	21

RSC Foundations No. 8

How to Use Machine Learning in Chemistry: An Introduction

By Hugh M. Cartwright

© Hugh M. Cartwright 2026

Published by the Royal Society of Chemistry, www.rsc.org

4	How Do We Know When Training Should End?	24
4.1	Connection Weights as an Indicator	24
4.2	Underfitting	25
4.3	Overfitting and Generalisation	27
4.4	Validation	30
4.5	The Test Set	35
5	The (Several) Roles of Random Numbers	37
5.1	Unpredictability	37
5.2	Where Random Numbers Are Used	38
5.2.1	Initial Connection Weights	38
5.2.2	Shuffling the Original Dataset	39
5.2.3	Shuffling During Training	41
5.2.4	Other Uses of Random Numbers	45
6	Hyperparameters: How to Make a Good Model Better	46
6.1	What Must Be Optimised?	46
6.2	Number of Input and Output Nodes	47
6.3	Number of Hidden Layers and Nodes	48
6.4	Number of Epochs	52
6.5	Validation and Test Set Sizes	53
6.6	Loss Functions	54
6.7	Learning Rate	56
6.8	Activation Functions at the Hidden Nodes	58
6.9	Activation Functions at the Output Layer	61
6.10	Batch Size	62
7	How to Bungle Training in a Few Easy Steps	64
7.1	Why Even Good Data Need Preprocessing	64
7.2	Before You Touch the Data	64
7.3	A Plan of Action	65

7.4	Cleaning	65
7.5	Noise	68
7.6	Insufficient Data	69
7.7	Excess Data	71
7.8	Encoding	72
7.9	Feature Selection	72
7.10	Splitting the Dataset	73
7.11	Scaling	74
7.12	Standardisation	77
7.13	Why Dataset Splitting Must Come Before Scaling and Standardisation	79
7.14	Bias	80
7.15	Finally...	81
	Reference	82

8 Further Ways to Improve the Model and Data 83

8.1	Biases	83
8.2	Regularisation	84
8.2.1	Early Stopping	84
8.2.2	Weights Penalties	85
8.2.3	Dropout and Dropconnect	85
8.2.4	Pruning	87
8.2.5	Target Penalties	88
8.3	Data Augmentation	88
8.4	<i>k</i> -Fold Cross-validation	90

9 Representation, Descriptors and Properties 93

9.1	Representation	93
9.1.1	3-D Coordinate Representations	94
9.1.2	String-based Representations	95
9.1.3	Graph-based Representations	96
9.1.4	Molecular Fingerprints	96
9.2	Molecular Descriptors	98
9.3	Chemical Properties	100
9.4	Internal or Learned Representations	100

10	Bayesian Optimisation	103
10.1	Another Black Box	103
10.2	Bayesian Optimisation – An Outline	104
10.2.1	Select Initial Values	105
10.2.2	Construction of the Surrogate Model	106
10.2.3	Significance of the Surrogate	108
10.2.4	The Search Strategy	108
10.3	A Complex Optimisation: The Automated Laboratory	109
10.3.1	The First Problem	111
10.3.2	The Second Problem	111
10.3.3	The Third Problem	112
10.4	What Can Be Done?	112
10.5	Challenges for Bayesian Optimisation	114
	References	114
11	Can We Understand How Machine Learning Models Reason?	115
11.1	The Need for Understanding	115
11.2	Interpretability and Explainability	117
11.3	Global and Local	117
11.4	Explanation	118
11.4.1	LIME	118
11.4.2	SHAP	120
11.4.3	Counterfactuals	120
11.4.4	Attention Mechanisms	120
11.4.5	Other Methods	120
11.5	Interpretation	121
11.6	Saliency	122
11.7	Challenges for Understandable Machine Learning	123
12	Other Types of Network	124
12.1	Feedforward Networks	124
12.2	Unsupervised Learning	124
12.3	Recurrent Neural Networks	125
12.4	Deep Neural Networks	126
12.5	Convolutional Neural Networks	127

12.6	Graph Neural Networks	127
12.7	Generative Neural Networks	129
12.8	Transformers	129
12.9	Reinforcement Learning	130
12.10	Transfer Learning	130

13 A Glance Ahead 132

13.1	Federated Learning	132
13.2	Large Language Models	134
13.3	Hybrid Models	135
13.4	Self-driving Labs	136
13.5	Explainable Models	136
13.6	Inexplicable Science	137
	References	138

Appendix: Q and A – And the Evolving Pattern 139

A.1	Problems with Training	140
A.2	Problems with Generalisation	142
A.3	Problems with Activation Functions	142
A.4	Problems with the Validation Loss	143
A.5	Problems with the Training Loss	144
A.6	General Problems	144