

## Contents

**Preface** xv

<b>1</b>	<b>From Genome to Actionable Insights in Biotechnology</b>	<b>1</b>
	<i>James Morrissey, Benjamin Strain, and Cleo Kontoravdi</i>	
1.1	Introduction	1
1.2	From Genome to Network	2
1.2.1	Metabolic Networks	3
1.2.1.1	Bottom-Up Approaches for Network Reconstruction	3
1.2.1.2	Top-Down Approaches for Network Reconstruction	4
1.2.2	Networks Beyond Metabolism	5
1.3	From Draft to Functional Network	6
1.3.1	Additional Reactions	6
1.3.1.1	Exchange Reactions	6
1.3.1.2	Demand Reactions	6
1.3.1.3	Transport Reactions	6
1.3.1.4	Spontaneous Reactions	7
1.3.1.5	Nongrowth Associated ATP Maintenance	7
1.3.1.6	Biomass Reaction	7
1.3.2	Network Validation	7
1.3.2.1	Manual Screening	8
1.3.2.2	Screening for Dead-End Reactions and Blocked Metabolites	8
1.3.2.3	Infinite Loops	9
1.3.2.4	Leaks and Siphons	10
1.4	From Functional Network to Model	10
1.4.1	Flux Balance Analysis	11
1.4.2	Flux Variability Analysis	12
1.4.3	Flux Sampling	13
1.5	From Model to <i>In Silico</i> Predictions	15
1.5.1	Constraints	15
1.5.2	Objective Function	16
1.5.3	Validating <i>In Silico</i> Predictions	16
1.5.3.1	Growth Rate Predictions	22
1.5.3.2	Amino Acid Auxotrophies	22
1.5.3.3	Gene Essentialities	22

1.5.3.4	Known Host Traits	23
1.5.3.5	Intracellular Predictive Accuracy	23
1.5.4	Toward Multilayer, Multiscale Metabolic Networks	23
1.5.4.1	Integrating Gene Regulatory Networks	24
1.5.4.2	Integrating Transcription and Translation	25
1.5.4.3	Integrating Signaling Networks	25
1.5.4.4	Multicellular and Multitissue Models	25
1.5.4.5	Multiscale Bioreactor Models	26
1.6	From Predictions to Actionable Insights in Biotechnology	26
1.6.1	Metabolic Engineering	26
1.6.2	Cell Line Development and Metabolic Profiling	27
1.6.3	Media and Feed Design	29
1.6.4	Gene Essentiality	30
1.6.5	Kinetic Parameter Estimation	30
1.6.6	Process Monitoring and Forecasting	30
	References	31
<b>2</b>	<b>Automated Approaches for the Development of Genome-Scale Metabolic Network Models</b>	<b>43</b>
	<i>Emma M. Glass, Deborah A. Powers, and Jason A. Papin</i>	
2.1	Introduction	43
2.2	Manual GSM Creation	44
2.3	Automated GSM Development	45
2.3.1	General Approach for Automated GSM Methods	45
2.3.2	GSM Construction Tools	46
2.3.2.1	From Raw Sequences	46
2.3.2.2	From Pre-annotated Sequences	52
2.3.2.3	From Reaction Database Information	53
2.3.2.4	Based on Existing GSMs	56
2.3.2.5	GSM Modification and Visualization Tools	57
2.4	Applications of Automatically-Generated GSM Collections	59
2.4.1	AGORA1 – 773 GSMs	59
2.4.2	EMBL GEMs – 5,587 GSMs	60
2.4.3	MetaGEM – 447 GSMs	61
2.4.4	AGORA2–7,302 GSMs	61
2.4.5	PATHGENN – 914 GSMs	62
2.5	Future Directions for the Field of Automated GSM Development	62
2.6	Conclusion	63
	References	63
<b>3</b>	<b>Machine-Guided Approaches for Synthetic Biology Part Design</b>	<b>67</b>
	<i>Marc Amil, Leandro N. Ventimiglia, and Aleksej Zelezniak</i>	
3.1	Introduction	67
3.2	Model-Guided Sequence Design Using Deep Learning	70

3.2.1	Predictive Models for DNA Function: CNNs in Regulatory Sequence Analysis	70
3.2.1.1	Data Considerations for Supervised Learning on Genomic Sequences	72
3.2.1.2	Primer on Convolutional Neural Networks for Supervised Genomic Sequence Modeling	73
3.2.2	Generative Sequence Modeling	75
3.2.2.1	Data Preparation for Unsupervised Learning	76
3.2.2.2	GANs for the Design of Biological Sequences	77
3.2.2.3	Transformer-Based DNA Models	80
3.2.2.4	Diffusion Models for the Design of Biological Sequences	84
3.3	Sources of Sequence–Function Data for Deep Learning	85
3.3.1	Native-Context Genome-Derived Datasets	86
3.3.2	Synthetic Datasets	87
3.4	Evaluating Synthetic Biological Parts Using Motif Analysis and Deep Learning	90
3.5	Current Challenges of Generative Part Design	93
3.6	Conclusion	93
	References	94
<b>4</b>	<b>Machine Learning for Sequence-to-Function Approaches</b>	<b>103</b>
	<i>Rana A. Barghout, Maxim Kirby, Austin Zheng, Lya Chinas, Marjan Mohammadi, Zhiqing Xu, Benjamin Sanchez-Lengeling, and Radhakrishnan Mahadevan</i>	
4.1	Introduction	103
4.2	Current State of Sequence-to-Function Modeling	105
4.2.1	Protein Function Prediction: From BLAST to Language Models	105
4.2.2	Gene Ontology	106
4.2.3	Enzyme Commission Numbers	106
4.2.4	Enzyme Activity	109
4.2.5	Protein Thermal Stability	109
4.2.6	Protein Toxicity	110
4.2.7	Protein Solubility	111
4.3	Tool Kits and Benchmarks	111
4.3.1	Overview of Open-Source Tools	111
4.3.2	Importance of Standardized Benchmarks	114
4.4	Emerging ML Methods	115
4.4.1	Contrastive Learning	115
4.4.2	Meta Learning	116
4.5	Case Studies	116
4.5.1	Prediction of Enzyme Activity and Substrate Specificity	116
4.5.1.1	Predicting $k_{\text{cat}}$ Using CPI-Pred	117
4.6	Challenges in Sequence-to-Function Mapping	118
4.6.1	Sparse Experimental Data	119
4.6.2	Interpretability	120

- 4.7 Conclusion and Future Directions 120  
References 121
- 5 Prediction of Enzyme Functions by Artificial Intelligence 131**  
*Ha Rim Kim, Hongkeun Ji, Gi Bae Kim, and Sang Yup Lee*
- 5.1 Introduction 131
- 5.2 Conventional Computational Approaches for Predicting Enzyme Function 132
- 5.3 Prediction of Enzyme Functions Using Machine Learning 133
- 5.3.1 Extraction of Enzyme Features from Amino Acid Sequences 134
- 5.3.2 Machine Learning-Based Approaches Algorithms for Enzyme Function Prediction 136
- 5.4 Prediction of Enzyme Functions Using Deep Learning 138
- 5.4.1 Convolutional Neural Network 139
- 5.4.2 Recurrent Neural Network 141
- 5.4.3 Transformer and Protein Language Models 142
- 5.4.4 Graph Neural Network 145
- 5.5 Concluding Remarks 147  
Acknowledgments 153  
References 153
- 6 Design of Biochemical Pathways via AI/ML-Enabled Retrobiosynthesis 161**  
*Hongxiang Li, Xuan Liu, and Huimin Zhao*
- 6.1 Introduction 161
- 6.1.1 Computer-Aided Synthesis Planning 161
- 6.1.2 Retrobiosynthesis 162
- 6.2 Retrobiosynthesis Tools 162
- 6.2.1 Template-Based Tools 162
- 6.2.2 Template-Free Tools 166
- 6.2.3 Searching Algorithm 167
- 6.2.4 Ranking 168
- 6.3 Enzyme Selection and Optimization 168
- 6.3.1 Enzyme Substrate Specificity 169
- 6.3.2 Enzyme Catalytic Efficiency 170
- 6.3.3 Enzyme Engineering 172
- 6.3.4 *De Novo* Enzyme Design and Discovery 172
- 6.4 Perspectives 174
- 6.4.1 Integrating Biocatalysis with Chemocatalysis 175
- 6.4.2 Toward Next-Generation AI for Retrobiosynthesis Planning 175
- 6.4.3 Enhancing Enzyme Prediction and Design Capabilities 176
- 6.4.4 Data Standardization and Model Interpretability 177  
Acknowledgments 178  
References 179

<b>7</b>	<b>Machine Learning to Accelerate the Discovery of Therapeutic Peptides</b>	<b>183</b>
	<i>Nicole Soto-Garcia, Mehdi D. Davari, and David Medina-Ortiz</i>	
7.1	Introduction	183
7.2	Peptides: Definitions and Main Characteristics	184
7.3	Benefits and Limitations of Therapeutic Peptides	185
7.4	Computational Design of Therapeutic Peptides	186
7.5	Data Sources for Peptide Discovery	187
7.6	ML-Based Strategies to Accelerate the Discovery of Therapeutic Peptides	189
7.6.1	Data-Driven Approaches	190
7.6.2	ML Strategies for Peptide Bioactivity Classification	192
7.6.2.1	Classification Models for Antimicrobial Peptides	192
7.6.2.2	Classification Models for Antiviral Peptides	193
7.6.2.3	Classification Models for Antifungal Peptides	194
7.6.2.4	Additional Classification Models for Therapeutic Peptides	194
7.6.3	Strategies for Building Toxicity Classification Models	195
7.6.3.1	Toxicity Prediction	196
7.6.3.2	Immunogenicity Identification	197
7.6.3.3	Hemolysis Evaluation	197
7.6.3.4	Other Toxic Adverse Effect Predictions	198
7.6.4	Data-Driven Strategies for Modeling Pharmacological Profiles	198
7.6.5	<i>De novo</i> Design of Therapeutic Peptides	199
7.6.5.1	Variational Autoencoder-Based Approaches	200
7.6.5.2	Generative Adversarial Networks	200
7.6.5.3	Transformer-Based Language Models	201
7.6.5.4	Diffusion Models	201
7.7	Next-Generation Peptide Design Through Multi-Agent Systems	201
7.8	Developing AI-Agent for Autonomous Therapeutic Peptide Design	203
7.9	Conclusion and Perspectives	205
	Acknowledgments	206
	References	207
<b>8</b>	<b>Machine Learning Approaches for High-Throughput Microbial Identification/Culturing</b>	<b>219</b>
	<i>Mohamed Mastouri and Yang Zhang</i>	
8.1	Introduction	219
8.2	High-Throughput (HTP) Techniques in Microbial Research	221
8.2.1	Definition and Scope of HTP Microbial Techniques	221
8.2.2	Metagenomics and Next-Generation Sequencing (NGS)	223
8.2.3	Mass Spectrometry-Based Proteomics (MALDI-TOF MS)	223
8.2.4	Flow Cytometry	224
8.2.5	Microfluidics and Lab-on-a-Chip Systems	224
8.2.6	High-Content Imaging and Phenotyping	225
8.3	Fundamentals of Machine Learning	226

- 8.3.1 Definition of Machine Learning and AI 226
- 8.3.2 Supervised vs. Unsupervised vs. Reinforcement Learning 226
- 8.3.3 ML Algorithms Commonly Used in Microbial Identification 229
- 8.4 Machine Learning Approaches for High-Throughput Microbial Identification 230
  - 8.4.1 Genomic and Metagenomic Data Processing 230
  - 8.4.2 Mass Spectrometry-Based Identification 234
  - 8.4.3 Imaging-Based Identification 236
- 8.5 Machine Learning Approaches for High-Throughput Microbial Culturing 237
  - 8.5.1 ML-Driven Microbial Growth Prediction 237
  - 8.5.2 AI in Microbial Cultivation Process Optimization 238
  - 8.5.3 Synthetic Biology and AI-Driven Strain Engineering 240
- 8.6 Challenges and Limitations of Machine Learning in HTP Microbial Research 241
- 8.7 Future Perspectives and Emerging Trends 242
- 8.8 Conclusion 243
  - Acknowledgments 244
  - References 244
  
- 9 Generative AI for Knowledge Mining of Synthetic Biology and Bioprocess Engineering Literature 253**  
*Zhengyang Xiao and Yinjie J. Tang*
  - 9.1 Introduction 253
  - 9.2 Text Mining Using Knowledge Graph Tools 254
    - 9.2.1 NEKO: A Lightweight Knowledge Graph Tool 254
    - 9.2.2 GraphRAG 256
  - 9.3 LLM-Automated Data Extraction for Machine Learning 258
  - 9.4 Current Limitations 259
  - 9.5 Conclusion 260
    - Acknowledgments 260
    - References 260
  
- 10 Metabolomics: Big Data Approaches 263**  
*Kenya Tanaka, Christopher J. Vavricka, and Tomohisa Hasunuma*
  - 10.1 Introduction 263
  - 10.2 Methods for Metabolomics 264
    - 10.2.1 Preparation of Samples 264
    - 10.2.2 Detection and Quantification 266
  - 10.3 Analysis and Application of Metabolomics Data for Biotechnology 267
    - 10.3.1 Metabolomics for Identification of Pathway Bottlenecks 267
    - 10.3.2 Absolute Metabolomics and Thermodynamic Analyses 272
    - 10.3.3 Evaluation and Optimization of Metabolic Flux 272
  - 10.4 Artificial Intelligence (AI)-Based Metabolomics Data Processing and Analysis 273

10.5	Future Direction of Metabolomics and Its Analysis	274
	References	278
<b>11</b>	<b>Strain Engineering, Flux Design, and Metabolic Production Using Big Data: Ongoing Advances and Opportunities</b>	<b>285</b>
	<i>Rafael S. Costa and Rui Henriques</i>	
11.1	Introduction	285
11.2	Big Data in Biotechnology: Prerequisites for ML-Based Approaches	287
11.2.1	Data Multimodality and Heterogeneity	287
11.2.2	Stratification and Data Transformations	288
11.2.3	Handling of Missing Values and Outliers	289
11.2.4	Longitudinal Studies	290
11.3	Types of ML-Based Approaches for the Design of Microbial Cell Factories	291
11.3.1	Machine Learning (ML) Models	291
11.3.2	Supervised ML	292
11.3.2.1	Neural Networks	292
11.3.2.2	Decision Trees and Ensembles	293
11.3.2.3	Alternative Predictive Approaches	294
11.3.2.4	Selected Case Studies	294
11.3.3	Unsupervised ML	296
11.3.3.1	Clustering	296
11.3.3.2	Biclustering	297
11.3.3.3	Representation Learning and Dimensionality Reduction	297
11.3.4	Hybrid ML and Constraint-Based Models	298
11.3.4.1	CBM-FBA as Input for ML	299
11.3.4.2	ML as Input for CBM-FBA	299
11.4	ML-Based Approaches in Microbial Cell Factory Design: Case Studies	300
11.4.1	ML-Based Approaches in Strain Engineering and Flux Design	301
11.4.2	Application of ML-Based Approaches for Metabolic Production	304
11.5	Conclusions and Perspectives	306
	References	308
<b>12</b>	<b>Next-Generation Metabolic Flux Analysis Using Machine Learning</b>	<b>317</b>
	<i>Ahmed Almunaifi, Richard C. Law, Samantha O'Keeffe, Kartikeya Pande, Tongjun Xiang, Onyedika Ukwueze, Aranaa Odai-Okley, Pin-Kuang Lai, and Junyoung O. Park</i>	
12.1	Introduction	317
12.2	Dynamic Nature of Metabolism	317
12.3	Flux Balance Analysis and Metabolic Flux Analysis	321
12.4	Incorporating Machine Learning into Metabolic Flux Analysis	324
12.4.1	Challenges in Applying ML to MFA	325

- 12.4.1.1 A Lack of Isotope Labeling Patterns and Flux Data for Training Machine Learning Models 325
- 12.4.1.2 Variable-Size Input of Isotope Labeling Patterns for Flux Prediction 326
- 12.4.1.3 Incorporation of Disparate Data Types 326
- 12.4.1.4 Initial Computational Cost 326
- 12.4.2 Available Computational Tools 326
- 12.4.2.1 Sampling Metabolic Fluxes for Training ML Models 327
- 12.4.2.2 Atom Mapping Throughout Metabolic Networks 327
- 12.4.2.3 Selection of Information-Rich Isotope Tracers 327
- 12.4.2.4 Other Helpful ML Tools 328
- 12.4.3 A General Workflow for Machine Learning-Based Metabolic Flux Analysis 328
- 12.4.3.1 Metabolic Model Construction 329
- 12.4.3.2 Acquiring Training Data 330
- 12.4.3.3 Training Machine Learning Models 330
- 12.4.3.4 Evaluation of Machine Learning Models 330
- 12.4.3.5 Execution: Computing Metabolic Fluxes Directly from Isotope Labeling Patterns 331
- 12.4.4 Improved Speed and Accuracy of ML-Based MFA 331
- 12.4.5 Toward Dynamic Flux and Isotope Labeling Analysis 332
- 12.5 Future Outlook 333
- References 334

### **13 Streamlining the Design-Build-Test-Learn Process in Automated Biofoundries 341**

*Enrico Orsi, Nicolás Gurdo, and Pablo I. Nikel*

- 13.1 Introduction 341
- 13.2 The Design-Build-Test-Learn Cycle (DBTLc) 342
- 13.2.1 The DBTLc Components 342
- 13.2.2 Application of the DBTLc for Strain Engineering 345
- 13.2.3 Description of Biofoundries and Their Operating Parts 345
- 13.2.4 Laboratory Workflows with Potential of Automation in Biofoundries 347
- 13.3 Geographical Distribution of Biofoundries Around the Globe 350
- 13.4 Challenges for Implementing Fully Automated Biofoundries 352
- 13.5 Perspectives and Outlook 355
- Acknowledgments 357
- Ethics Declarations 357
- References 357

### **14 Machine Learning-Enhanced Hybrid Modeling for Phenotype Prediction and Bioreactor Optimization 367**

*Oliver Pennington, Yirong Chen, Youping Xie, and Dongda Zhang*

- 14.1 Bioprocess Modeling and Optimization 367
- 14.1.1 Challenges in Bioprocess Modeling 367

14.1.2	Machine Learning for Bioprocess Modeling	369
14.1.2.1	Principal Component Analysis and Partial Least Squares	369
14.1.2.2	Artificial Neural Networks	370
14.1.2.3	Gaussian Processes	370
14.1.2.4	Ensemble Learning	371
14.1.2.5	Reinforcement Learning	371
14.1.2.6	Future Prospects for Machine Learning	371
14.1.3	Introduction to Hybrid Modeling	372
14.1.3.1	Literature Review	372
14.1.3.2	Hybrid Modeling for Biosystem Optimization	373
14.2	Methodology for Hybrid Modeling	378
14.2.1	Mechanistic Model Construction	378
14.2.2	Time-Varying Parameter Estimation	380
14.2.3	Machine Learning Model Construction	381
14.2.4	Hybrid Model Uncertainty Estimation	383
14.2.5	Considerations for Hybrid Model Construction	384
14.2.5.1	Hybrid Model Greyness	384
14.2.5.2	Discrepancy Hybrid Modeling	384
14.2.5.3	Machine Learning Component	385
14.2.6	Dynamic Optimization Using Hybrid Modeling	385
14.2.6.1	Optimal Feeding in Fed-Batch Bioprocesses	386
14.2.6.2	Dynamic Optimization Under Uncertainty	386
14.3	Hybrid Modeling of Microalgal Lutein Production	387
14.3.1	Introduction to Microalgal Lutein Production	387
14.3.2	Experimental Setup and Data Availability	388
14.3.3	Constructing a Hybrid Model for Microalgal Lutein Production	389
14.3.3.1	Preliminary Kinetic Modeling of Microalgal Lutein Production	389
14.3.3.2	Artificial Neural Network Implementation for Time-Varying Parameters	391
14.3.4	Dynamic Optimization of Microalgal Lutein Production	394
14.4	Summary	397
	References	398

<b>Index</b>	407
--------------	-----

