

Contents

Preface — V

Introduction — XIII

Technical setup and naming conventions — XV

- 1 Data science: introduction — 1**
- 2 Data science: the “fourth paradigm” of science — 4**
- 3 Relations to other domains and cheminformatics — 8**

Part A: IT, Data science, and AI

IT basics (cloud, REST, edge)

- 4 Cheminformatics application landscape — 13**
- 5 Cloud, fog, and AI runtime environments — 15**
- 6 DevOps, DataOps, and MLOps — 19**
- 7 High-performance computing (HPC) and cluster — 23**
- 8 REST and MQTT — 28**
- 9 Edge devices and IoT — 34**

Programming

- 10 Python and other programming languages — 41**
- 11 Python standard libraries and Conda — 45**
- 12 IDE's and workflows — 55**
- 13 Jupyter notebooks — 61**
- 14 Working with notebooks and extensions — 68**
- 15 Notebooks and Python — 73**
- 16 Versioning code and Jupyter notebooks — 80**
- 17 Integration of Knime and Excel — 83**

Data engineering

- 18 Big data — 91**
- 19 Jupyter and Spark — 96**
- 20 Files: structure representations — 102**
- 21 Files: other formats — 106**

22	Data retrieval and processing: ETL — 111
23	Data pipelines — 114
24	Data ingestion: online data sources — 117
25	Designing databases — 125
26	Data science workflow and chemical descriptors — 130

Data science as field of activity

27	Community and competitions — 139
28	Data science libraries — 143
29	Deep learning libraries — 145
30	ML model sources and marketplaces — 148
31	Model metrics: MLFlow and Ludwig — 152

Introduction to ML and AI

32	First generation (logic and symbols) — 161
33	Second generation (shallow models) — 164
34	Second generation: regression — 170
35	Decision trees — 179
36	Second generation: classification — 182
37	Second generation: clustering and dimensionality reduction — 190
38	Third generation: deep learning models (ANN) — 196
39	Third generation: SNN – spiking neural networks — 204
40	xAI: eXplainable AI — 206

Part B: Jupyter in cheminformatics

Physical chemistry

41	Crystallographic data — 213
42	Crystallographic calculations — 215
43	Chemical kinetics and thermochemistry — 219
44	Reaction paths and mixtures — 223
45	The periodic table of elements — 227
46	Applied thermodynamics — 230

Material science

- 47 **Material informatics — 235**
- 48 **Molecular dynamics workflows — 237**
- 49 **Molecular mechanics — 241**
- 50 **VASP — 245**
- 51 **Gaussian (ASE) — 249**
- 52 **GROMACS — 251**
- 53 **AMBER, NAMD, and LAMMPS — 255**
- 54 **Featurize materials — 261**
- 55 **ASE and NWChem — 266**

Organic chemistry

- 56 **Visualization — 273**
- 57 **Molecules handling and normalization — 280**
- 58 **Features and 2D descriptors (of carbon compounds) — 287**
- 59 **Working with molecules and reactions — 294**
- 60 **Fingerprint descriptors (1D) — 299**
- 61 **Similarities — 306**

Engineering, laboratory, and production

- 62 **Laboratory: SILA and AnIML — 313**
- 63 **Laboratory: LIMS and daily calculations — 319**
- 64 **Laboratory: robotics and cognitive assistance — 325**
- 65 **Chemical engineering — 330**
- 66 **Reactors, process flow, and systems analysis — 334**
- 67 **Production: PLC and OPC/UA — 338**
- 68 **Production: predictive maintenance — 342**

Part C: Data science

Data engineering in analytic chemistry

- 69 **Titration and calorimetry — 349**
- 70 **NMR — 352**

- 71 X-ray-based characterization: XAS, XRD, and EDX — 354**
- 72 Mass spectroscopy — 359**
- 73 TGA, DTG — 361**
- 74 IR and Raman spectroscopy — 364**
- 75 AFM and thermogram analysis — 368**
- 76 Gas chromatography-mass spectrometry (GC-MS) — 373**

Applied data science and chemometrics

- 77 SVD chemometrics example — 379**
- 78 Principal component analysis (PCA) — 383**
- 79 QSAR: Quantitative structure–activity relationship — 386**
- 80 DeepChem: binding affinity — 389**
- 81 Stoichiometry and reaction balancing — 393**

Applied artificial intelligence

- 82 ML Python libraries in chemistry — 399**
- 83 AI in drug design — 405**
- 84 Automated machine learning — 409**
- 85 Retrosynthesis and reaction prediction — 412**
- 86 ChemML — 419**
- 87 AI in material design — 423**

Knowledge and information

- 88 Ontologies and inferencing — 429**
- 89 Analyzing networks — 436**
- 90 Knowledge ingestion: labeling and optical recognition — 443**
- 91 Content mining and knowledge graphs — 450**

Part D: Quantum computing and chemistry introduction

- 92 Quantum concepts — 461**
- 93 QComp: technology vendors — 471**
- 94 Quantum computing simulators — 477**
- 95 Quantum algorithms — 480**
- 96 Quantum chemistry software (QChem) — 485**

Quantum computing applications

- 97 Application examples — 491
- 98 Simulating molecules using VQE — 495
- 99 Studies on small clusters of LiH, BeH₂, and NaH — 500
- 100 Quantum machine learning (QAI) — 504

Code index — 511

Index — 519