

Molecular Wave Mechanics

*A Probabilistic and Quantum
Approach to Molecular Biology*

From De Broglie Waves to Genomics,
Protein Structure, and Immunotherapy

Abdelkader BOUSABAA

University of Paris-Saclay

Volume I — First Edition — 2026

Contents

Preface	xiii
1 Why Molecular Biology Needs a New Language	1
1.1 The data avalanche of the post-genomic era	2
1.2 The four hard problems that data alone has not solved	3
1.3 Three communities, three vocabularies, one object	5
1.4 The diagnosis: what current methods get wrong	6
1.5 Why neural networks need a theory	9
1.6 The thesis of this book: one wave function, three threads	10
1.7 Roadmap of the book	11
1.8 How to read this book, and the central conjecture in plain sight . .	13
I Mathematical Foundations	15
2 Probability and Statistics for the Molecular Sciences	17
2.1 Sample spaces and biological outcomes	18
2.2 Random variables for sequencing reads and expression counts . . .	19
2.3 Limit theorems and bulk-versus-single-cell averaging	20
2.4 Estimators, bias, and the maximum-likelihood principle	21
2.5 Resampling, bootstrap, and permutation	22
2.6 A wave-mechanical preview: probability and the Born rule	23
2.7 Outlook: from probability to information	24
3 Information Theory and Entropy	27
3.1 Shannon entropy of a biological alphabet	28
3.2 Differential entropy and continuous distributions	30
3.3 Mutual information and co-evolution	30
3.4 KL divergence, cross-entropy, and the maximum-likelihood loss . .	31
3.5 Sequence motifs as information bottlenecks	33
3.6 A wave-mechanical preview: fidelity, overlap, and entanglement entropy	33
3.7 Outlook: from information to optimisation	35
4 Optimization and Variational Methods	37
4.1 Convex versus non-convex objectives	38
4.2 Gradient descent, momentum, and adaptive methods	39
4.3 Stochastic gradient and the noise that helps	39

4.4	Constrained optimization and Lagrangians	40
4.5	Variational principles in physics, machine learning, and biology . .	41
4.6	A wave-mechanical preview: the Rayleigh quotient as the ur- variational principle	42
4.7	Outlook: from optimisation to inference	44
5	Bayesian Inference	45
5.1	Bayes' rule, the posterior, and the evidence	46
5.2	Conjugate priors and credible intervals	47
5.3	Markov chain Monte Carlo	49
5.4	Variational Bayes and amortised inference	50
5.5	Hierarchical models and empirical Bayes	50
5.6	A wave-mechanical preview: the posterior as a wave function on parameter space	51
5.7	Outlook: Part I closes, Part II opens	53
II	Machine Learning and Neural Networks	55
6	Statistical Learning Theory: Generalisation and the Variational Certificate	57
6.1	Bias, variance, and the generalisation gap	58
6.2	PAC learning and sample complexity	59
6.3	Regularisation as the imposition of a prior	60
6.4	Double descent and the modern picture	60
6.5	Wave-mechanical perspective: the variational gap as a certificate .	62
6.6	Outlook	63
7	Neural Networks and Backpropagation: The Variational Bridge to Wave Mechanics	65
7.1	Neural networks as universal function approximators	67
7.2	Backpropagation from the chain rule	67
7.3	The variational principle and neural quantum states	68
7.4	The bidirectional bridge	69
7.5	Worked example: a neural quantum state for the folding double-well	70
7.6	Why depth helps: composition, hierarchy, and coarse-graining . . .	71
7.7	Wave-mechanical perspective	73
7.8	Outlook: the bridge in action	74
8	CNNs, RNNs, and Transformers: Architectures as Wave-Function Ansätze	75
8.1	Convolutional networks: locality and translation invariance	76
8.2	Recurrent networks: sequentiality and its limits	77
8.3	Self-attention and the transformer block	77
8.4	Inductive biases and the no-free-lunch theorem	78
8.5	Self-attention as the pair-correlation operator	79
8.6	Wave-mechanical perspective	81

8.7	Outlook	81
9	Generative Models: VAEs, GANs, Diffusion	83
9.1	Latent variable models and the ELBO	84
9.2	Variational autoencoders	85
9.3	Generative adversarial networks	86
9.4	Diffusion and score-based models	87
9.5	The Wick-rotation bridge: diffusion as imaginary-time wave mechanics	90
9.6	Wave-mechanical perspective	92
9.7	Outlook: from generation to foundation	93
10	Foundation Models and Self-Supervised Learning	95
10.1	Self-supervision: masked, autoregressive, contrastive	96
10.2	Scaling laws and emergent abilities	97
10.3	Transfer, fine-tuning, and prompting	98
10.4	The biological foundation-model zoo	99
10.5	The universal wave function	101
10.6	Wave-mechanical perspective	102
10.7	Outlook: from the bridge to the territory	103
11	Machine Learning for Biological Sequences and Structures	105
11.1	Sequence embeddings: from one-hot to language-model features . .	106
11.2	Geometric deep learning for structures	107
11.3	Multi-modal models: sequence, structure, expression	108
11.4	Transfer learning across organisms and modalities	109
11.5	The synthesis of Part II: one bridge, four bricks	109
11.6	Wave-mechanical perspective	111
11.7	Outlook: crossing into Part III	112
III	De Broglie Waves in Molecular Biology	113
12	De Broglie Waves for Biomolecules	115
12.1	From Einstein to Arndt: the experimental bridge to biology . . .	116
12.2	The thermal de Broglie wavelength	118
12.3	The migration of wave nature into internal degrees of freedom . .	121
12.4	Decoherence in the cellular environment	123
12.5	Coherence length and the biological Goldilocks zone	125
12.6	Three regimes: a working classification	126
12.7	Empirical fingerprints of the wave nature in living systems	128
12.8	From kinematics to dynamics	130
13	A Biological Schrödinger Equation	131
13.1	Why biology needs a Schrödinger-like equation	132
13.2	The biological Schrödinger equation	133
13.3	Coarse-graining: from atoms to slow coordinates	134
13.4	The folding potential $V(q)$	135

13.5	Stationary states as native folds	136
13.6	Time evolution and conformational sampling	138
13.7	The Wick rotation: bridge to classical biology	139
13.8	From dynamics to applications	141
14	Tunnelling, Decoherence, and Enzymatic Catalysis	143
14.1	The phenomenon: how barriers leak	144
14.2	The WKB approximation	146
14.3	Decoherence and tunnelling: when does the barrier leak in a cell? .	148
14.4	Enzymatic catalysis as wave-assisted passage	149
14.5	Kinetic isotope effects: the empirical signature	150
14.6	Electron tunnelling and Marcus theory	151
14.7	Outlook: from tunnelling to coherent transport	152
15	Cellular Wave Function and Stochastic Gene Expression	155
15.1	Gene expression is stochastic — and surprising	156
15.2	The chemical master equation	157
15.3	From the master equation to the cellular wave function	158
15.4	Bursting dynamics as wave-packet collapse	160
15.5	Cell-fate priming and superposition	161
15.6	Decoherence in cell populations	162
15.7	Empirical signatures in single-cell data	163
15.8	Outlook	163
16	Entanglement and Long-Range Biomolecular Correlations	165
16.1	From local to non-local: the question	166
16.2	The mathematics of entanglement	167
16.3	Allosteric correlations as non-factorisable wave functions	168
16.4	Co-evolutionary couplings and direct coupling analysis	169
16.5	Long-range information transfer in biomolecules	172
16.6	Entanglement signatures in MSA Transformers and AlphaFold . .	173
16.7	Outlook: closing Part III	173
IV	Sequence: Genomics and Transcriptomics	175
17	NGS Pipelines and the Wave Function over Sequence Space	177
17.1	Sequencing as the empirical input to modern biology	178
17.2	From reads to alignment: the canonical pipeline	179
17.3	Coverage, deduplication, and base quality	180
17.4	Sequence space as a high-dimensional Hilbert space	182
17.5	The wave-mechanical reading of an NGS pile-up	183
17.6	Wave-mechanical perspective	184
17.7	Outlook	184
18	Variant Calling: From Likelihoods to Quantum Measurements	187
18.1	GATK and likelihood-based calling	188

18.2	Deep learning callers: DeepVariant and successors	190
18.3	HMM-based and graph-based callers	191
18.4	Calibration, recall, and the long tail of rare variants	192
18.5	Wave-mechanical perspective	192
18.6	Deepstrand application: construct QC against PCR-introduced mutations	193
18.7	Outlook	195
19	RNA-seq Differential Expression and Wave Packets in Transcript Space	197
19.1	Count models: Poisson, negative binomial, and over-dispersion . . .	198
19.2	DESeq2, edgeR, and limma-voom	199
19.3	Empirical Bayes shrinkage of dispersions	200
19.4	The volcano plot and multiple testing	201
19.5	Wave-mechanical perspective	202
19.6	Deepstrand application: predicting shRNA knockdown efficiency .	203
19.7	Outlook	205
20	Single-Cell Trajectories: Cell States in an Expression Hilbert Space	207
20.1	From bulk to single-cell: the resolution revolution	208
20.2	Pre-processing: normalisation, dropout, and batch correction . . .	209
20.3	Dimensionality reduction and clustering	210
20.4	Trajectory inference and pseudotime	212
20.5	Wave-mechanical perspective	213
20.6	Outlook: from scRNA-seq to immune repertoires	214
V	Structure: Proteins and Macromolecules	215
21	AlphaFold, ESMFold, and the Energy Landscape as Effective Hamiltonian	217
21.1	Levinthal's paradox revisited	218
21.2	The AlphaFold2 architecture: Evoformer and structure module . .	220
21.3	ESMFold and the language-model-only regime	221
21.4	What AlphaFold does not solve	222
21.5	Wave-mechanical perspective	223
21.6	Deepstrand application: validating the cloned protein fold	223
21.7	Outlook	225
22	Docking and Binding Affinity: Tunnelling at Transition States	227
22.1	Classical scoring functions and free energy	228
22.2	DiffDock and ML-based docking	229
22.3	Alchemical free-energy methods: FEP, TI, and BAR	230
22.4	Transition states and tunnelling at binding	231
22.5	Wave-mechanical perspective	232
22.6	Outlook	233

23 Protein Design: Variational Principle for the Folded State	235
23.1 RFdiffusion and structure generation	236
23.2 ProteinMPNN and inverse folding	237
23.3 Functional design: binders, enzymes, switches	238
23.4 Validation: from in silico to wet lab	239
23.5 Wave-mechanical perspective	240
23.6 Deepstrand application: codon optimisation via inverse folding . .	241
23.7 Outlook	242
24 Allostery and Dynamics: Phonons and Coupled Oscillators	245
24.1 Elastic network models and normal modes	246
24.2 Anisotropic and Gaussian network models	247
24.3 ML-based protein dynamics: DynamicBind, AlphaFlow, ESM-Flow	249
24.4 Cryptic pockets and ligand-induced allostery	250
24.5 Wave-mechanical perspective	251
24.6 Outlook: closing Part V	252
VI Signal: Cancer and Immunotherapy	253
25 TCR and BCR Repertoires as Wave Packets	255
25.1 V(D)J recombination and the source of diversity	256
25.2 Repertoire sequencing: TCR-seq and BCR-seq	257
25.3 Diversity metrics: richness, evenness, Hill numbers	258
25.4 Clonotype tracking and antigen-specific expansion	259
25.5 Wave-mechanical perspective	261
25.6 Outlook	262
26 Neoantigen Prediction as a Superposition of Binding Modes	263
26.1 HLA-peptide binding: from motifs to predictors	264
26.2 Antigen processing and presentation	265
26.3 TCR-pMHC recognition	266
26.4 Pipelines, benchmarks, and the false-positive burden	266
26.5 Wave-mechanical perspective	268
26.6 Outlook	269
27 CAR-T Design as an Engineered Wave Function	271
27.1 CAR architecture: scFv, hinge, transmembrane, signalling	272
27.2 Antigen selection: target validity, escape, off-target	273
27.3 Affinity tuning by ML and directed evolution	274
27.4 Toxicity: cytokine release, neurotoxicity, on-target/off-tumour . .	275
27.5 Wave-mechanical perspective	276
27.6 Deepstrand application: assembling CAR constructs	277
28 Clinical Response Prediction: The Measurement Problem in the Clinic	279
28.1 Survival analysis: Kaplan–Meier, Cox, accelerated failure time . .	280

28.2	Machine learning for response prediction	282
28.3	Biomarkers, signatures, and digital twins	283
28.4	Causality and counterfactuals	283
28.5	Wave-mechanical perspective	284
28.6	Outlook: closing Part VI	285
VII	Perspectives	287
29	Quantum Biology: Empirical Evidence	289
29.1	Photosynthetic coherence: the FMO complex and beyond	290
29.2	Avian magnetoreception and the radical-pair mechanism	291
29.3	Enzymatic tunnelling: kinetic isotope effects beyond the classical limit	294
29.4	Olfaction: the vibrational theory and its critics	295
29.5	The honest scoreboard	295
29.6	Open empirical questions and how to test them	296
29.7	Outlook: from where biology is quantum to whether quantum computers can help	297
30	Quantum Machine Learning for Genomics and Structure	299
30.1	Quantum circuits, gates, and the noisy hardware era	300
30.2	Variational quantum eigensolvers and the quantum neural quantum state	301
30.3	Quantum kernels and feature maps	303
30.4	Quantum-enhanced sampling and Hamiltonian dynamics	304
30.5	Where the realistic biological wins are	305
30.6	Honest limits of the NISQ-era programme	306
30.7	Outlook: from the quantum hardware horizon to the universal- wave-function conjecture	307
31	Foundation Models as Universal Wave Functions: Open Problems	309
31.1	The conjecture, stated precisely	310
31.2	Why foundation models converge to wave-like representations . . .	311
31.3	The state of the partial evidence	312
31.4	The unsolved problems: dynamics, multi-scale, causality	314
31.5	A research agenda for molecular wave mechanics	315
31.6	The path to Volume II: Numerical Recipes for Molecular Wave Mechanics	316
31.7	Coda: the wave is launched	317
	Bibliography	319

Preface

This book emerged from the meeting of two currents that, for many years, I believed to be distinct: more than two decades of multi-domain teaching at the University of Paris-Saclay, and a five-year passage through molecular oncology that I owe almost entirely to a single act of trust from a single colleague at the Institut Curie. The book is the place where these two currents finally meet. It would not exist without either of them, and it would certainly not exist in its present form without the second.

The first current is teaching. For more than twenty years I have lectured at the University of Paris-Saclay and in adjacent *grandes écoles* across a deliberately broad set of fields — applied mathematics, stochastic methods, scientific computing, data science with Python, and, increasingly over the last decade, the artificial-intelligence techniques that have transformed every quantitative discipline they have touched. The courses I have given at Paris-Saclay span domains that the textbooks usually keep apart: the same neural-network architectures applied to physical, financial, and biological data; the same Bayesian machinery used to infer market dynamics one week and protein conformations the next. The data-science course in particular is one of the direct ancestors of this book; its slides, worked examples, Python notebooks, and analogies were tested on cohort after cohort of students before they reached these pages. The students of Paris-Saclay, with their relentless, unembarrassed “but *why?*”, kept the cross-domain analogies alive in my mind for years before the biological applications matured into the chapters that follow.

The second current — the decisive one — is the one I owe most directly. Between the spring of 2018 and the late summer of 2023, on a path I could not have foreseen from the lecture room, I led for five and a half years a biotech company created within the orbit of the *Institut Curie*. The mission was as ambitious as the science was new: to use computational methods, and in particular the artificial-intelligence techniques I had developed and was teaching at Paris-Saclay, to advance the design of T-cell-based immunotherapies against cancer — the CAR-T frontier as it was then opening up.

I had no formal training in immunology. I had read about T cells, I had a working understanding of antigen recognition, but I had never seen, before that period, what immuno-oncology data actually look like once they leave the laboratory: small cohorts, fewer patients than features, signals buried in noise, and decisions of life-and-death importance to be made nonetheless. The conventional wisdom, at

the time and to a large extent still today, was that machine learning required tens of thousands of examples and that biology rarely supplied them. The conventional wisdom was wrong, and a single person at the Institut Curie was willing to bet on it being wrong.

That person is **Sebastian Amigorena**. INSERM director, head of an immunology research unit at the Institut Curie, internationally known for his work on antigen presentation and cross-presentation by dendritic cells, Sebastian had no obligation to trust a quantitative analyst with a background in applied mathematics and data science. He could have insisted that the AI methods I proposed be validated on the kinds of large cohorts that the machine-learning community considered respectable. Instead, he trusted me to apply those methods to the data we actually had — a small set of tumour-infiltrating lymphocyte (TIL) repertoires from a modest number of patients — and to make them speak. The methodology that emerged from that collaboration, refined under the pressure of biological reality and of stubborn clinical questions, is the methodological core of Part VI of this book. Every chapter on TCR and BCR repertoires (Chapter 25), neoantigen prediction (Chapter 26), CAR-T design (Chapter 27), and clinical response prediction (Chapter 28) carries Sebastian’s intellectual fingerprint. The errors that remain are mine; the courage to attempt the work in the first place was his.

To Sebastian, and to the Institut Curie teams who patiently taught a quantitative analyst the language of molecular oncology, I am more grateful than these pages can express. The clinicians and biologists who walked me through their data, the post-docs who corrected my misunderstandings of immune signalling, the technicians who explained how a TIL extraction actually works at the bench, the colleagues in adjacent groups who shared their reagents and their time — all of them are present in this book, even where I have not named them individually. The Institut Curie is, in its everyday operation, a remarkable laboratory of collaborative science, and I was lucky to be allowed to inhabit it for a while.

A second tangible outcome of those five and a half years was **Deepstrand**, a molecular biology software toolkit I built to support the experimental cloning workflows of the laboratory. Deepstrand automates the routine operations of construct design: creating PCR-based fragments, generating shRNA constructions on pre-defined scaffolds, building constructs, designing synthetic adaptors, extracting sub-sequences either by selection or between restriction sites, supporting Gateway™ cloning in both BP and LR reactions, aligning two sequences with a local algorithm, and assembling final constructs from PCR fragments, adaptors and shRNAs, or directly from restriction-bounded sub-sequences chosen graphically. Deepstrand was the bench-side companion to the AI work I was doing on TIL repertoires; it is the tool through which the methodological ideas of this book met the daily reality of the molecular biology bench.

The chapters of Parts IV through VII of this book are, in many ways, the methodological extension of Deepstrand. Each addresses a problem that the wet lab and its tooling could only partially solve in the 2018–2023 period: validating that a cloned construct carries the intended sequence and no spurious mutations

(Chapter 18), predicting the knockdown efficiency of a designed shRNA before committing to the bench (Chapter 19), checking the predicted fold of an assembled protein construct (Chapter 21), choosing codons that maximise expression (Chapter 23), engineering TCR and BCR chains for therapeutic use (Chapter 25), assembling CAR-T constructs with quantified off-target risk (Chapter 27), and ultimately driving the next generation of bench tools with biological foundation models (Chapter 31). Each of these chapters proposes a wave-mechanical or AI-based enhancement that the next generation of Deepstrand could absorb; each is offered as a methodological recipe, not only a theoretical statement.

Looking back, I can see that the two currents were always connected. The data-science and AI methods I had been teaching at Paris-Saclay — Bayesian inference, deep learning, foundation models, the whole modern toolkit — turned out to be precisely what the molecular oncologists of the Institut Curie needed in order to extract structure from small, noisy clinical datasets. The biology and the mathematics were two readings of the same underlying object: a noisy, partially observed system whose probability density is the square of a complex amplitude. The wave-mechanical reading developed in Part III of this book is the unified language those two readings share.

This book is the attempt to make that unity explicit. It is written for three audiences at once. The data scientist will find a familiar formalism (Parts I and II) applied to an unfamiliar domain. The biologist or clinician will find a structural account of why neural-network predictors work as well as they do (Parts IV and V), and why the residual mystery in immunotherapy response (Part VI) is exactly the kind of mystery that the wave-mechanical reading was designed to address. The physicist or applied mathematician will find a coherent dynamical framework (Part III) that connects the kinematics of de Broglie waves to the empirical machinery of modern molecular biology.

I have tried to make the book three things at once: rigorous (every result is derived, not asserted), intuitive (every formula is accompanied by an analogy or a visualisation), and practical (every chapter is built around concrete biological examples drawn either from the published literature or from the work I conducted at the Institut Curie and elsewhere). Whether I have succeeded is for the reader to judge.

A word on how to read this book. Parts I and II recall the data-science foundations — probability and statistics, information theory, optimisation, Bayesian inference, neural networks, foundation models — in the dialect that biology actually speaks. Readers comfortable with this material may skim it. Part III introduces the wave-mechanical formalism in five focused chapters; this is the formal core, and the chapters of Parts IV through VI lean on it explicitly. Parts IV (genomics), V (structure), and VI (immunotherapy) take the formalism to work on the applied questions. Part VII looks forward.

To the colleagues, students, and patients whose questions, data, and patience made this book possible: thank you. To Sebastian Amigorena, to whom I owe

the second current entirely: thank you. To the Institut Curie, whose corridors I will always associate with the moment my quantitative training began to speak a biological language: thank you.

Any errors that remain are mine alone. The book is offered not as a finished theory but as a beginning — a first exploration of a territory that I believe is rich with undiscovered structure, and that the immunotherapy of the next decade will, I hope, help us to map.

Abdelkader Bousabaa

Paris, May 2026

The genome is the sheet of music. The cell is the concert. The wave function is the music.

AlphaFold, ESM, scGPT, RFdiffusion — read carefully, the most successful methods of biological machine learning are all one construction: a neural network used as a variational approximation to a wave function on a biological configuration space. Naming that construction is what this book does, and it follows it from the mathematics to the clinic.

Volume I builds the framework in seven parts: the mathematics and the machine learning — probability, information theory, variational methods, neural networks, transformers, generative and foundation models; then the wave mechanics proper — de Broglie wavelengths for biomolecules, a biological Schrödinger equation, decoherence and enzymatic tunnelling, the cellular wave function; then the bench — NGS pipelines and variant calling, RNA-seq and single-cell trajectories, structure prediction, docking and protein design, T-cell and B-cell repertoires, neoantigen prediction, CAR-T engineering and clinical response. The companion **Volume II**, *Numerical Recipes*, supplies the theorems, the derivations and the solved exercises, chapter by chapter.

Abdelkader BOUSABAA has taught applied mathematics, stochastic methods, scientific computing and artificial intelligence at the University of Paris-Saclay for more than twenty years. Between 2018 and 2023 he led a biotech company in the orbit of the *Institut Curie*, applying these methods to the design of T-cell immunotherapies against cancer — with the cohorts biology actually supplies: more features than patients, and decisions of consequence to be made anyway.