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

Seven Lions Editions

Contents

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.

Seven Lions Editions