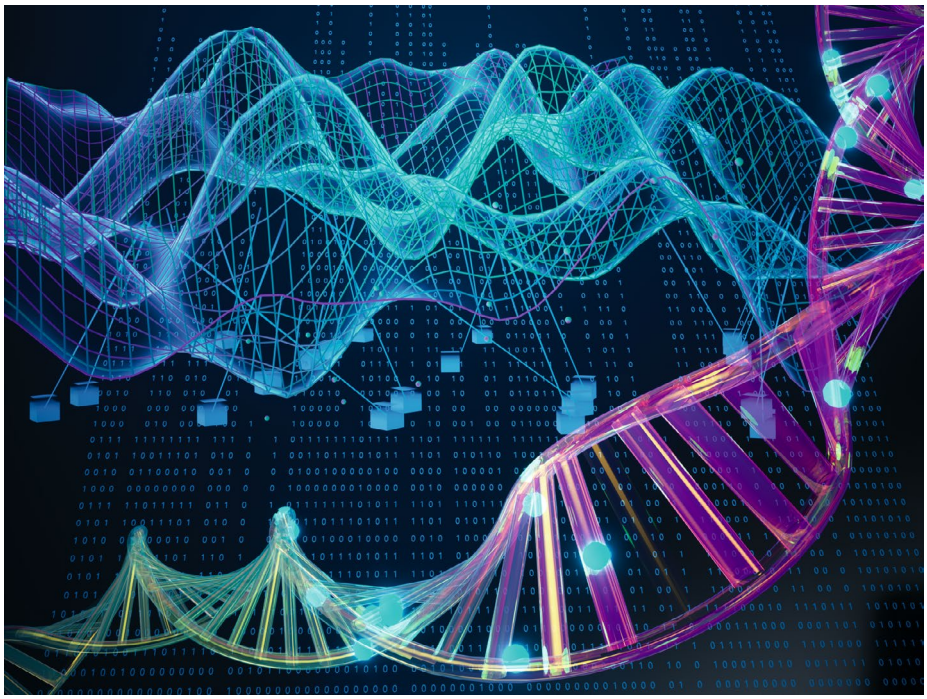
Pooling the strengths of data and models

The availability of higher-quality biomedical and clinical data is widening the reach and usefulness of data-fitted biophysical models and of data-driven mathematical and statistical modelling.

Scarce samples, small data and the limited availability of computing capacity and data storage have long constrained the acquisition of scientific knowledge. Yet this didn't prevent Charles Darwin, James Clerk Maxwell, Gregor Mendel, and Rosalind Franklin, James Watson and Francis Crick from describing how natural selection works, the behaviour of electric and magnetic fields, how genetic inheritance functions, and how DNA stores information. Most scientific laws describing aspects of the natural world (and scientific theories explaining it) have been originally put forward on the basis of limited actual information and via careful reasoning and maths, and then validated over time with more observations and repeated experiments. The strengths of mathematical and physical models are indisputable.

However, in a world increasingly awash with data, it is tempting to leave biophysical models behind and let model-agnostic computational techniques (often using machine learning) find patterns that describe the data. After all, not everything can be put down into tractable mathematical formulae. And the bigger and messier the datasets, the harder is to work out formal relationships that explain what is going on. Nonetheless, when available or feasible, data-driven mathematical and statistical models can provide mechanistic understanding, generalization and explainability, and can be used to predict outcomes. And even when necessarily simplistic, analytical models can capture the essence of the problem or be used to derive insight from the data.

Many biomedical problems and systems of pressing relevance are too complex to be feasibly or practically described via analytical biophysical models. Instead, pure data-driven modelling does not require predefined rules or hypotheses about the problem or system (and the biases that come with such assumptions). But data-driven models can be blind to all sorts of biases in the data, may find irrelevant or unphysical patterns, or fail to converge to scientifically meaningful solutions. This may not matter when the intention is to create a practical tool that accurately predicts outputs within the bounds of the data inputs, but we may be lead astray when pursuing discovery or meaning.



Credit: artwork by Kate Zvorykina, design by Michael Koldobskiy and Andrew Feinberg

When big data and mathematical modelling grounded on physical and biological knowledge can be judiciously combined, the power of raw information and the robustness and predictive strength of mathematical or statistical relationships can bring about the best of both worlds. Five papers in this issue of *Nature Biomedical Engineering* showcase how this can be achieved for a range of problems in neuroscience and oncology.

Understanding brain function involves finding patterns in the connectivity between regions of the brain that share a function or that are engaged in the same task at the same time. Such functional connectivity patterns have been biophysically modelled as neural circuits. Yet such models typically apply to population-level responses rather than to responses observed in a patient or individual, are disease-specific, and cannot be easily applied to control closed-loop electrical stimulation for treating neurological disorders (such as Parkinson's disease and epilepsy). Maryam Shanechi and colleagues report the development of linear input–output models to predict how multiregional brain networks respond to continuous stimulation, by using neural recordings from two awake (yet head-restrained) monkeys (which experienced stochastic changes in stimulation amplitude and frequency). Using cross-validation, the researchers fitted the parameters of the data-driven models to the datasets generated, and then evaluated their forward-prediction accuracy. By using numerical simulations, they also show that the fitted models could be used in closed-loop neuromodulation systems.

Amit Etkin and colleagues also used a data-driven approach, yet with purely statistical modelling (unsupervised sparse K-means clustering; that is, the clustering of observations into a K number of groups after filtering out uninformative features), to identify subtypes of post-traumatic stress disorder and of major depressive disorder on the basis of functional connectivity patterns by analysing data in four resting-state electroencephalography datasets that in aggregate included hundreds of patients with the two psychiatric disorders as well as healthy controls. They validated the

discovered neurophysiological subtypes with datasets of connectivity features from resting-state functional magnetic resonance imaging.

Finding genetic and epigenetic drivers of cancer is becoming increasingly feasible because of big data generated by array-based techniques. For example, analyses of DNA methylation have shown that some cancers are associated with specific patterns of epigenetic dysregulation. By performing whole-genome analysis of methylation stochasticity (using data from whole-genome bisulfite sequencing of patient samples covering a large fraction of CpG sites in the genome) for four subsets of paediatric acute lymphoblastic leukaemia, Andrew Feinberg, John Goutsias and colleagues have now found a relationship between methylation entropy and gene-expression variability. The analysis involved statistical-physics modelling, via a 'potential-energy landscape' of DNA-methylation patterns (pictured) following a probability distribution akin to that of a one-dimensional Ising model, of the probability that a specific methylation pattern is observed within a specific genomic region. They found that a set of genes involved in in-frame chromosomal translocations has methylation levels that carry information about the molecular subtype of leukaemia.

Biophysical modelling can also be leveraged to help patients. Franziska Michor, Eric Holland and colleagues show that a stochastic agent-based computational model of the dependence of cell location in the glioblastoma microenvironment (the perivascular niche) on the cells' sensitivity to radiotherapy and the concurrent administration of the chemotherapeutic temozolomide (the standard-of-care treatment for glioblastoma) can be used to optimize the treatment schedule so as to maximize survival. The model consists of a collection of autonomous 'agents' (in this case, cells) that interact with their immediate environment according to pre-specified rules, and was parametrized and validated using mouse data (yet can be matched to human data).

Predicting the time course of patients receiving checkpoint inhibitors, ideally before they start treatment, is currently limited by the absence of accurate biomarkers and clinical criteria. Zhihui Wang, Vittorio Cristini and colleagues developed a mathematical model that predicts tumour burden over time by using a set of differential equations describing the rates of tumour growth and of immune activation, tumour–immune-cell interactions, and the efficacy of immune-mediated cytotoxicity. They parametrized the model with published clinical data, and validated its performance in the stratification of patients according to long-term tumour burden with data from patients across four additional clinical trials.

Clearly, the pooling of models and data via data-driven modelling or model-driven data analysis can be leveraged for scientific discovery and to help solve practical problems. However, working out which types of data and model to use to best address specific problems leverages experience and perseverance; that is, it often draws on informed intuition more than on raw maths or computation. ❐

Published online: 16 April 2021
https://doi.org/10.1038/s41551-021-00721-0