The Upcoming Market for Structured Biological Data

@fvderop

November 2024

The best way to evaluate the quality of any information is to use that information to make a prediction. Inversely, if a model consistently makes accurate predictions, it can be assumed that it is based on high-quality information. Drugs and therapies are biology's most coveted predictions. After all, a drug is an intervention on an ill person's future - "if we administer you this drug, you will get better". Disease and death thus generate a large demand for accurate biological predictions, which can only be realised with quality information. Under normal circumstances, markets will attempt to satisfy large aggregate demands by increasing supply. This happens by re-allocating resources away from less efficient processes, and increasing production efficiency through technological advance.

Despite large amounts of resources allocated to pharma, and broad technological progress across the industry, the rate of new drug discoveries has not increased. In fact, the efficiency of drug approval has *decreased* both in time and per unit of money spent, with cost per drug doubling approximately every nine years. This phenomenon has been dubbed "Eroom's law", an antithetical reference to Moore's law[1]. When markets do not operate efficiently, we can examine how freely they are allowed to operate. The regulatory burden pharma and biotech operate under can be observed in how much time and money is spent on regulatory affairs: the mean duration of discovery and pre-clinical research is around 5 years[2], while the ensuing clinical trials take approximately 7 years[3], and the regulatory approval procedure adds another year. Satisfaction of regulatory oversight thus takes longer than the preceding research. Failure to pass the clinical stage is also common: the estimated industry-wide cost of failed oncology trials is $50–$60 billion annually. Exact calculations are difficult to obtain, but one study estimates that only around 7%[4] of the total cost of approved drugs is spent on non-clinical research!

Can regulatory pressure be the main driver in reducing efficiency? Let us compare pharma/biotech with finance, which also operates under regulatory

[1] Scannell et al., *Nat Rev Drug Discov*, 2012.

[2] Singh et al., *Front. Drug Discov.*, 2023.

[3] Martin et al., *Nat Rev Drug Discov*, 2017.

[4] Sertkaya et al., *JAMA Netw Open*, 2024.

scrutiny. Both industries attempt to make predictions in complex and stochastic systems. The financial industry has long shifted to a data driven approach. Real estate in Mahwah, New Jersey commands a premium due to proximity to the New York Stock Exchange's data center, giving traders a latency - and thus data/execution - advantage. Companies like Experian and Equifax sell in excess of $6 billion worth of consumer data annually. A subscription to a Bloomberg terminal provides real-time price data for stocks, bonds, commodities, currencies and derivatives, and costs around $25,000 per year - and this is only one of many providers. The finance industry is also the largest private consumer of satellite data, which is used to examine weather trends, crop growth, migration, energy consumption. All of these traded datasets are well-structured, and usually large in size. Structured data is thus a commodity that is valued by actors who wish to make accurate predictions. Moreover, it appears that this strategy is efficient since algorithmic traders have outpaced value investing[5].

If the same developments in algorithms and models are available to everyone, and incentives are just as great, then why has a data-driven strategy not penetrated as deeply in biotech and pharma as it has in finance? We could consider that biological models do not scale as well with data and compute as financial or language models have. Perhaps the matter is just that much more complicated. However, machine learning models have recently begun outperforming humans in a variety of medical tasks: LLMs have outperformed physicians at some diagnostic tasks[6], and computer vision models have outperformed pathologists at interpreting medical imaging[7]. More relevant to R&D, the field of bioinformatics has produced a number of new predictive tools. Alphafold allows for rapid prediction of protein structure and even molecular interactions based just on amino-acid sequence. High-dimensional sequencing data has been used to predict the relationship between genomes and gene expression[8]. We must seek the answer to our question in the type of data used, and perhaps more importantly, how it is generated. Whereas finance ingests structured data, biotech and pharma rely mostly on iterative experimental campaigns: small rounds of experiments are used to generate different data types, such as imaging, biochemical assays, and in recent decades, various -omics. These datasets are combined and interpreted by skilled and experienced professionals. If high-dimensional data is presented, it is usually first processed and dimension-reduced until palatable to humans. A new hypothesis (prediction) is proposed and tested, and this process iteratively continues until a hypothetical drug is found to be effective. Training a model on a large number of heterogeneous datasets is challenging. In addition, this step-wise approach demands flexibility of execution, which is why human operators are preferred. This in turn injects a number of confounding effects into the data: differences between operators or even the same operator at different times, time lag between experiments, biased reporting, ... all multiplicatively increase noise in the data. Experienced professionals are deployed

[5] McGowan, *Duke Law & Tech Rev*, 2010.

[6] Rutledge, *Learn Health Syst*, 2024.

[7] Bhave et al., *European Heart Journal*, 2024.

[8] Janssens et al., *Nature*, 2022.

to fill data gaps with intuition - which is why they are so highly valued - but this strategy does not scale.

We can observe that high levels of noise, custom reagents, and the challenge of describing protocols translate to low reproducibility in academic research[9]. Even in bioinformatics, an extreme minority of results can be reproduced from published code and data[10]. Given that academic talent flows towards biotech, we can expect that the industry has inherited some of these flaws. Lab automation could theoretically provide relief and help increase robustness, but small experiment scales make it difficult to justify the cost and lead time associated with it. Paradoxically, the small size of experiments is itself a consequence of widespread uncertainty. As a result, most lab automation capabilities are currently used for either manufacturing, diagnostics, or variants of high-throughput hit screening. We could consider doubling down and simply becoming better at automating low-n work, but I think this path is sub-optimal in the long run. Wide variety in protocols and techniques complicates low-n automation, which is why generalists such as Emerald Cloud Labs and Strateos have failed. I expect that specialists who focus on one single assay or service, such as Plasmidsaurus and Adaptyv Bio, will perform much better. Once these platforms reach critical mass, they will be able to benefit from economies of scale and have room to finetune their quality. Furthermore, they will be able to scale their services horizontally to fill large batch orders. This is the basic idea that we all need to understand: shifting to simpler experiments that scale well will allow us to generate larger data, which we know will improve model performance in the foreseeable future[11]. Scale has been able to pull this off, providing large datasets for generative AI, automotive industries and government. The demand for large, well-curated, structured datasets in industries where ML models have penetrated sufficiently deep is vast.

Silicon and energy have been commodified, but biological training data has not. The biotech industry should have the confidence to generate large (in the order of $10M) structured, high-dimensional datasets across a large number of technical and biological replicates with the explicit intent to feed this data into machine learning models. Such large datasets could even be "rented" out to customers who train models on them without ever seeing the data by using homomorphic encryption. When it is proven that this mode of data collection will lead to better models, and predictive power grows, this will also attract talent from other industries. Workers prefer to spend their time efficiently, and noisy data and inefficient collection act as a negative filter on this natural pressure. Remove this filter, and the system will self-reinforce. This evolution has already been set in motion, and I am hopeful that we will see a new generation of model-derived medicine and therapies enter clinical trials before the end of the decade.

[9] Rodgers et al., *eLife*, 2021.

[10] Ziemann et al., *Briefings in Bioinformatics*, 2023.

[11] Kaplan et al., *arXiv preprint arXiv:2001.08361*, 2020.