ECLIPSEBIO
THE RNA SUCCESS COMPANY

Whitepaper

Top 10 data problems for RNA therapeutic development

Introduction

Data problems are one of the quietest bottlenecks for successful RNA therapeutic development right now. The science is moving fast, but the data ecosystem is messy, fragmented, and often misleading. In this whitepaper, we review the top 10 data problems that the RNA therapeutic space currently faces.

Top 10 data problems in the RNA space

1. **Scarcity of relevant training data**
2. **Lack of high-throughput, multimodal data generation assays**
3. **Disconnect between design and effect space**
4. **Lack of chemically modified data**
5. **Translatability**
6. **Context-dependence**
7. **Proprietary data silos**
8. **Data fragmentation**
9. **Lack of standardized, relevant benchmarks**
10. **Immunogenicity data gap**

1. Scarcity of relevant training data

The mapping from mRNA sequence to therapeutic function is dramatically less well-characterized than for proteins. Protein language models can train on millions of sequences with solved structures, but RNA therapeutics lack equivalent large-scale, high-quality labeled datasets. Data relevance extends beyond structure. Therapeutically relevant outcomes are diverse, and data generation is expensive and time-consuming.

2. Lack of high-throughput, multimodal data generation assays

Generating labeled RNA data is expensive and time-consuming, which is why the field is moving toward massively parallel reporter assays (MPRAs). MPRAs enable simultaneous measurement of thousands of sequence variants in a single experiment, dramatically increasing throughput. Sequencing-based characterization approaches add another dimension—rather than single-point readouts, they capture richer information about expression kinetics, ribosome occupancy, and degradation patterns. These methods are essential for building datasets at the scale that machine learning requires.

3. Disconnect between design and effect space

Computational metrics like CAI, MFE, and GC content are commonly used as evaluation benchmarks for RNA models, but these are proxies, not outcomes. Optimizing for what you can compute is not the same as optimizing for therapeutic performance. The biological effect space (expression, stability, immunogenicity in vivo) follows rules we don't fully understand, and these proxy metrics capture only a fraction of what matters. This is a call for the field to develop more predictive metrics and steer data generation toward therapeutically relevant outcomes.

4. Lack of chemically modified data

Most therapeutics use N1-methylpseudouridine, but most datasets are endogenous and treat RNA as unmodified. This has significant impacts on structure, translation, stability, and immunogenicity.

5. Translatability

Models trained on in vitro data fail to predict in vivo performance. Potential solutions include more relevant models (organoids, patient-derived tissues) or approaches designed for data-poor environments like active learning.

6. Context-dependence

mRNA performance varies dramatically based on delivery vehicle and composition, target cell type, tissue microenvironment, dosing route, and more. Most exogenous datasets are generated under specific conditions that aren't captured and don't generalize.

7. Proprietary data silos

Most high-quality RNA therapeutic data sit inside pharma and biotech companies and never get published. Academic datasets are small and use inconsistent experimental protocols. This fragmentation makes it difficult to build generalizable models or compare methods across the field.

8. Data fragmentation

mRNA is a multi-component system — 5' UTR, coding sequence, 3' UTR, poly(A) tail — but most datasets optimize one element while holding others constant. We lack systematic data on how these regions interact. A UTR that performs well with one coding sequence may fail with another, yet current data rarely captures these dependencies.

9. Lack of standardized, relevant benchmarks

RNA benchmarks exist, such as BEACON and RNAGym, but they measure proxies like structure prediction accuracy rather than therapeutic outcomes. If the field optimizes for the wrong metrics, this leads to misguided designs and poor outcomes. The emphasis needs to shift from "can we predict structure" to "can we predict expression, stability, and immunogenicity in relevant contexts."

10. Immunogenicity data gap

Immunogenicity is a critical quality attribute, but current detection methods are inadequate. Existing dsRNA detection and quantification methods, such as gel electrophoresis, ELISA, or homogeneous time-resolved fluorescence (HTRF), have low sensitivity or are time-consuming. Newer approaches like biolayer interferometry (BLI) using FHV B2 protein and lateral flow strip assays are improving detection, but the structural heterogeneity of dsRNA (including length and structure) and its precise immunomodulatory mechanisms affecting vaccine safety are poorly understood. Better immunogenicity readouts are essential for closing the loop between sequence design and safety.

Addressing these problems with Eclipsebio

Although these data problems currently act as limitations for therapeutic programs, Eclipsebio and others are working to alleviate them through assay development and database generation. For example, our eVERSE database contains thousands of data points on different dimensions of RNA biology, including secondary structure and miRNA binding to assist in model training, while our eSENSE dsRNA profiles dsRNA species to solve the immunogenicity data gap. Contact us today to learn how Eclipsebio can support your data needs for successful RNA drug development.