Before the Next PDB: Can a Field Converge Before It Builds?

By Peter K. Koo
Cold Spring Harbor Laboratory
April 2026

Funders want to invest in AI for science. They do not yet know where.

Over the past two years I have sat in several convenings on AI-readiness for science, including ones organized by Schmidt Sciences and the UK Department for Science, Innovation and Technology. Each had roughly the same brief: identify the next foundational dataset, the one that would do for its field what the Protein Data Bank did for protein structure.

The meetings are well-intentioned. The people in them are the right people. The money is real and waiting. And the meetings end the same way. The discussion sprawls. Everyone agrees the opportunity is real. No one agrees on what to build. The funders leave without a target, and the capital stays on the sidelines or gets scattered across small bets that do not add up to a foundational resource.

Regulatory genomics, my own field, illustrates the gap. Models are increasingly able to absorb systematic perturbation and functional genomics data at scale, but we have no agreed mechanism for deciding which perturbations, in which cellular contexts, and at what depth would produce the most generally useful public resource.

The Protein Data Bank is the standard reference for what a foundational dataset looks like. The PDB was not designed to train AlphaFold. It was built slowly, over decades, because crystallographers deposited structures they had already solved, and because journals and funders eventually turned deposition into a requirement. It captured static structure. It did not capture the dynamics, complexes, and functions that protein biologists might have argued for if asked to design the perfect resource. Its power came from being a thin, well-specified slice that later proved to be the right slice. If you asked protein biologists today to design the PDB's successor, you would not get the PDB. You would get dynamics, complexes, and function all at once, and almost certainly, nothing would get built.

The foundational dataset is the one a field is least equipped to design by committee, and the committee is what we keep convening.

The simplest way to see why is to ask where, in the current scientific ecosystem, the convergence step actually lives. By convergence step I mean the moment when a field looks across the datasets it could build and chooses one, with a metric and a plan. Agencies fund proposals written by individual investigators, and even their large consortium programs typically execute a target rather than choose among candidates. Philanthropies can convene around the problem, but the output is usually a roadmap that lists everything a field wants rather than a commitment to what it needs most. Industry funds what it can capture, which excludes precompetitive foundational data. The convergence step is structurally orphaned.

The same gap appears at the institutional level. Focused Research Organizations and coordinated philanthropic programs have shown that dedicated teams can turn ambitious public-goods ideas into executable programs. The function that remains underdeveloped sits one step earlier: helping a field choose among competing targets under explicit metrics and pilot decision thresholds.

My hypothesis is that some fields fail to produce foundational public resources because no institution is responsible for helping them choose before large-scale capital is deployed. Money is available. Technical capacity exists. What's missing is a staged decision process that forces a field to narrow before it builds.

The experiment has three stages: proposal selection, constrained convening, and scaling pilot. Proposal selection tests whether a field can define a resource question clearly enough to justify a convening. The constrained convening tests whether the field can turn that question into a ranked portfolio of candidate datasets, each with a metric, scaling hypothesis, pilot design, and decision threshold. The scaling pilot then tests whether any candidate improves the agreed model metric as more data are generated. The purpose is not to choose the final full-scale resource at the convening, but to turn a diffuse wish list into a small, testable portfolio with explicit decision rules.

First, funders issue a call for short proposals. Each proposal names the resource question, the needed participants, and why the field is ready to narrow. A review panel selects two or three convenings to fund. This selection step is itself informative: if a subfield cannot define a compelling resource question on paper, it is unlikely to converge in a room.

Second, the funded convenings run under fixed constraints. Each must produce a ranked list of three to five candidate datasets instead of a general roadmap. For each candidate, the group must specify a metric, a scaling hypothesis explaining why more data should improve that metric, a falsifiable pilot, and a decision threshold for advancing. An independent review group then judges whether the ranking is defensible, the scaling hypothesis testable, and the pilot executable. Success means a ranked portfolio specific enough to fund, test, or reject. A satisfied room and another report would not count.

To make this concrete in regulatory genomics, the goal would be to identify which kinds of perturbation data would most improve models of how DNA sequence controls gene expression. A passing portfolio might compare three plausible public resources: dense saturation mutagenesis of cis-regulatory elements, systematic CRISPRi perturbation of candidate enhancers with paired expression readout, and combinatorial transcription factor perturbations to probe regulatory logic. Each targets a different layer of regulatory control, and each would have to state its cellular context: a single well-characterized line, a small panel, or a broader atlas. The group might use held-out perturbation prediction as the agreed metric, because it tests whether models can predict the effects of unseen interventions in regulatory DNA rather than merely interpolate among measurements already collected. Each candidate would need a numerical scaling hypothesis, a pilot that could falsify it within months, and a decision rule naming the improvement threshold required to advance. A failing portfolio would simply name attractive datasets without metrics, invoke scaling without numbers, or propose pilots whose results would not change anyone's mind.

Any convening that produces a target must also produce an attribution protocol. The PDB worked partly because depositing a solved structure was cheap, and because journals and funders eventually made deposition routine. Foundational AI data has neither property. A single perturbation experiment may not be publishable on its own. Its value appears only when many such measurements are combined into a resource a model can learn from, while a reviewer may not see any single contributor's piece as a major scientific output. A field has not converged on a buildable resource unless it can say how contributors will be credited through downstream model use. That means persistent contributor identifiers, required citation in model cards and downstream papers, and review policies that recognize infrastructure output.

The same asymmetry shapes who must participate. A funded convening must seat both the experimentalists who would generate the data and the AI researchers who would model it. Experimentalists carry the cost of generation and the career risk of producing infrastructure rather than hypothesis-driven papers. AI researchers benefit from the data with little marginal cost. If that asymmetry is ignored, the dataset will be specified by the people least responsible for building it.

The goal is a defensible choice the field can stand behind. Convenings should be screened for scope before they are funded, favoring domains with tractable questions and credible participation from both data generators and model builders. A group asked to decide too much will not decide anything; a group asked to decide too little will define a resource too narrow to matter. If a workshop cannot meet the prespecified bar, it should not advance. That failure is useful. It tells funders that the field may not be ready, the scope may have been mis-set, the

community may be split on what to build, or the credit structure may not yet be workable. Each is worth learning before committing the kind of capital these resources require.

Third, portfolios that pass review compete for modest support to run scaling pilots. Once a field has narrowed to serious candidates, the next question is whether any of them are worth scaling. A pilot would measure how the agreed metric improves as more data are generated, estimate the cost of reaching target performance, and detect whether the proposed data type shows early signs of saturating too quickly to justify a larger build. The scaling curve is the decision. The point is to replace a nine-figure guess with a measured answer about whether scaling is warranted.

Such pilots would not be cheap in absolute terms, but they would be cheap relative to full-scale infrastructure. In regulatory genomics, a serious scaling pilot might cost hundreds of thousands to a few million dollars, depending on the assay, replication, number of cellular contexts, and sequencing depth. That is the right order of magnitude for deciding whether a data type has a plausible scaling curve before committing tens or hundreds of millions of dollars to build it.

If a pilot succeeds, full-scale execution should be coordinated without becoming monolithic. Multiple teams could be funded toward the same agreed target, with common metrics, shared interfaces, explicit non-overlap in work packages, and public release under the attribution protocol. Competition should be real, but it should occur inside a shared specification. The program should continue tracking the scaling curve as the build proceeds. If improvement flattens, or if another data type begins to produce a steeper gain, the plan should adapt rather than continue by inertia. Without that shared specification and ongoing measurement, the default is what we already have: adjacent partial datasets that cannot be combined into a single resource a model can learn from.

Regulatory genomics is only one instance of a broader structural pattern. The same issue appears across protein complexes, cell state under systematic perturbation, and systems neuroscience. Protein biology faces choices among cryo-EM atlases, interactome maps, and complex dynamics; cell-state biology among perturbation atlases and spatial readouts; systems neuroscience among whole-brain recording, connectomics, and behavioral data. Each field can name plausible resources, but none has a clear process for choosing among them before large-scale capital is committed. The experiment reveals whether a field is ready to narrow before tens or hundreds of millions are spent building the wrong thing.

The timing is right because we can finally test scaling behavior before committing to full-scale infrastructure. Models are now strong enough to make scaling curves informative, and assay pipelines scale far better than they did. A modest pilot can therefore measure whether a proposed dataset improves the agreed metric and at what cost, and whether returns saturate quickly enough to make a larger build unwise. This was much harder when models were weaker, assays were less scalable, and the value of additional data could not be measured as directly.

The funding landscape has also shifted. Renaissance Philanthropy's AI for Science Datasets RFP supports dataset concept development and early validation rather than full build-out, while the Chan Zuckerberg Biohub's $500M Virtual Biology Initiative shows that major capital is moving toward open datasets for predictive models of the cell. These are serious efforts, and they make the narrowing problem more urgent: fields need a way to decide which specific resources should be scaled, under what metrics, and with what shared specification.

The next foundational AI-for-science datasets are unlikely to emerge as modest archives assembled incidentally over decades. They are deliberate infrastructure projects that can require nine-figure budgets. Before building at that scale, funders should test whether a field can narrow, whether the proposed data type improves model performance as it scales, and whether the community will hold to a shared specification once the money is on the table.

We will probably not name the next PDB from a conference table. The experiment worth running is whether a field can choose under constraint before funders default to the loudest or most familiar ambition.