Prediction Before Interpretation: A Criterion for Field-Like Constructs in Complex Systems

Flyxion
Independent Researcher

Abstract

This essay proposes a criterion by which field-like theoretical constructs—terms such as *capacity*, *agency*, or *resilience* that grammatically behave as measurable quantities—may be judged to have earned empirical, rather than merely rhetorical, standing. The criterion is stated independently of any particular framework and is then applied, as a test case, to a capacity–transport–dissipation decomposition proposed for active Brownian particle systems near the onset of motility-induced phase separation (MIPS). Both possible outcomes of that empirical test—success and failure—are interpreted here, in advance of any data, in order to prevent their meaning from being negotiated after the fact. The essay does not defend the correctness of the decomposition under study. It defends the standard by which such a decomposition should be judged.

1 Introduction: Vocabulary and Measurement

A certain kind of framework has a recognizable signature. It introduces a term—*capacity*, *agency*, *obligation*, *utility*, *resilience*, *alignment*—that sounds like it refers to something with a value: a quantity that could in principle be larger here, smaller there, changing over time. The word is doing the grammatical work of a measurable thing. But no procedure accompanies it. Nothing tells you how two people, looking at the same system, would independently arrive at the same number, or even agree on which direction it moved.

This is not a complaint about vagueness as such. Each of these words has a real technical home somewhere—utility in decision theory, capacity in information theory and thermodynamics, agency in control theory, resilience in ecology. The trouble is a second, looser use that borrows the authority of the technical sense while discarding its operational content. A paper can define a quantity with full mathematical formality—an integral, a functional, a coupling term—and still be making this move, if the definition

was never subjected to the question that gives definitions their bite: what observation would distinguish a system that has more of this from one that has less?

That is the diagnostic. Take any candidate construct and ask: if two competent observers examined the same system and disagreed about its value, is there a procedure that could in principle settle the disagreement? If yes, the construct has entered the domain of empirical claims, whatever its philosophical pedigree. If no, it is still doing rhetorical work—organizing intuition, suggesting analogies, lending vocabulary—but it has not yet done empirical work, no matter how precisely it is written down.

This is not an appeal to naive verificationism. Constructs are allowed to precede their measurement procedures. Atomic theory existed before anyone could observe an atom; econometrics routinely reasons about confounders no one has measured. The claim here is narrower and comes with a time index: a construct earns *predictive* standing only once a procedure for recovering it from data exists and has been tested against alternatives. Before that point, it is a hypothesis about what might turn out to be measurable—worth stating, worth exploring, but not yet a result. The failure mode is not proposing such constructs. It is treating the proposal as though it were already the result.

Stated as a slogan, because the rest of this essay is going to lean on it: new terminology is not new predictive content. A framework can rename existing observables in more evocative language without adding anything an outside skeptic should update on. The only way to tell the difference between renaming and discovery is to make the new term earn its keep the same way any other empirical quantity does.

What follows is an attempt to state that criterion precisely, and then to apply it—deliberately, in public, before the outcome is known—to a single construct in a single well-studied system. Not because the construct is expected to survive. Because subjecting it to the test is the only way its survival, or its failure, would mean anything.

2 An Identification Criterion

The diagnostic in Section 1 asks whether a construct is, in principle, adjudicable—whether two observers who disagreed about its value could be brought to agreement by some procedure. That question is a floor, not a ceiling. It rules out the clearest cases of vocabulary doing rhetorical work, but it does not by itself tell a researcher what to do. What follows is an attempt to convert that diagnostic into something usable: a criterion specific enough that a proposed field-like construct could actually be tested against it, and could actually fail.

A proposed field earns predictive standing only if it satisfies four conditions jointly. First, it must be reconstructible: there must exist a procedure that recovers the quantity from observational data, stated with enough precision that another researcher could apply it independently and obtain the same values. Second, where the underlying intuitive concept admits more than one reasonable formalization, the competing formalizations must be defined and compared against one another rather than one being adopted by fiat. Third, the coefficients or parameters identified for the construct must remain stable, or transform in a predictable and principled way, across variation in what an out-

side observer would regard as a nuisance parameter—coarse-graining scale, sampling resolution, choice of estimator. Fourth, the construct must improve prediction on data withheld from whatever process fit it, relative to the strongest baseline already available, rather than merely fitting the data it was calibrated on.

None of these four conditions is sufficient on its own, and the failures they each guard against are worth naming, because each is a real and common failure mode in its own right. Reconstructibility alone guards against nothing: almost any quantity can be smoothed, kernelled, or averaged into a spatial field, whether or not the field means anything. A perfectly well-defined reconstruction procedure can still produce noise dressed as structure. Comparability alone guards against nothing either, if it stops at comparison: a framework that offers three competing definitions of a concept and never adjudicates between them has not made the concept more rigorous, only more elastic, since any observed pattern can be attributed to whichever variant fits it after the fact. Stability alone is cheapest of all to satisfy and least informative—a constant is perfectly stable and utterly uninformative, and stability across coarse-graining scale is a property many uninteresting quantities share. Predictive improvement alone, isolated from the other three, is the condition most vulnerable to overfitting: a construct fit and evaluated without an independent reconstruction procedure or a principled account of its own stability can appear to predict well simply by having enough free parameters to track the specific dataset it was built on. The criterion does its work only as a conjunction. A construct that satisfies three of the four clauses and fails the fourth has not partially succeeded; it has failed a different, specific way, and the failure is worth naming precisely rather than rounding up to "promising." The graded categories introduced later, in Figure 1, classify different forms of empirical success and failure only after this threshold has already been crossed; they do not relax the requirement that all four clauses be satisfied before a construct is judged to have earned predictive standing at all.

The connection back to Section 1 is direct. The two-observer test asked whether independent agreement was possible in principle. Clauses one and two are what make that agreement possible in practice: reconstructibility supplies the procedure two observers could each run, and comparability prevents the construct from being defined so loosely that agreement is available only by stipulation. Clauses three and four go further than Section 1's diagnostic strictly required, because adjudicability is a necessary condition for empirical meaning but not, on its own, a very demanding one—plenty of adjudicable quantities are still useless. Prediction is the stronger and more useful standard, and it is the standard this project is actually prepared to be judged by.

It is worth being explicit about what this criterion is not claiming, since the natural objection is that it smuggles in a narrow instrumentalism—that nothing counts unless it predicts. That is not the claim. Explanatory unification, conceptual clarification, and mathematical elegance remain legitimate reasons to develop a theoretical framework, and frameworks have historically earned their place in a field's vocabulary for reasons well short of predictive superiority over an existing model. The claim here is narrower and applies to a specific rhetorical situation: when a framework introduces field-like vocabulary—a term that behaves grammatically like a measurable quantity with location and magnitude—that vocabulary is not yet entitled to be believed as a description

of the world until something like these four conditions has been checked. Elegance and clarification are reasons to keep developing an idea. They are not yet reasons to think the idea has found something.

Because this criterion is stated here, before any system has been named, Section 3 can introduce active matter as a deliberate application of an independently motivated standard, rather than as a test case selected and then justified backward once its outcome was already known.

3 Active Matter as a Test Case

A criterion is only as convincing as the test it is willing to survive. Before naming a system, it is worth stating what would make a system a fair one—otherwise the choice of test case risks looking, in retrospect, like it was picked because it was likely to cooperate.

A good test case for the criterion in Section 2 needs three properties. The underlying process must be directly observable, not inferred through a proxy that could itself be the source of any apparent signal—trajectories, not summaries of trajectories. The phenomenon under study must have a genuine transition or event to predict, defined by criteria independent of whatever new construct is being tested, so that "improves prediction" has something concrete and pre-existing to attach to. And strong, well-validated competing models must already exist, so that outperforming them is a real achievement rather than the trivial feat of explaining something nobody had previously tried to explain.

Motility-induced phase separation in active Brownian particle systems satisfies all three, and it is worth being explicit about why. The trajectories are exact: simulated ABP systems yield the position and orientation of every particle at every timestep, with no measurement noise and no missing data to blame ambiguity on—a rare condition, and one that removes an entire category of excuse for a null result. The transition is genuine and multiply characterized: MIPS onset shows up independently in density bimodality, in the growth of the low-wavenumber structure factor, and in the fraction of particles belonging to the largest connected cluster, which means a target variable can be defined without reference to any of the candidate capacity fields—the label does not know what it is being used to test. And the baselines are not weak: kinetic and hydrodynamic theories of active matter already give a reasonably complete account of MIPS without invoking anything resembling a capacity–dissipation vocabulary. This is not an empty explanatory space waiting to be filled by whatever framework arrives first.

That last point is worth dwelling on, because it is easy to mistake for a weakness of the test case rather than its central strength. If MIPS were poorly understood, almost any new decomposition would look like it was adding something, simply because so little had been tried. Because MIPS is already well explained, this is a hard test rather than a soft one. A new construct that adds nothing here has genuinely added nothing—there is no gap in existing theory for it to be quietly filling by default. A new construct that does add something here has added something real, against a standard already strong enough to make the addition meaningful. This is the setting in which Section 1's

slogan—new terminology is not new predictive content—is hardest to dodge, which is exactly why it is the right setting to test it in.

The competing capacity hypotheses can now be stated concretely, as the direct instance of Section 2's comparability clause rather than as a single definition asserted by fiat. The imposed variant treats capacity as the propulsion force built into each particle—a quantity that, for a constant-speed ABP model, is nearly uniform in space and therefore carries little informative gradient on its own; it functions mainly as a control field, a way of confirming the reconstruction pipeline behaves sensibly on a field expected to be nearly featureless. The signed realized variant treats capacity as the propulsion actually converted into displacement—active work, positive or negative, which can go negative when collisions push a particle backward against its own orientation. The usable-forward variant restricts this to its non-negative part—realized propulsion in the forward direction only, discarding the sign. None of the three is presented as correct. The point of naming three, rather than one, is that a single hand-picked definition would look exactly like the failure mode Section 1 describes—vocabulary asserted rather than earned—while three competing candidates, tested against the same criterion and the same data, is what taking that criterion seriously actually looks like in practice.

The falsifiable form of the hypothesis, stated precisely, is this: a capacity–transport–dissipation decomposition, reconstructed from ABP trajectories by the procedures this project specifies, contains identifiable dynamical information about impending MIPS that is not reducible to contemporaneous density structure. It is worth stating with equal precision what this excludes. It is not the claim that this decomposition is the correct description of active matter, nor that the process-first framework it is drawn from is thereby validated. It is a narrower and more modest claim about incremental information content—whether this particular way of cutting up the system's dynamics tells you something about the future that the system's present density configuration alone does not.

What remains—the reconstruction procedures, the sparse-regression identification, the prospective-prediction protocol, the intervention design—belongs to the empirical paper, not to this essay. This essay's task ends here, at the point of stating the test cleanly enough that running it is the only thing left to do.

4 What Success Would Mean

It is worth fixing, before any identification study is run, exactly what a positive result would license and what it would not. The temptation to let a successful outcome expand its own meaning after the fact is real, and the only defense against it is to write the boundary down now, while nothing is at stake.

Success, precisely, means this: at least one of the competing capacity reconstructions—the imposed variant, the signed realized variant, the usable-forward variant—survives every clause of the criterion stated in Section 2. It reconstructs stably from trajectory data. Its coefficients remain stable, or transform predictably, as the coarse-graining scale changes. And the resulting models improve held-out prediction of MIPS onset beyond

the strongest available baseline of conventional structural observables.

What does not follow from that result is worth stating with equal precision, because a successful test of one construct is easily mistaken for a broader vindication it did not earn. Success would not validate the process-first framework this construct is drawn from—it would show that one operationalization, in one system, cleared one bar. Success would not mean existing kinetic or hydrodynamic theories of active matter were wrong, or even incomplete in any general sense—only that they were missing one piece of predictive information, in one regime, that a different decomposition happened to supply. And success would not license extrapolating the result to other systems—ecological, economic, cognitive—without independently rerunning the same identification procedure there. What would transfer is the method's demonstrated worth, not its conclusion.

There is also a distinction worth drawing between two kinds of success, because they warrant different language in whatever paper eventually reports them. A result can be observationally strong—predictive improvement that holds across a full parameter sweep—without being mechanistically established, which requires surviving the intervention protocol: perturbing activity or dissipation directly and confirming the transition probability changes in the predicted direction while density is held fixed. A construct that predicts well but has not been tested under intervention has earned the description *early-warning variable*. It has not yet earned the description *driver*. Collapsing that distinction is one of the more common ways a good empirical result gets oversold in its own write-up.

One further caution belongs here, close in spirit to the caution that will govern Section 5. Success would establish the empirical usefulness of a particular operationalization of capacity, not the necessity of that operationalization. If the signed realized variant survives the test and the other two do not, that is evidence this reconstruction procedure works—not evidence that "capacity" *means* signed realized power, as though the word had a hidden correct definition the experiment had merely uncovered. The criterion in Section 2 does not require all three variants to succeed for the exercise to be worthwhile; it requires only that the intuitive concept earn its keep through at least one procedure that beats the null. A single successful reconstruction is a complete result on its own terms.

Figure 1. Pre-Registered Interpretation Scheme

Outcome	Interpretation	Permitted Claim
Strong success	Predictive improvement holds across the parameter sweep and survives intervention	Evidence for a mechanistically relevant capacity reconstruction in this system
Weak success	Predictive improvement holds, but intervention support absent or untested	Early-warning variable; predictive but not established as driver
Conditional success	Improvement confined to specific regimes	Local predictive utility; domain-limited result
Failure	No predictive improvement beyond baseline	Capacity reconstruction not empirically justified in this system
Clean failure	Reconstruction unstable or non-identifiable across coarse-graining scale	Construct fails before predictive testing is possible
Indeterminate	Insufficient statistical power to distinguish among the rows above	No claim permitted pending further data

This table is not itself a finding. It is fixed here, before the identification study exists, so that whichever row the eventual result falls into, the language used to describe it will have been chosen before anyone had a reason to prefer one description over another.

5 What Failure Would Mean

Failure deserves the same precision as success, if not more, because failure is where the temptation to redescribe a result after the fact is strongest.

Failure, precisely, means this: the capacity–transport–dissipation decomposition, reconstructed by the procedures this project actually attempts, does not improve held-out prediction of MIPS onset beyond the strongest conventional baseline, or its coefficients fail to remain stable across coarse-graining scale, or fail to transfer from one region of parameter space to another.

Several conclusions do not follow, and it is worth listing them so they cannot be smuggled in later under cover of disappointment. Failure would not vindicate existing active-matter theory at this construct's expense—the kinetic and hydrodynamic accounts serving as baselines were never under test; they were the standard this construct was asked to beat, and beating them was never guaranteed to be possible for any candidate. Failure would not mean capacity, dissipation, or the broader vocabulary this project draws on are meaningless in general—only that this operationalization, in this

system, did not clear the bar set in Section 2. A construct can fail one empirical test and remain reasonable elsewhere, provided the failure is treated as informative rather than argued away. And failure would not discredit looser, qualitative uses of the same vocabulary in other contexts, since those uses were never staked on this outcome—that separation is the entire argument of Section 1.

What failure does establish is narrower and, for that reason, more useful: that this specific reconstruction—this capacity variant, this kernel family, this candidate library—failed to earn predictive standing under a criterion stated in advance. That is a scoped, technical finding about one attempt. It is not a verdict on an idea.

The distinction between clean and dirty failure, named in Figure 1, matters because the two cases point toward different next steps rather than toward the same conclusion. A clean failure—stable fields, stable coefficients, no predictive gain—says the conventional observables already contained what mattered, and the new decomposition, however elegant, added nothing a predictor could use. That is a real answer. A dirty failure—coefficients that will not settle, that drift with coarse-graining scale or refuse to transfer between runs—says something narrower: that no clean dynamical signal may be recoverable at the resolution attempted, a statement about the limits of the identification procedure at least as much as about the construct it was identifying. Mistaking a dirty failure for a clean one would understate what remains possible; mistaking a clean one for dirty would overstate it.

Symmetry with Section 4 requires one final point. Just as success would establish the usefulness of one operationalization rather than the correctness of a definition, failure establishes the inadequacy of one operationalization rather than the falsity of an intuition. The word "capacity" is not on trial here. One procedure for measuring it is.

6 Conclusion: The Limits of the Present Argument

The arc of this essay can be stated in a single paragraph. Section 1 named a pathology: vocabulary that grammatically behaves like a measurable quantity while carrying no procedure for measuring it. Section 2 converted the diagnostic this pathology suggests into a checklist specific enough to fail. Section 3 selected a test case chosen for its difficulty rather than its hospitality, and stated a single falsifiable hypothesis in its terms. Sections 4 and 5, together with the interpretation scheme standing between them, fixed the meaning of both possible outcomes before either exists. Nothing in this argument has depended on whether the capacity–transport–dissipation decomposition under discussion is, in the end, correct.

It is worth stating plainly what this essay has not done. It has not run the identification study described in Section 3. It has not reconstructed a single field from data. It takes no position on which row of Figure 1 the eventual result will occupy, and it should not be read as hedging toward one. Its authority, such as it is, comes from having no stake yet in which way the result goes—and that condition is worth naming directly rather than leaving implicit, since it will not hold once data exist.

The criterion in Section 2 was built for field-like constructs in physical systems with

directly observable trajectories. It is not obviously portable, unmodified, to domains where observation is a harder notion—ethics, economics, or any setting in which the quantity of interest is partly constituted by the act of measuring or reporting it. Extending the criterion to such domains would require its own argument, one this essay does not make. Claiming otherwise here would repeat, at the level of the essay itself, exactly the failure mode Section 1 diagnosed in others: vocabulary reaching past what it has actually earned.

A criterion that no proposed construct could survive would be useless; a criterion every construct survives would be equally useless. The purpose of the standard proposed here is that both outcomes remain genuinely possible, and that which one obtains is not yet known.

A separate document, *Capacity, Dissipation, and the Onset of Motility-Induced Phase Separation*, will develop the experimental framework associated with the criterion proposed here. Its central hypothesis is the sentence given in Section 3, but its purpose will not be to defend the criterion itself. Rather, it will specify the measurements, analyses, and computational procedures through which that hypothesis may be evaluated.

Experimental results are not part of the present essay. They will be reported separately once the analyses have been completed and the outcomes assessed.

The three components address different questions. This essay asks why a standard of this kind should be demanded at all. The companion document will ask how that standard can be operationalized and tested. The eventual results paper will ask what the experiments show. The last thing worth saying is the plainest: it is the criterion, not the construct, that this essay is asking the reader to accept.