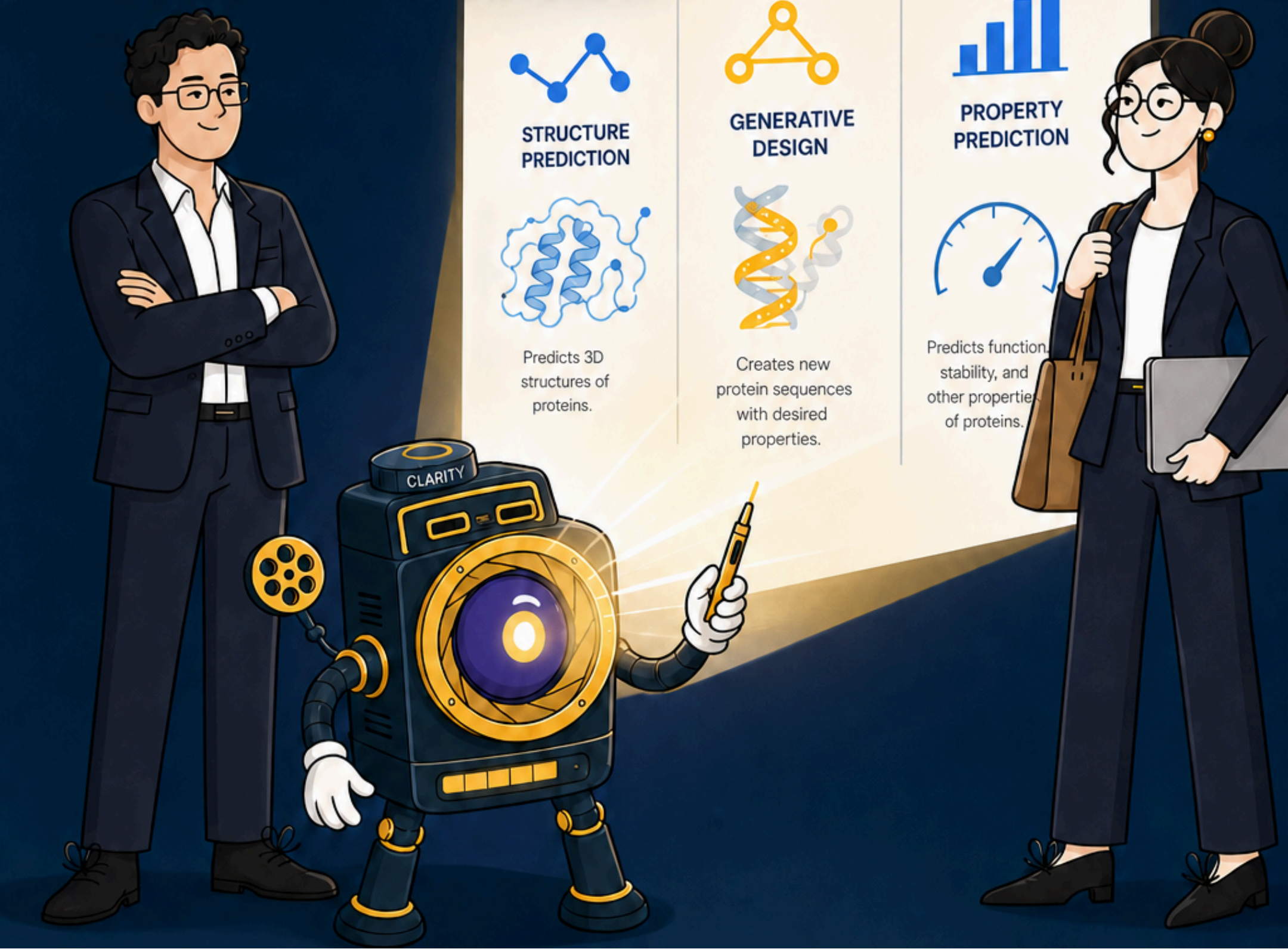
Three AI Models Reshaping Protein Design:

A Plain-English Investor Guide

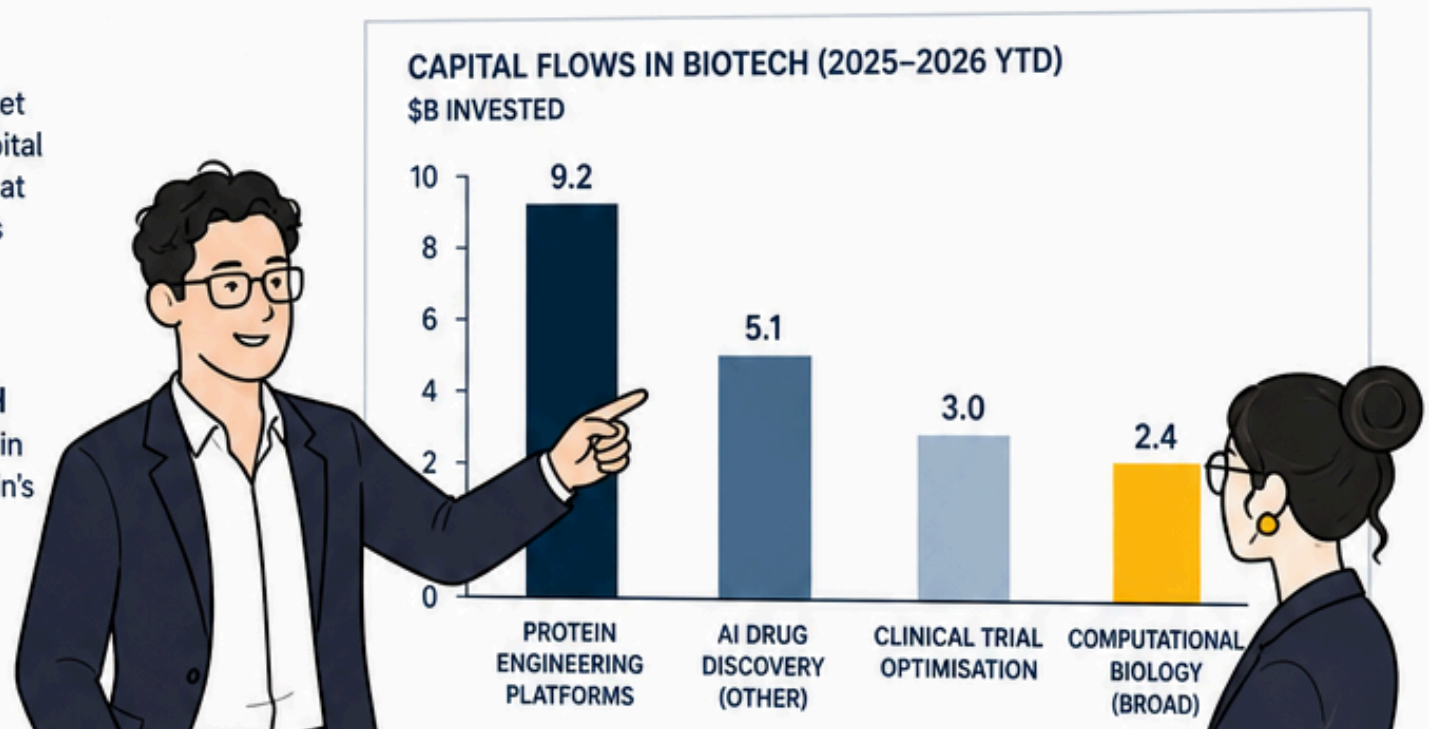


Why protein design matters to your portfolio right now

MARKET GROWTH
The AI protein design market is growing rapidly, with capital flowing into applications that deliver engineered proteins with specified functions.

SPEED BREAKTHROUGH
AI reduces the time to obtain a working model of a protein's 3D structure from months or years to hours or days.

UNDERSTAND THE AI
There are three distinct architectures—and they're not interchangeable.



Institutional capital is moving into commercial platforms that produce engineered proteins with specified functions.

Source: PitchBook, CB Insights, company reports (2025–2026 YTD)

Model type 1: Protein language models (PLMs)

A protein language model is conceptually the most intuitive of the three, because it's closely related to the large language models you already know. Instead of training on words and sentences, a PLM is **trained on protein sequences**: the **long chains of amino acids**, represented as letters, that encode every protein in the known biological world.

The training objective is essentially the same as it is for a text model. Given a partial sequence, **predict what amino acids** are most likely to appear next, or fill in a masked position. Do that billions of times across millions of real protein sequences from across evolutionary history, and the model internalises something remarkable: the **deep statistical grammar** of what makes a protein a protein.

- Which residue combinations are **physically stable**.
- Which patterns appear in proteins that **bind to specific targets**.
- Which sequences **fold reliably** versus which ones collapse into **useless tangles**.

Profluent achieved a notable landmark in designing OpenCRISPR-1, the **first AI-generated CRISPR system** demonstrated to be functional in human cells. That milestone matters for investors: it's a demonstration that a model trained on biological sequences can output something that **works in biological reality**, not just in computational space.

The broader field — including earlier work from groups such as Salesforce Research on ProGen — has progressively demonstrated that language models trained on protein sequences can **generate novel, functional proteins**.

Protein language models (PLMs): What they're commercially good for

PLMs excel at two specific commercial tasks.

1 Directed evolution acceleration

Traditional directed evolution screens thousands of random mutants to find a variant with better stability, activity, or specificity. A PLM can score which mutations are likely to improve function, dramatically shrinking the search space before any wet-lab work begins. Companies like Cradle Bio apply this approach to help pharmaceutical clients improve the performance of biologics and enzymes.

Thousands of random mutants → PLM scores & ranks → Top candidates for testing

2 Sequence-space exploration

Because a PLM has learned what "looks like a real protein" from millions of examples, it can generate novel sequences that have no natural counterpart but are statistically likely to fold correctly and perform a specified function. EvolutionaryScale, for instance, is developing large-scale language models for biology, with ESM3 capable of generating novel protein sequences with specified structural and functional properties.

Trained on millions of natural sequences → PLM generates novel sequences → Likely to fold & perform target function

The investor-relevant constraint:
PLMs reason primarily in sequence space. They don't inherently operate in three-dimensional atomic geometry. That's not a flaw; it's a design choice. But it does mean that for applications requiring precise control over three-dimensional structure from scratch, you need a different tool.

Model type 2: Diffusion models

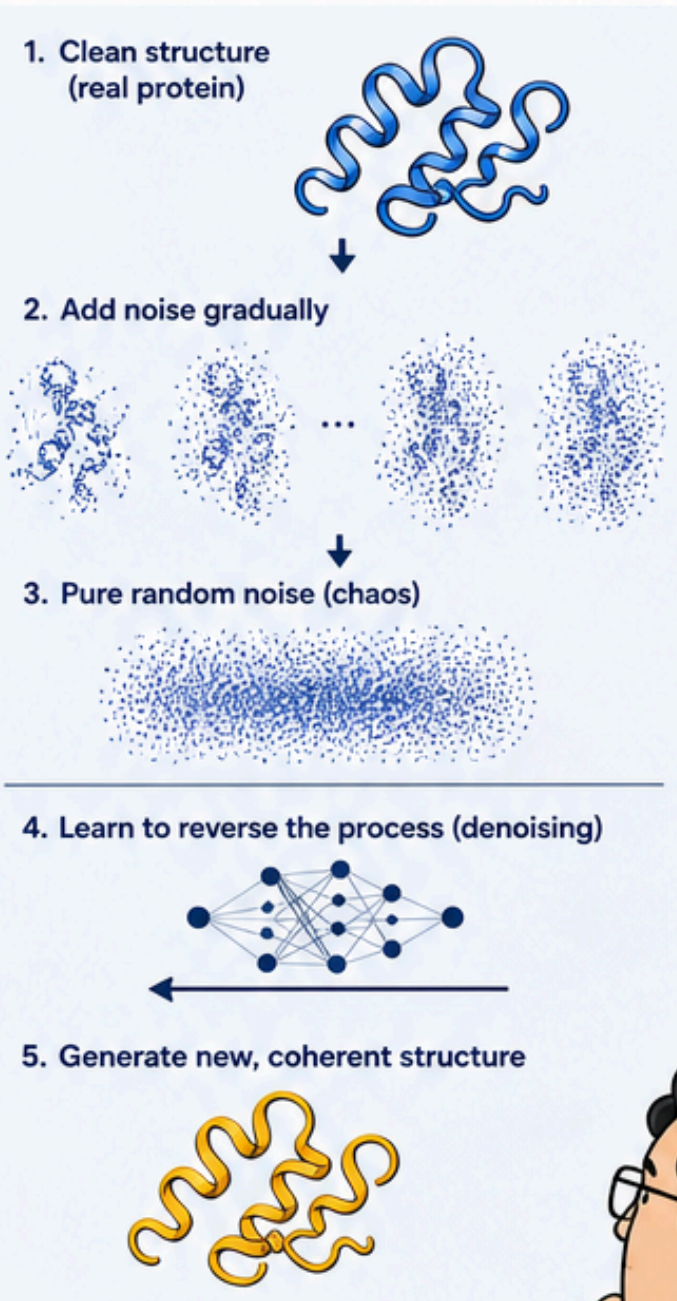
What they are

Diffusion models work by learning to reverse a noise process. The training procedure takes a known, clean data point, gradually adds random noise until it becomes indistinguishable from chaos, and then trains a neural network to denoise it step by step. Once the network learns to reverse noise reliably, you can start from pure random noise and generate a new, coherent data point that looks like something from the training distribution.

For protein design, the "data" isn't text tokens. It's three-dimensional atomic coordinates: the precise positions of every backbone atom in three-dimensional space. Diffusion models offer an edge in generating high-diversity folds, which can be conditioned through a wide variety of inputs or design objectives.

The most cited example is RFdiffusion, developed at the Baker Lab. RFdiffusion is a generative diffusion model for protein structure design, using a RoseTTAFold-based network as its denoising backbone. It can generate novel protein backbones conditioned on functional site geometries, target shapes, or binding requirements, and has been applied to scaffold enzyme functional sites and generate de novo binders.

Experimental validation reported in Watson et al. (Nature, 2023) demonstrated success rates that varied by task type — reaching around 20% in some binder design contexts from small sets of expressed candidates — which the authors noted as substantially higher than classical computational design approaches. Readers should consult the original paper for the full breakdown by task, as results varied considerably across applications.



DIFFUSION MODELS: WHAT THEY'RE COMMERCIALLY GOOD FOR

- Novel therapeutic targets
- Enzyme scaffolds for industrial biotechnology
- Binder proteins with precise geometric complementarity

THE CONSTRAINT:
A diffusion model outputs a backbone geometry. It doesn't automatically give you a sequence that will fold into that geometry. Pipeline integration is essential.

> "Structure-based methods such as RFdiffusion consistently perform well across most scenarios, with RFdiffusion showing particular robustness in generating a high number of designable scaffolds."
>
> –23-1

Model type 3:
Structure prediction backbones

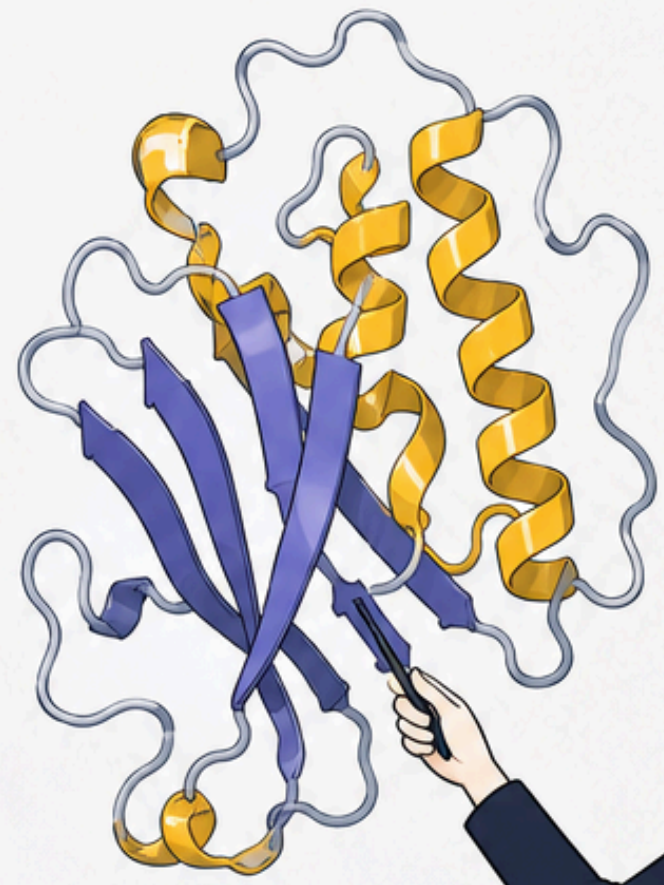
This is the category most investors have heard of, because it contains AlphaFold.

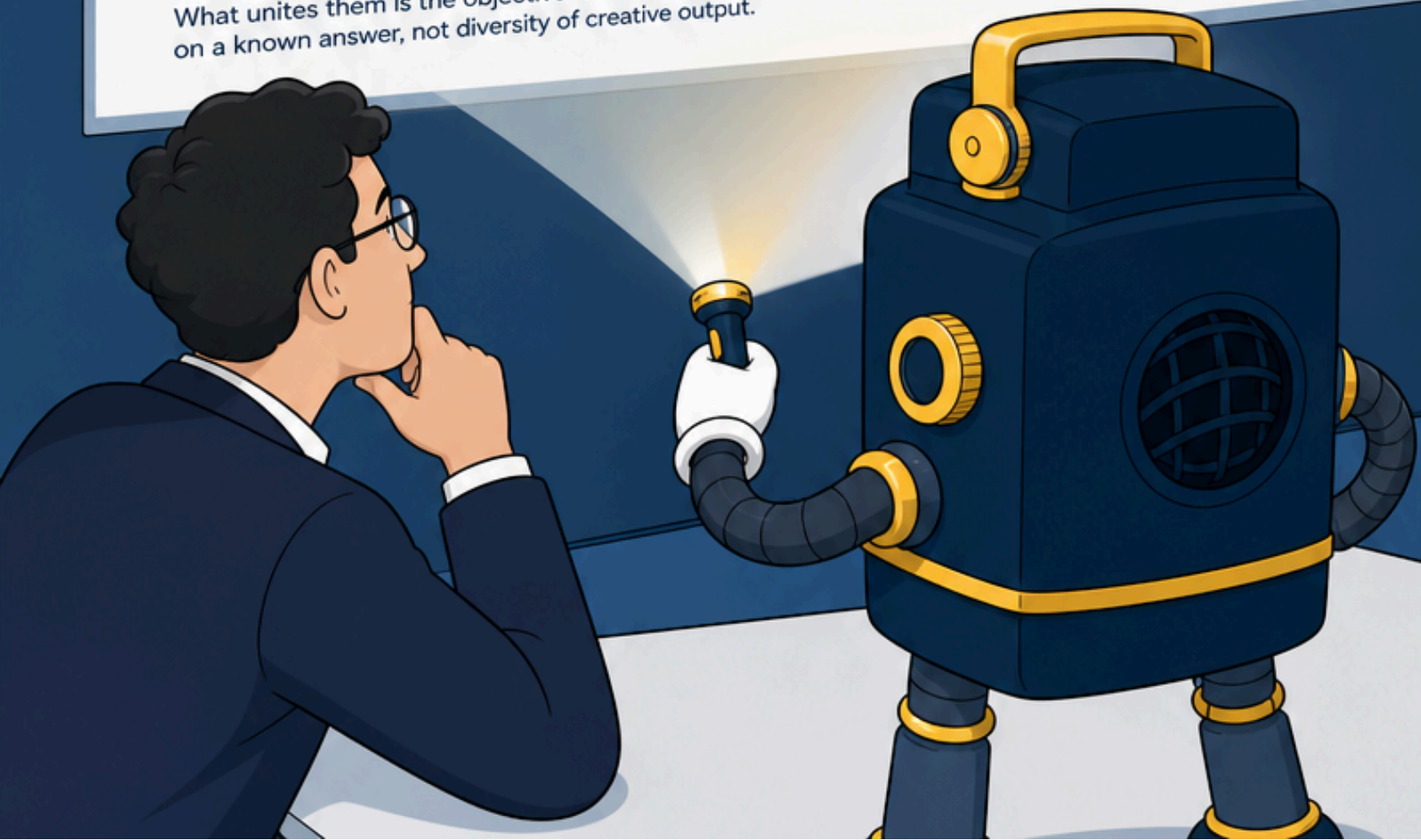
Discriminative models are trained differently from generative ones. Rather than learning to create diverse outputs, they're trained to predict a single, most-accurate answer to a specific question. Discriminative models learn conditional probability distributions and focus on predicting the most likely output for a given input. AlphaFold 2 is the canonical example: it predicts a protein's three-dimensional structure from its amino acid sequence but does not generate new sequences.

AlphaFold2's performance at the CASP14 competition represented a step-change in the field — solving structures that had resisted experimental determination for decades and achieving GDT scores far exceeding prior methods. Its developers subsequently reported confident structural predictions covering approximately 98.5% of the human proteome, a figure widely cited in the literature though one that reflects prediction confidence thresholds rather than experimental validation of every modelled structure.

AlphaFold3 extended this further. AlphaFold3 extends predictions to protein complexes, interactions with DNA, RNA, small molecules, ions, and other biomolecules using a diffusion-based framework. This matters commercially because drug design usually involves predicting not just a protein's shape in isolation, but how it interacts with a small molecule or another protein.

The broader AlphaFold-lineage includes RoseTTAFold, ESMFold, and Chai-1, each with slightly different architectural choices and computational trade-offs. What unites them is the objective function: accuracy on a known answer, not diversity of creative output.





Structure prediction backbones:
What they're commercially good for

500 candidate sequences

PLM / Diffusion model

Structure predictor

Top 20 high-confidence structures

Generated in **seconds**
Cost: fraction of a cent

Evaluated in **hours**
Cost: fraction of a cent

Synthesise only the top 20
Save millions & months

The part that actually matters for investors: hybrid pipelines

A modern AI protein design pipeline typically looks like this:

1 Protein Language Model

Proposes candidate sequences based on statistical patterns learned from evolutionary data, or explores mutations around a known starting point.

2 Diffusion Model

Generates novel three-dimensional backbone scaffolds for cases where no natural starting point exists, or conditions sequence generation on a target geometry.

3 Structure Predictor

Discriminative structure predictor (AlphaFold-lineage) acts as the in-silico filter, predicting which of the generated candidates are likely to fold correctly and validating structural plausibility before any synthesis decision is made.

In practice: RFdiffusion can produce multiple backbone structures per prompt, ProteinMPNN can generate multiple sequences per backbone, ESM models propose sequence variants, and AlphaFold-style predictors return multiple structural candidates.

This is not merely a workflow convenience. The compression effect is central to the commercial thesis. The rationale for heavy investment in these platforms is that any successful platform could dramatically shorten the path to first-in-man trials.

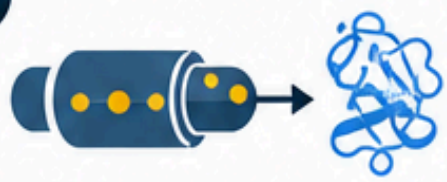
The critical investor question to ask isn't "do they use AI?" It's: at which stage does the human loop close? How many wet-lab experiments per in-silico cycle? What's the measured hit rate on synthesised candidates? Companies that can answer those questions with data, not just a pipeline diagram, are the ones building real moats.

	Protein language models	Diffusion models	Discriminative predictors
Trained to do	Learn sequence patterns from evolutionary data	Generate novel 3D backbone geometries	Predict the single most accurate structure for a given sequence
Input	Protein sequence (partial or masked)	Random noise in 3D coordinate space	Complete amino acid sequence
Output	Novel or optimised sequences	Novel 3D backbone structures	Predicted 3D structure
Primary commercial use	Directed evolution acceleration, sequence exploration	*De novo* scaffold design, novel fold creation	*In-silico* candidate filtering and validation
Key examples	ESM3, EvoDiff, Profluent's models	RFdiffusion, Chroma, Framediff	AlphaFold 2/3, RoseTTAFold, ESMFold, Chai-1
Generates diversity?	✓ Yes	✓ Yes	✗ No, optimised for accuracy

What to look for in a diligence conversation

When you're sitting across from a synthetic biology or biotech founder claiming AI protein design advantage, the three questions that separate real platform differentiation from a well-dressed literature review:

1



Where in the pipeline is your novel contribution?

A company that's built genuine IP around how it orchestrates the generative-then-discriminative pipeline is more defensible than one that's stitched together publicly available models in the obvious sequence.

2

What's your wet-lab hit rate?

The entire value proposition of hybrid pipelines is that in-silico filtering improves the quality of candidates sent to synthesis.[13] Investors should heed what exactly the AI is modelling and how it fits into therapeutic development; companies must show not just algorithms but real endpoints, novel molecules, and preclinical success.[13,14]

3

What does your proprietary training data look like?

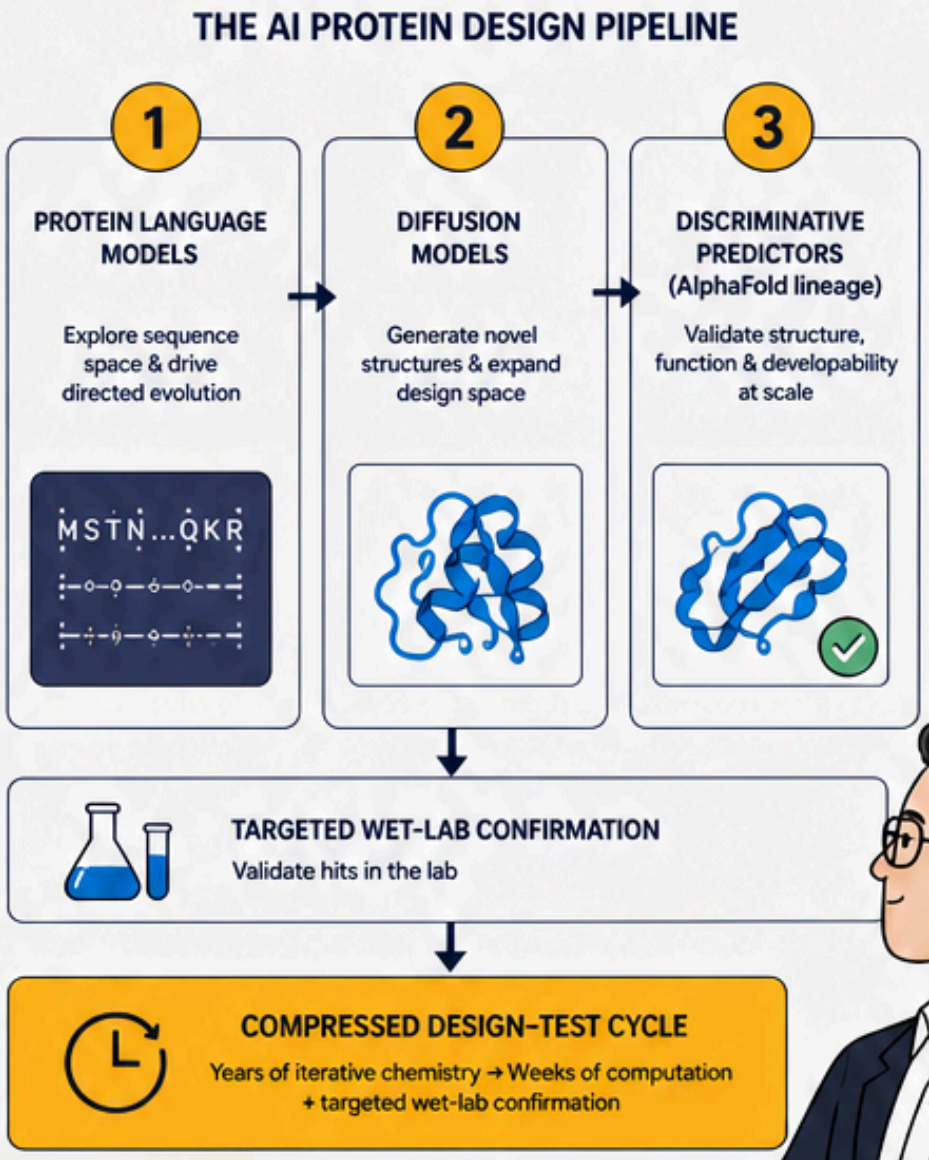
[22] Challenges still exist pertaining to modelling sequence-structure-function relationships and ensuring robust generalisation beyond the regions of protein space spanned by training data.
A company with a proprietary experimental dataset that feeds back into model retraining is compounding a data advantage that pure software platforms can't replicate.

The bottom line

The AI protein design space is real, it's growing, and the underlying science is genuinely compelling. But "AI protein design" as a category descriptor is too broad to support a thesis on its own. The three paradigms do different things, at different points in the pipeline, with different implications for what can be defended commercially.

Protein language models accelerate sequence-space exploration and directed evolution. Diffusion models open up entirely new structural territory. Discriminative predictors, the AlphaFold lineage, are the validation engine that makes the other two commercially viable at scale. Together, in a well-orchestrated pipeline, they compress the design-test cycle from years of iterative chemistry into weeks of computation followed by targeted wet-lab confirmation.

That compression is where the investment thesis lives. The question isn't whether a company uses these tools. It's whether they've built a pipeline that uses all three in a way that genuinely changes what's achievable, at what cost, and how fast.



infrairis

https://startups.infrairis.com/