Development of realistic and safe AI/ML applications for biotherapeutic characterization and process development

Jeremy Shaver

Senior Director, Research Data Science

Biotherapeutics Process Sciences (BTxPS)

Pfizer

27 Jan 2025 – CASSS CMC Forum

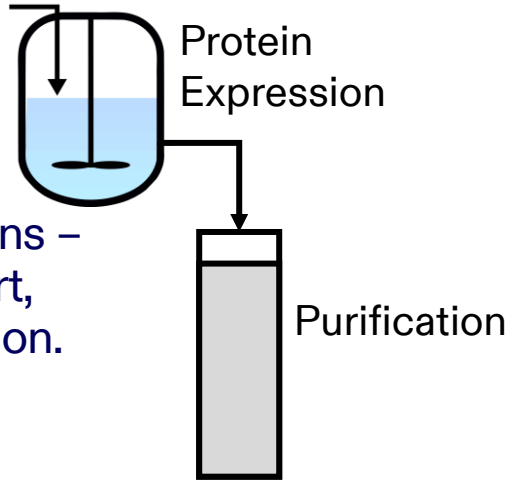
“In an era of AlphaFold, protein language models, and other powerful neural networks, why aren’t we doing all process development *in silico*?”

Some AI/ML opportunities in the process development space…

Liability prediction – when and how a molecule will require a non-platform process.

Formulation and stability prediction – auto-designing stable drug products and knowing the stability.

Identifying key manufacturing conditions – reduced process characterization effort, automatic process design, and execution.



Analytical methods development – Accelerated method development, better reagents and release methods, faster.

nature › news feature › article

NEWS FEATURE | 06 November 2024

The antibodies don't work! The race to rid labs of molecules that ruin experiments

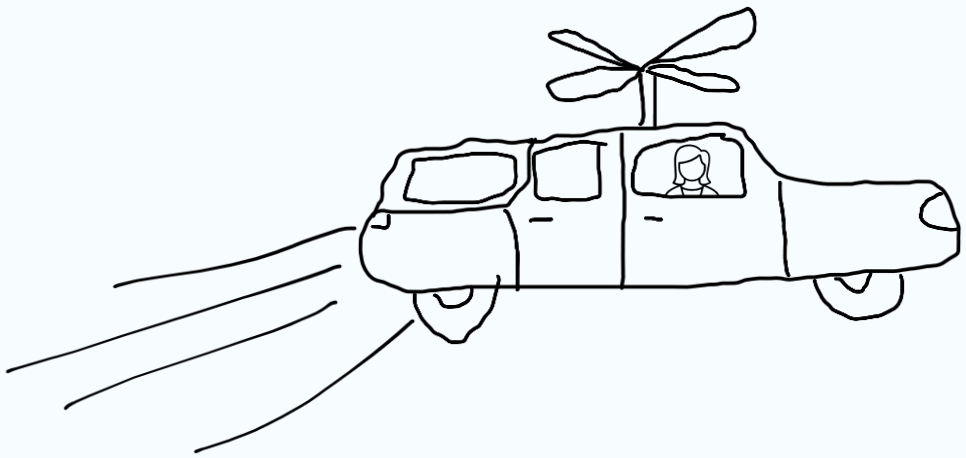
Poorly performing antibodies have plagued biomedical sciences for decades. Several fresh initiatives hope to change this.

By Diana Kwon

DOI: 10.1038/d41586-024-03590-0

We have self-driving cars. Why don't we have flying cars?

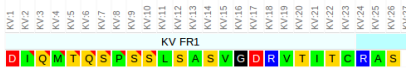
We have the technology but there are practical considerations that make it not yet prudent to use widely



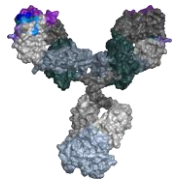
- Failure modes are sometimes unexpected
- Users need extensive training to safely handle normal operations and failure conditions
- They aren't well designed for some specific but important conditions
- The safety margins in ~~flying~~ process development are more stringent

Complexity of sequence-to-behavior deep learning tasks

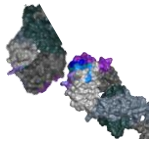
Sequence



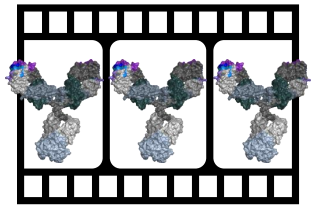
Molecular structure



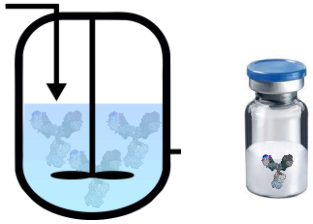
Protein-protein interactions



Protein dynamics



Protein Language Models (PLMs) and other NNs feeding bespoke models



Biophysical behavior in process development

Behavior in complex biological systems

Some of the Sequence to Structure models released in 2024

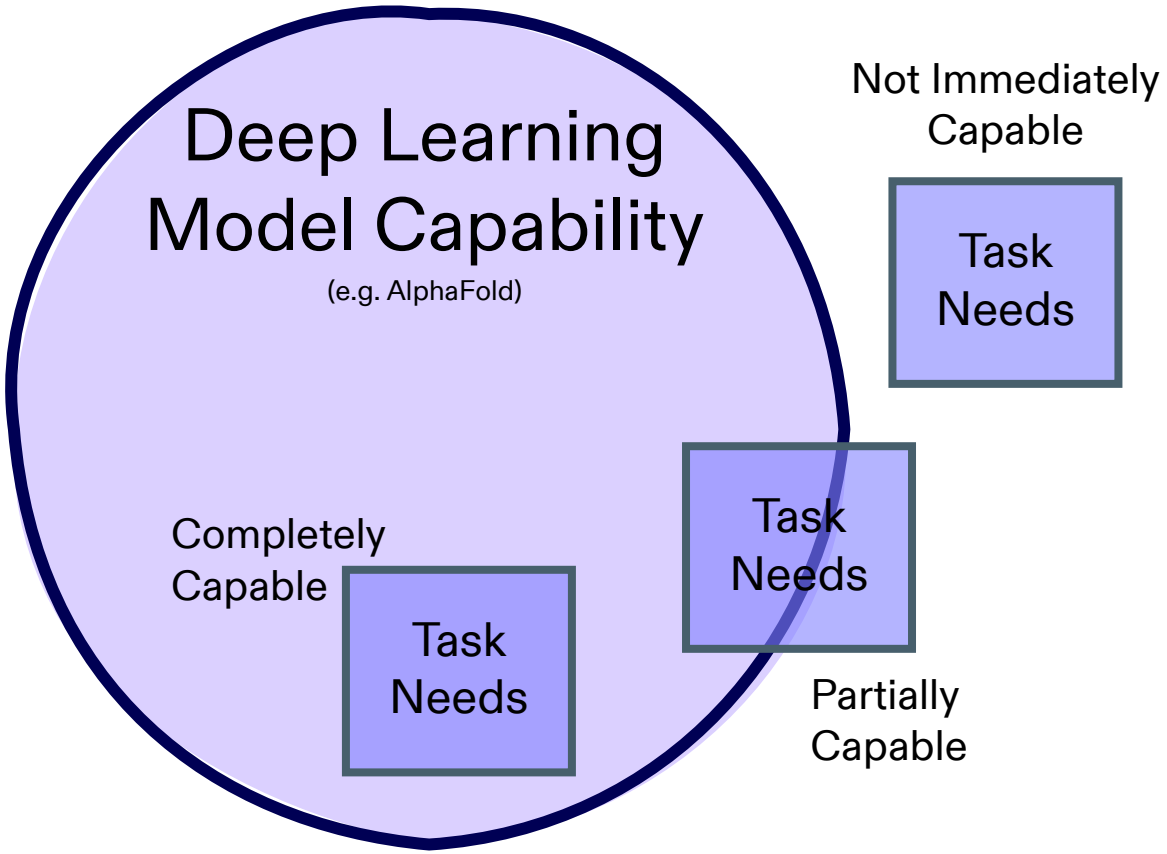
Date	Software
2024-05	Alphafold 3
2024-08	HelixFold3
2024-10	Chai-1
2024-11	Protenix
2024-11	Boltz

http://blog.booleanbiotech.com/alphafold3-boltz-chai1.html

Increasing complexity ***(not to scale!)***

Deep learning models are “learning” physical science

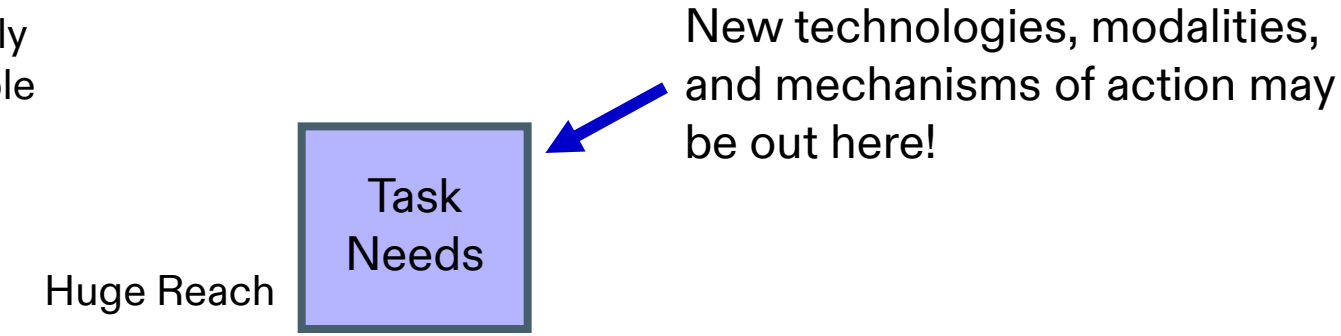
But how broad are the rules they’re learning?



We’re training the models to infer physics, chemistry, and biology rules. The data it learns from scopes the rules it learns and tasks it is capable of:

- Relevant tasks: in silico estimates possible
- Similar tasks: partially capable (approximate)
- Adjacent tasks: not capable, extrapolatable?
- Distant tasks: not capable – too different

- Does the science and data exist yet to learn the necessary rules?



Dickinson lab protein–protein interaction challenge

15 binding proteins and 15 target proteins
(+ decoys)

Experimental data: nearly perfect selectivity between targets and binders. (Top Right)

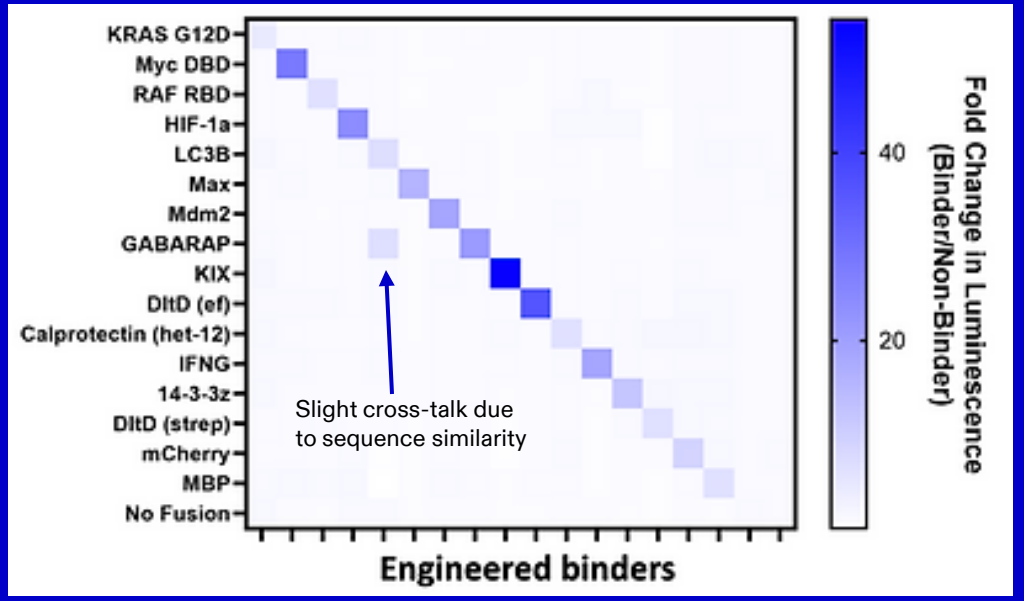
Computational data: AlphaFold3 does very poorly. No accurate predictions. (Bottom Right)

Community-challenge results: other approaches were 25% accurate *at best.*

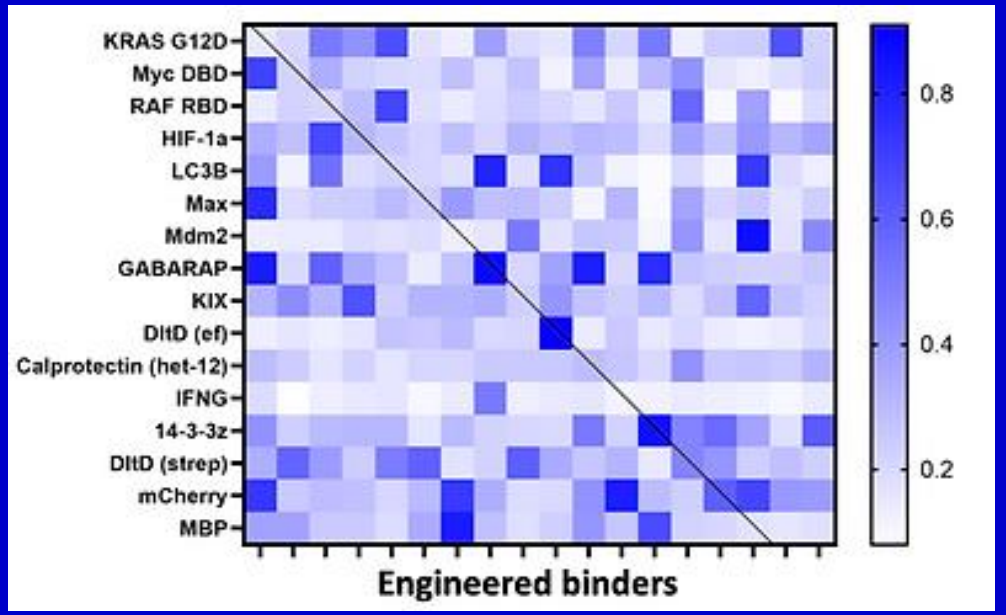
Used with permission from
Professor Bryan Dickinson, University of Chicago

https://www.dickinsonlab.uchicago.edu/ppi-challenge

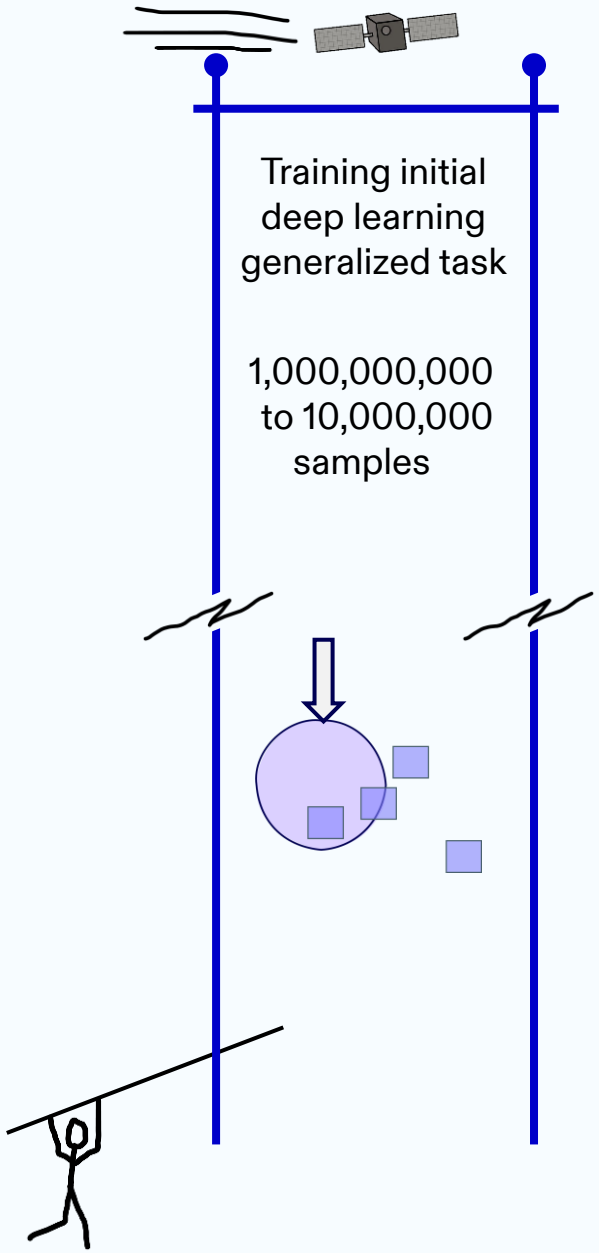
Measured Binding Selectivity

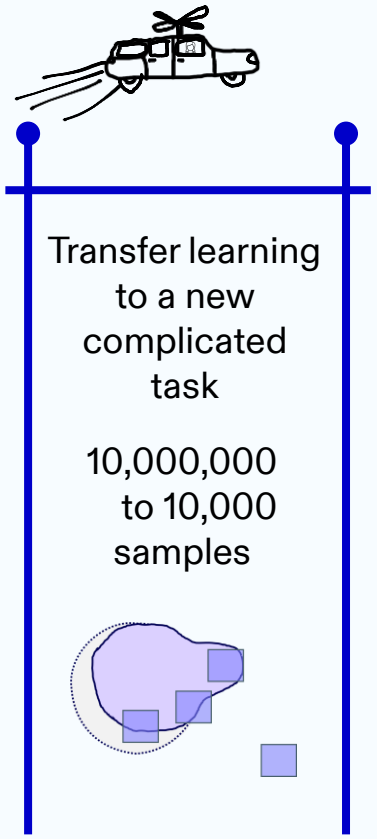


AlphaFold3 Predicted Binding Selectivity



Vast differences in data required by modeling approaches





1. The amount of data we need to train models is large
2. Most of us are using models trained by others – need transparency on what data was used
3. Many tasks are still only approximated by these models

A-Alpha Protein-Protein Interaction models

Evidence of the amount of data it takes

Protein language model to predict how to mutate an antibody into a better binder for a specific antigen

1. Train generalized binding model ("AlphaBind") on ***7.5 million*** affinity measurements across diverse set of antibody-antigen pairs

2. Fine-tune AlphaBind model to a specific antibody-antigen system measuring **30 thousand** antibody variants against the target antigen

3. Make predictions for promising new variants

   Repeat step 2 for each specific antibody-antigen system of interest

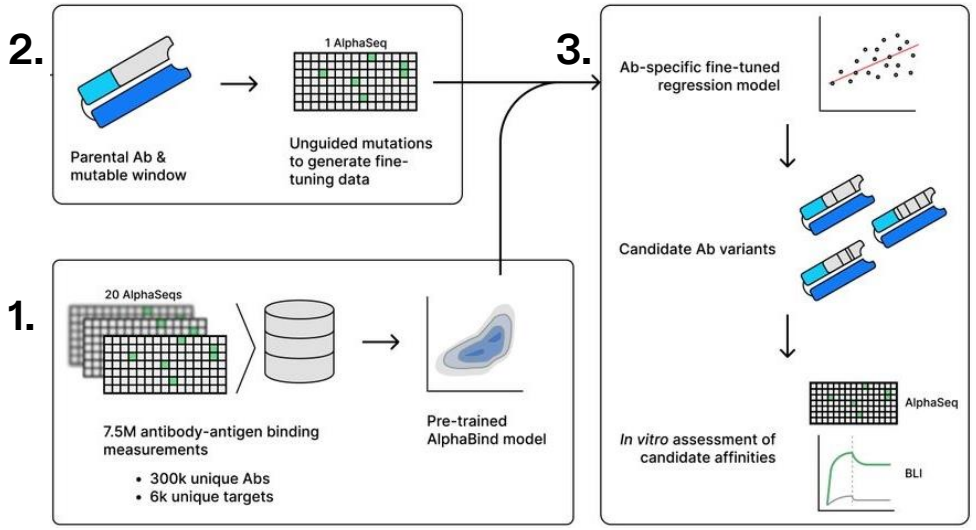


Figure 1B from preprint used with permission

November 2024 Biorxiv Preprint

AlphaBind, a Domain-Specific Model to Predict and Optimize Antibody-Antigen Binding Affinity

Aditya A. Agarwal, James Harrang, David Noble, Kerry L. McGowan, Adrian W. Lange, Emily Engelhart, Miranda C. Lahman, Jeffrey Adamo, Xin Yu, Oliver Serang, Kyle J. Minch, Kimberly Y. Wellman, David A. Younger, Randolph M. Lopez, Ryan O. Emerson

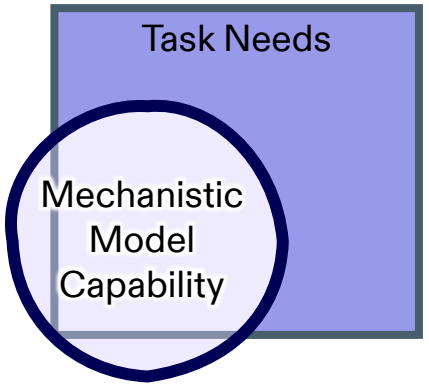
doi: https://doi.org/10.1101/2024.11.11.622872

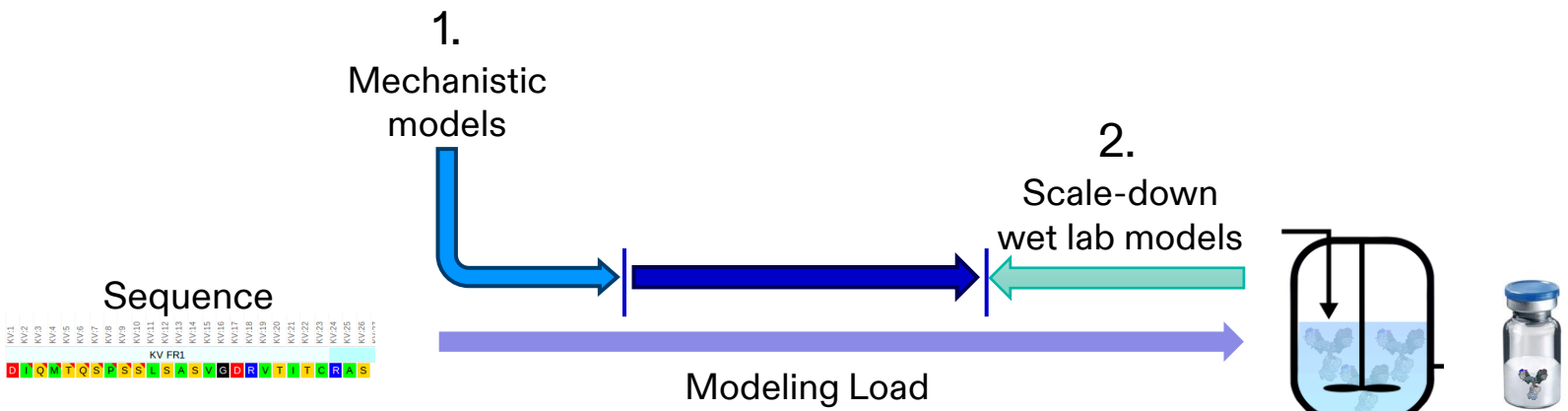
Making it easier: first-principle and wet lab scale-down models

1. Mechanistic models (based on first-principle science) can reduce complexity by providing approximation of portions of the task.

2. Task complexity can be reduced by targeting scale-down wet lab models where prior knowledge connects those models to full behavior. (e.g. surrogate assays)

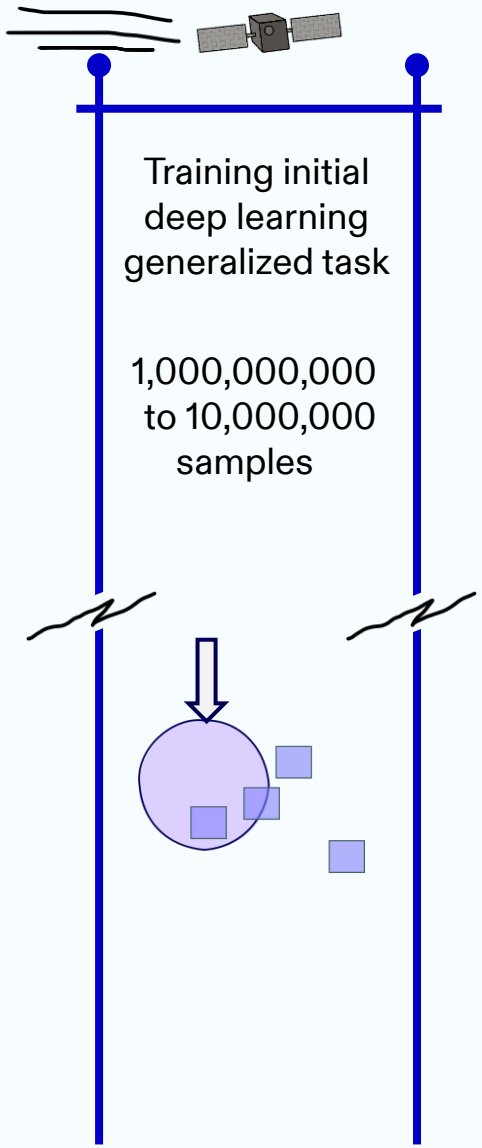
Mechanistic modeling examples:
- Chromatographic retention behavior
- Metabolomic pathways
- Molecular modeling
- Physics Informed Neural Networks (PINN)

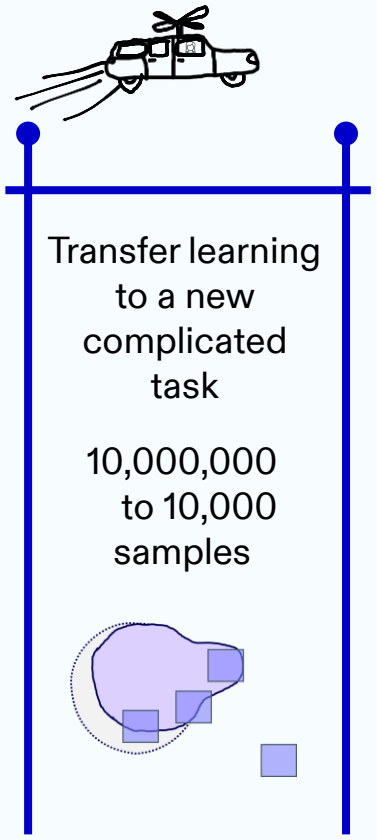


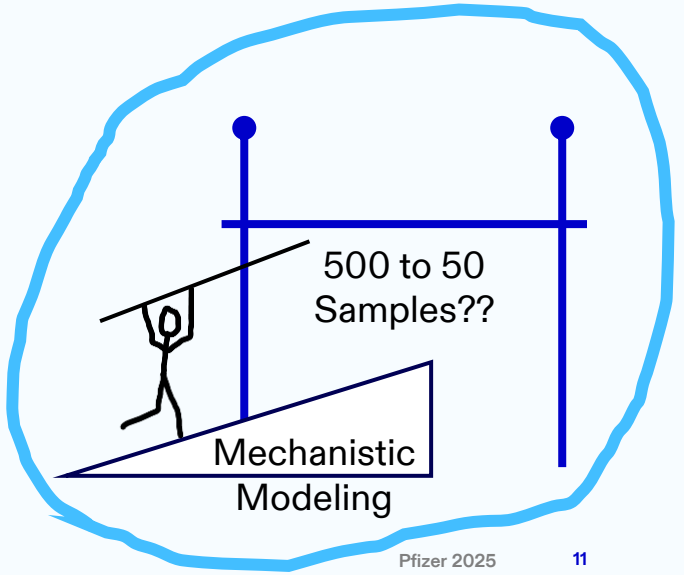


Biophysical behavior in process development

Vast differences in data required by modeling approaches

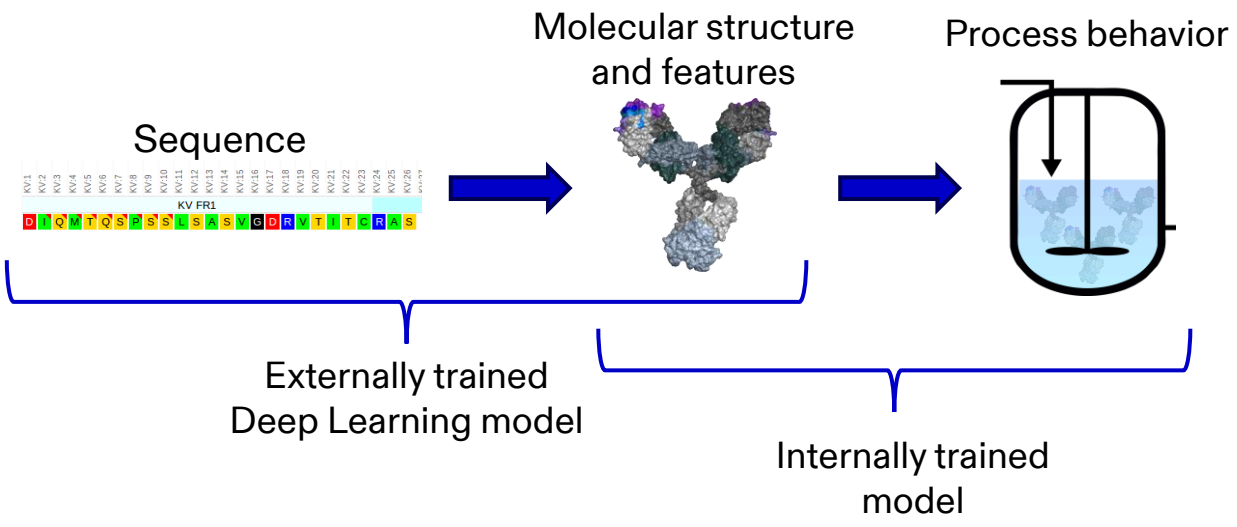


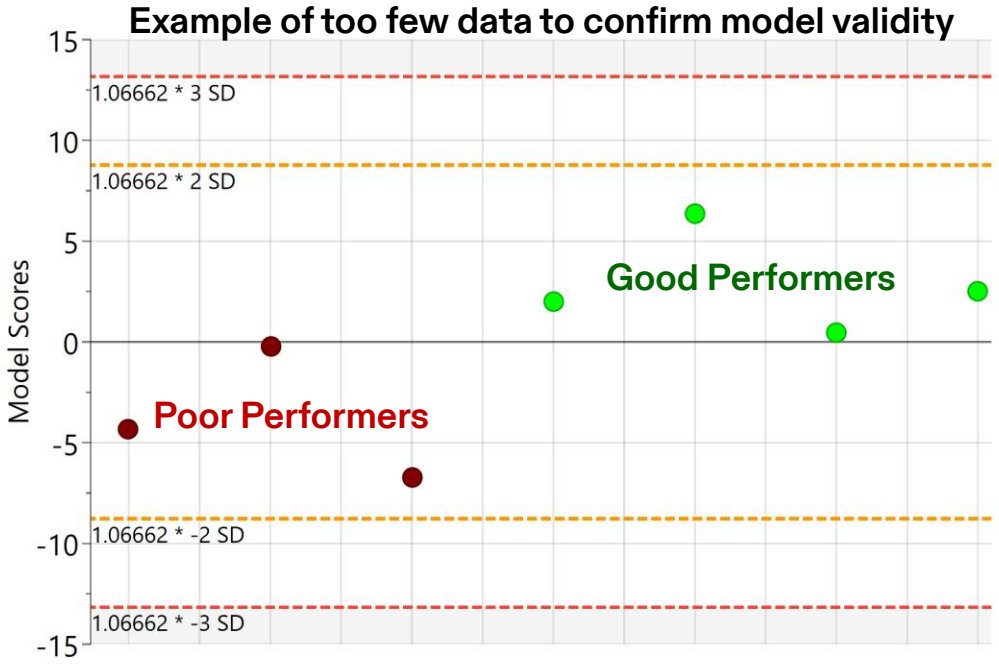




Good practices: Assessment of capability requires sufficient data

Even if the deep learning model is correct, verification of the use of its outputs still requires sufficient data to verify the output.





Akin to implementing advanced process monitoring technology:

pH probe < hyperspectral instrumentation < deep learning model

Good practices: Successful modeling requires appropriate data

While collecting data is important, collecting *diverse* data is critical.

We *must* capture problematic cases, not just successful cases

Example 1 : Challenging antibodies

Jain et al. 2017 : published the measured biophysical property assays for 137 commercial antibodies and their sequences.

Numerous papers show training machine learning models to predict antibody “developability”.

The set contains no molecules that were dropped due to problematic behaviors.

RESEARCH ARTICLE | APPLIED BIOLOGICAL SCIENCES |

Biophysical properties of the clinical-stage antibody landscape

Tushar Jain, Tingwan Sun, Stéphanie Durand, +15, and K. Dane Wittrup Authors Info & Affiliations

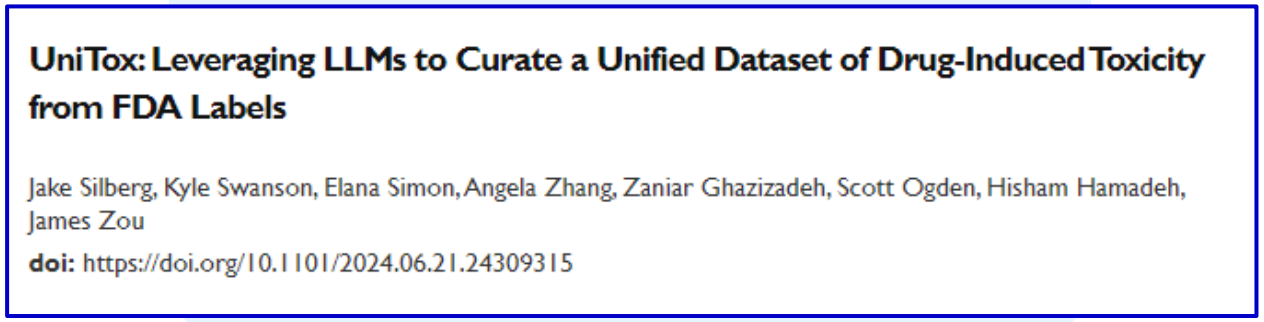
Edited by James A. Wells, University of California, San Francisco, CA, and approved December 13, 2016 (received for review October 2, 2016)

January 17, 2017 | 114 (5) 944-949 | https://doi.org/10.1073/pnas.1616408114

Example 2: Toxicity data

Silberg et al. 2024 – “Drug-induced toxicity is one of the leading reasons new drugs fail clinical trials. Machine learning models that predict drug toxicity from molecular structure could help researchers prioritize less toxic drug candidates.”

Created: “UniTox” dataset of 2,418 ***FDA-Approved drugs*** with drug-induced toxicity summaries.



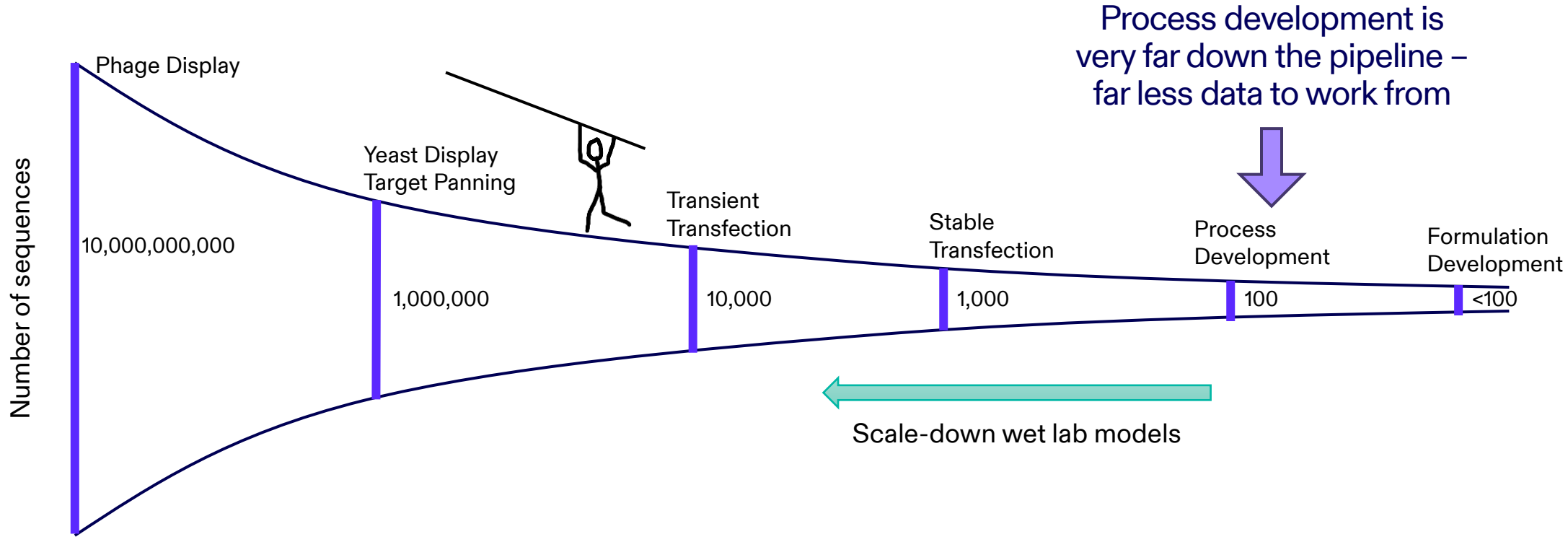
UniTox: Leveraging LLMs to Curate a Unified Dataset of Drug-Induced Toxicity from FDA Labels

Jake Silberg, Kyle Swanson, Elana Simon, Angela Zhang, Zaniar Ghazizadeh, Scott Ogden, Hisham Hamadeh, James Zou

doi: https://doi.org/10.1101/2024.06.21.24309315

But why do we hear so much about in silico methods revolutionizing the pharma industry?

Discovery is orders of magnitude larger, and even inaccurate models can help reduce development risks.

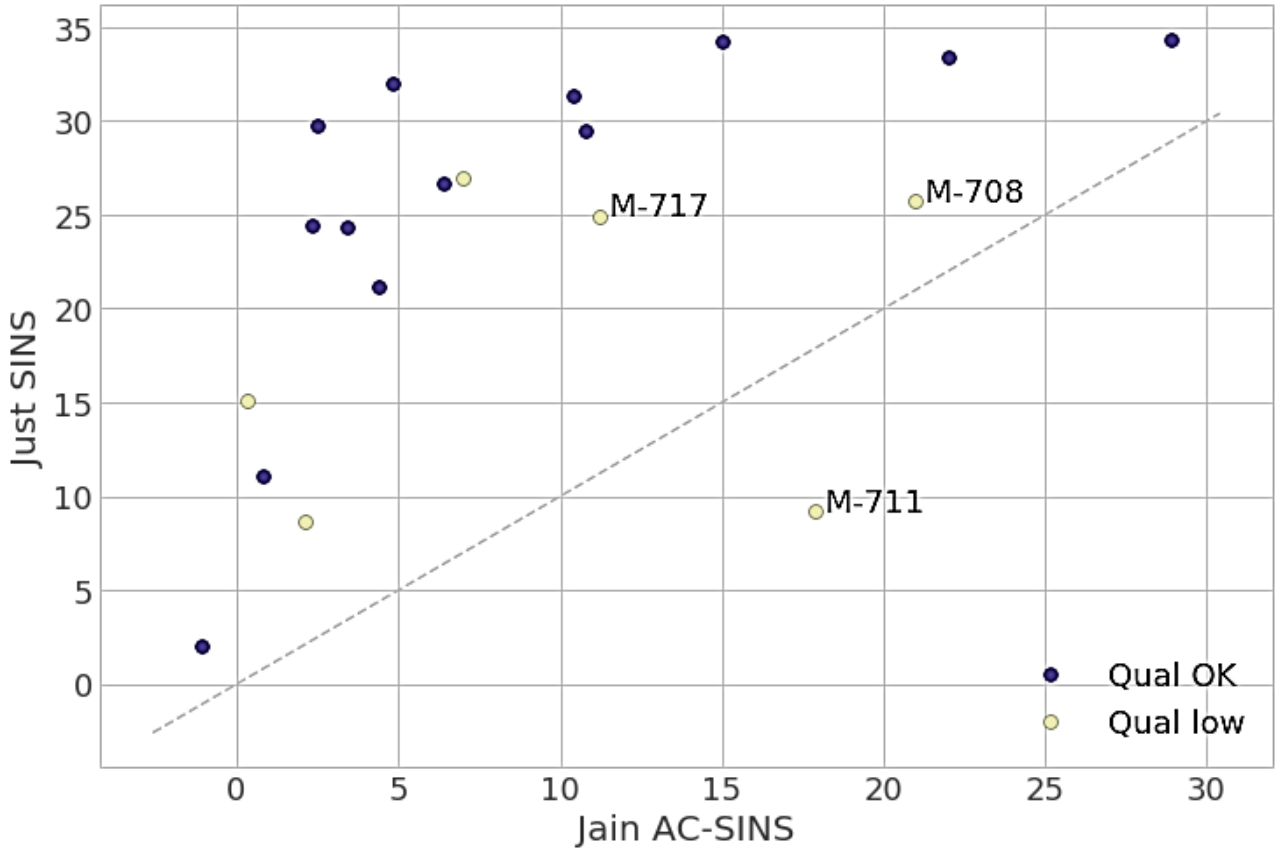


Federated learning?

Aggregation assay example

Two labs with published, comparable biophysical assays comparing identical sequences. Results showed significant non-linearity.

Advantage of extra data may be offset by significantly more complicated model.



Results presented by J. Shaver at CCD 2018 – Courtesy Just - Evotec

How do we expand use of *in silico* methods in process development?

Don't...

- Don't under-appreciate the complexity of the science
- Don't over-estimate the amount of known scientific knowledge
- Don't loose appreciation for tempered "traditional" machine learning methods

Don't expect deep learning tools to:

- Eliminate work completely
- Eliminate the need for critical thought
- Work "out of the box" without some development* and testing

**YET*

Do...

- Do collect more data, embrace high-throughput methods and incorporate known science
- Do collect the right data – positive and negative and note knowledge limitations
- Do push to have "best practices" around both traditional and advanced *in silico* methods

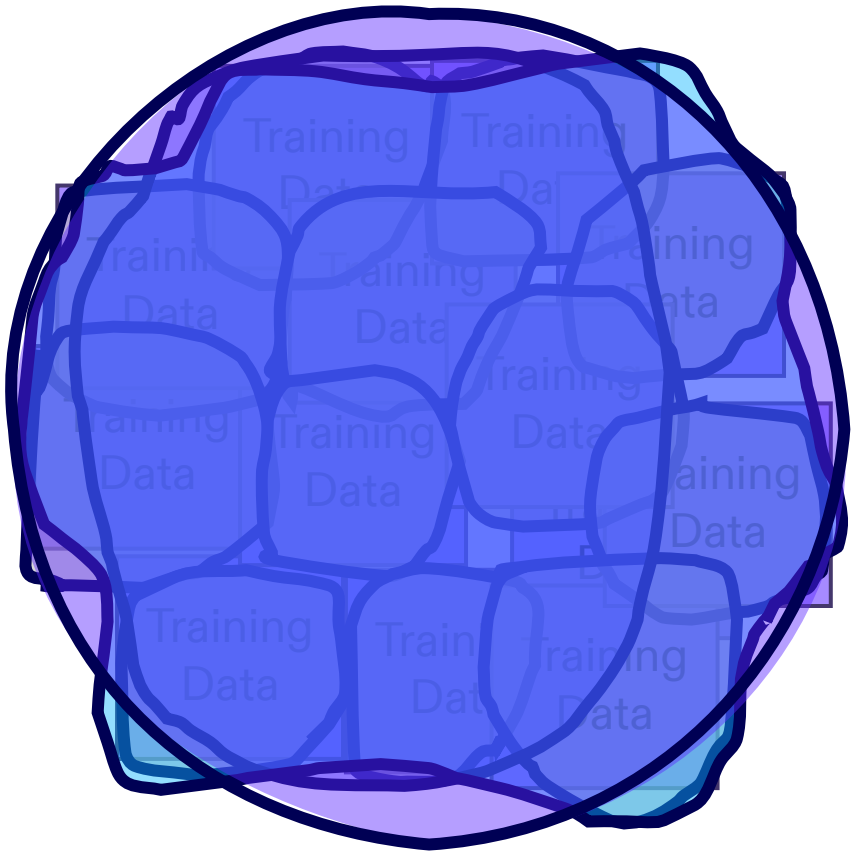
Do expect deep learning tools to:

- Reduce the number of experiments (once we have relevant data – prior knowledge)
- Deepen our understanding of the science
- Identify where we have holes in our knowledge and data

Thank You

Deep learning models are “learning” physical science

The data we train with matters



Different models may extract different kinds of information.

But the amount of information *in* those data doesn’t change.

Currently many of the models in this space are based on nearly identical data sources.