WELCOME

Practical Introduction to Data Science Methods, Using AI and GenAI

Dr Raminderpal Singh
raminderpal@hitchhikersai.org

TECHNOLOGY·PROCESS·CULTURE·DATA
20/15 VISIONEERS

slas
Come Transform Research

DON'T PANIC!
THE ANSWER IS 42

Thank you for joining today

There will be 4 topics covered. Plus several quick discussions & activities.

In the afternoon, there will be a lab. In groups of 2-3, using your **laptops**.

The activities and lab leverage ChatGPT (and/or Claude)* in your browser. No software installs will be needed.

Please open your laptops and make sure you are connected to the WiFi.

Remember to add any notes to the Workshop Notes doc, for everybody to learn from.

For your Starter Kit, I will tag the slides in this presentation.

This presentation: https://shorturl.at/EMOvv

*Note: today's discussions and activities use publicly available data. In your workplace, you may need to use secure applications behind the firewall.

Note

This course is less about theory and more about kickstarting pragmatic starting points for you **to develop an opinion and to take advantage** of the AI technologies out there.

Please immerse yourself in the topics and ask lots of questions throughout.

There are no “bad questions”.

Sources for the content in these slides

Much of the material in this presentation has been cut & paste from the references provided in the eBooklet. Please refer to them for more details.

Schedule for today

8:30am	Welcome
9am	TOPIC 1 - Big picture, trends, challenges, reality check
9:30am	TOPIC 2 - ML, Knowledge Graph, LLMs
10am	Coffee Break
10:30am	TOPIC 2 (con't)
	TOPIC 3 - Data Quality, Culture, FAIR
12pm	LUNCH
1pm	TOPIC 4 - No-code Scientific Workflow; Intro to Lab
2pm	Lab
2:30pm	Coffee Break
3:30pm	Team readouts for Lab
4 - 4:30pm	End-of-day; Questions

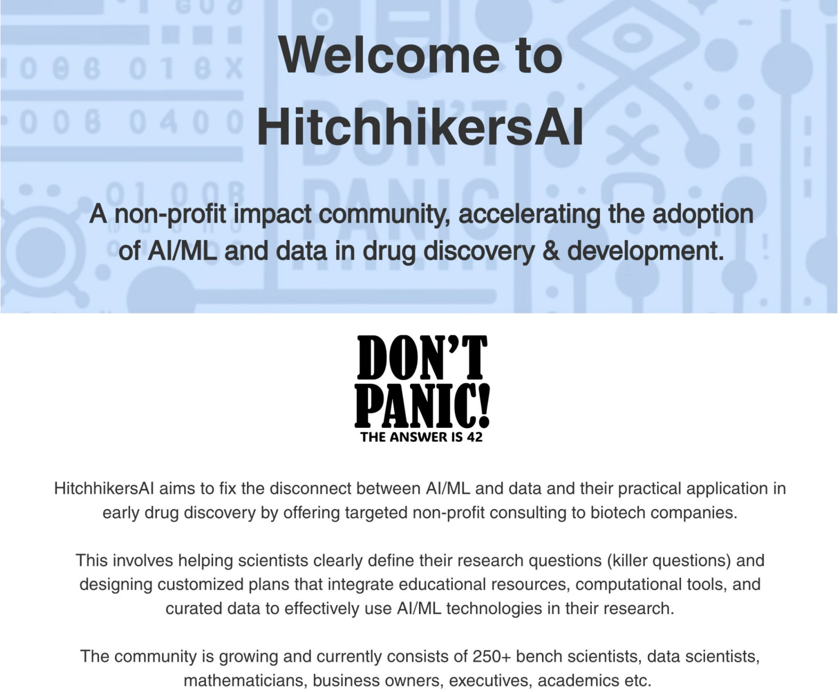
The times for covering each of the four TOPICs are for guidance only.

Break and Lunch areas

AM Break:	10:00 am - 10:30 am; Cityside Foyer – Second Level
Lunch:	12:00 pm - 1:00 pm; Center Terrace
PM Break:	2:30 pm - 3:00 pm; Cityside Foyer – Second Level

Please do not leave valuables in the course rooms when you are not there. Course rooms will NOT be locked or monitored during breaks.

Please join our grassroots community



Welcome to HitchhikersAI

A non-profit impact community, accelerating the adoption of AI/ML and data in drug discovery & development.

DON'T PANIC!
THE ANSWER IS 42

HitchhikersAI aims to fix the disconnect between AI/ML and data and their practical application in early drug discovery by offering targeted non-profit consulting to biotech companies.

This involves helping scientists clearly define their research questions (killer questions) and designing customized plans that integrate educational resources, computational tools, and curated data to effectively use AI/ML technologies in their research.

The community is growing and currently consists of 250+ bench scientists, data scientists, mathematicians, business owners, executives, academics etc.

https://www.hitchhikersai.org/

Please also check out my regular articles on AI in drug discovery

Drug
TARGET REVIEW

https://www.drugtargetreview.com/?s=raminderpal+singh

ARTICLE

Part two: the impact of poor data quality

26 August 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

ARTICLE

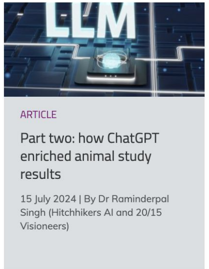
Bridging science and technology: a biotech CEO's perspective

14 October 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

ARTICLE

An industry leader's perspective on the complexity of scientific data

24 October 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)



ARTICLE

Part two: how ChatGPT enriched animal study results

15 July 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

ARTICLE

Part four: an industry leader's perspective on managing data quality

24 September 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

ARTICLE

Scientific workflow for hypothesis testing in drug discovery: Part 2 of 3

14 January 2025 | By Dr Raminderal Singh (Hitchhikers AI and 20/15 Visioneers, Nina Truter

ARTICLE

Part one: an introduction to data quality

14 August 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

ARTICLE

Kickstarting the use of AI for biotechs: part two

24 May 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

ARTICLE

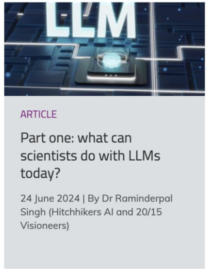
Kickstarting the use of AI for biotechs: part one

15 May 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

ARTICLE

Kickstarting the use of AI for biotechs: part three

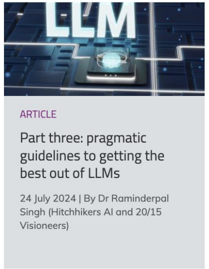
20 June 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)



ARTICLE

Part one: what can scientists do with LLMs today?

24 June 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

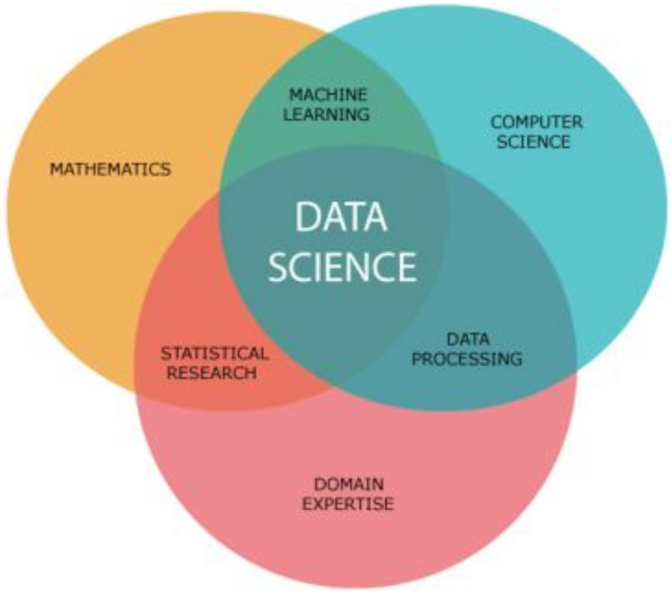


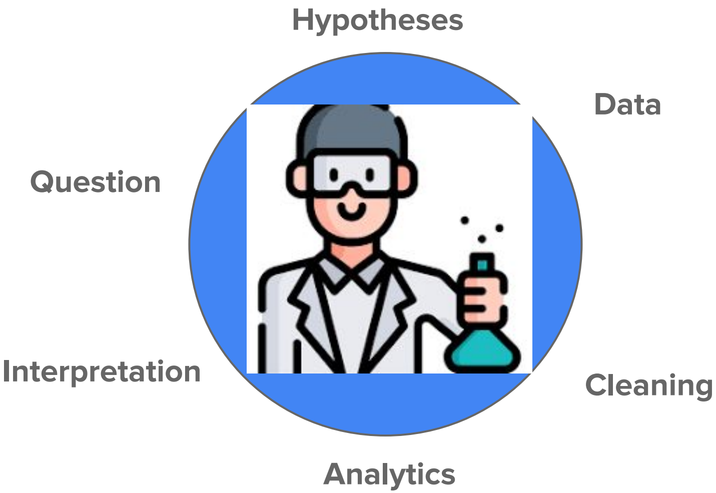
ARTICLE

Part three: pragmatic guidelines to getting the best out of LLMs

24 July 2024 | By Dr Raminderpal Singh (Hitchhikers AI and 20/15 Visioneers)

Before we get going … where do your skills lie?





TOPIC 1

Big picture, trends, challenges, reality check

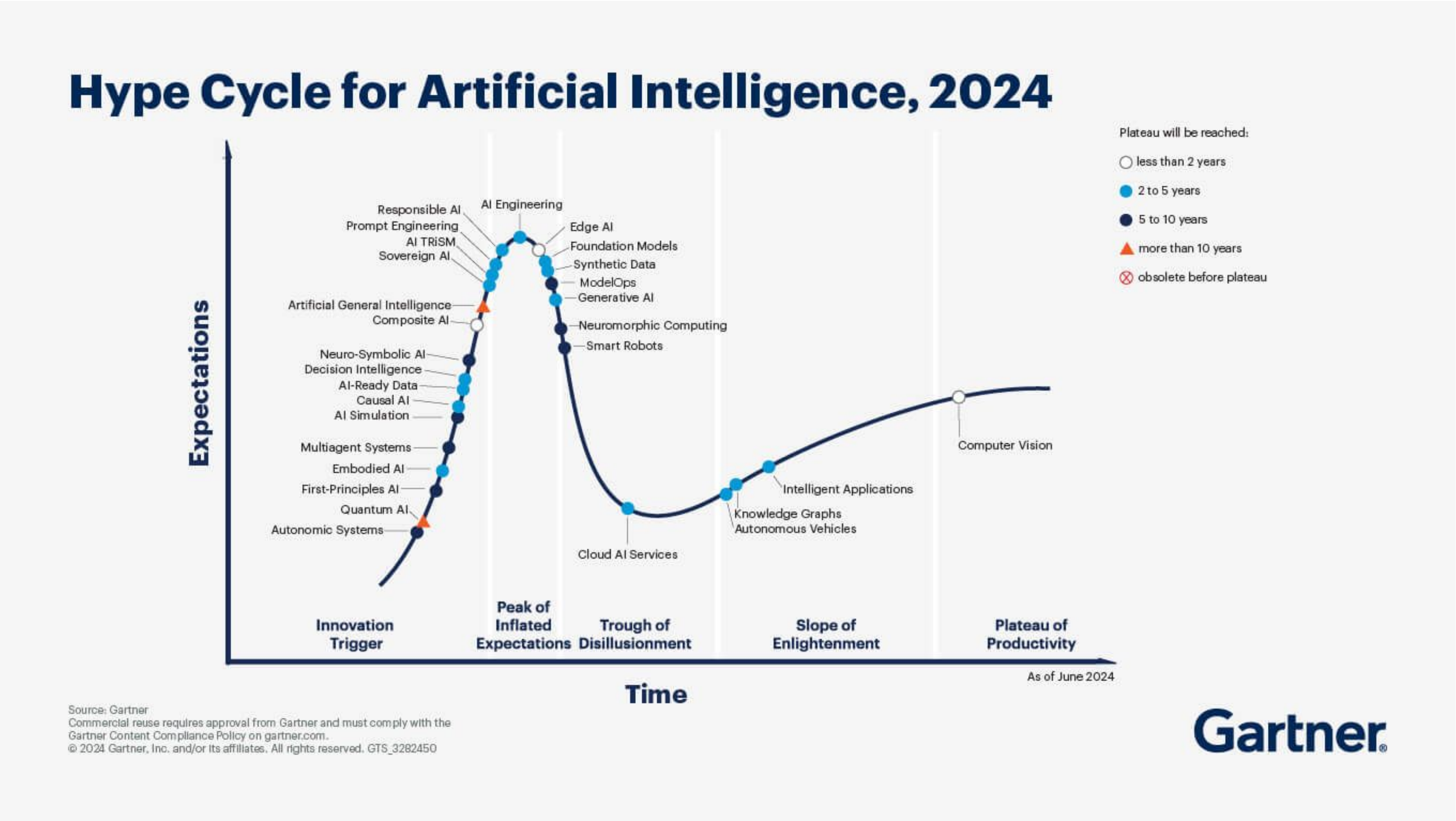
Principles:

ROI

Risks

Impact

Strategy



Hype Cycle for Artificial Intelligence, 2024

As of June 2024

Source: Gartner
Commercial reuse requires approval from Gartner and must comply with the Gartner Content Compliance Policy on gartner.com.

https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence

Takeaways from JPM 2025

The industry's attitude toward AI continues to evolve, perhaps even mature

I'll admit this is a weak sort of a takeaway, but the tenor of conversations about AI at this show has changed in the last couple of years. Now that generative AI (GenAI) has become so commonplace in the workplace in general, pharma people are starting to be a lot more specific with what they mean when they say AI – the sophisticated algorithms used in drug discovery and development, the GenAI-based chatbots for patient-facing use cases, or even "physical AI" – combining AI with robotics for next-generation applications in surgery and prosthetics, something NVIDIA was talking about at the show.

This is something of a relief for journalists and technologists, who were starting to become frustrated with the vague and hype-laden use of the term AI over the last couple of years.

There's also a growing recognition of machine intelligence as something distinct from human intelligence – not better or worse in every way, but better at some tasks and worse at others. Flagship Pioneering CEO Noubar Afeyan discussed the concept of "polyintelligence", the idea of recognising three distinct kinds of intelligence: human, machine, and natural. The latter describes, for instance, the pseudo-intelligence of viruses and bacteria rapidly evolving to counter vaccines and treatments meant to stop them.

Agree or disagree with the framing, it's a great example of an industry that finally understands AI well enough to thoughtfully consider its applications and its place in the larger picture of the industry.

When investors talked about AI at the various events I attended, they spoke about it the same way: as a tool, as something they look for their companies to use intelligently, but not generally as a product or service in and of itself to invest in.

https://pharmaphorum.com/rd/5-takeaways-jp-morgan-2025

Key Status & Trends of AI in Drug Discovery

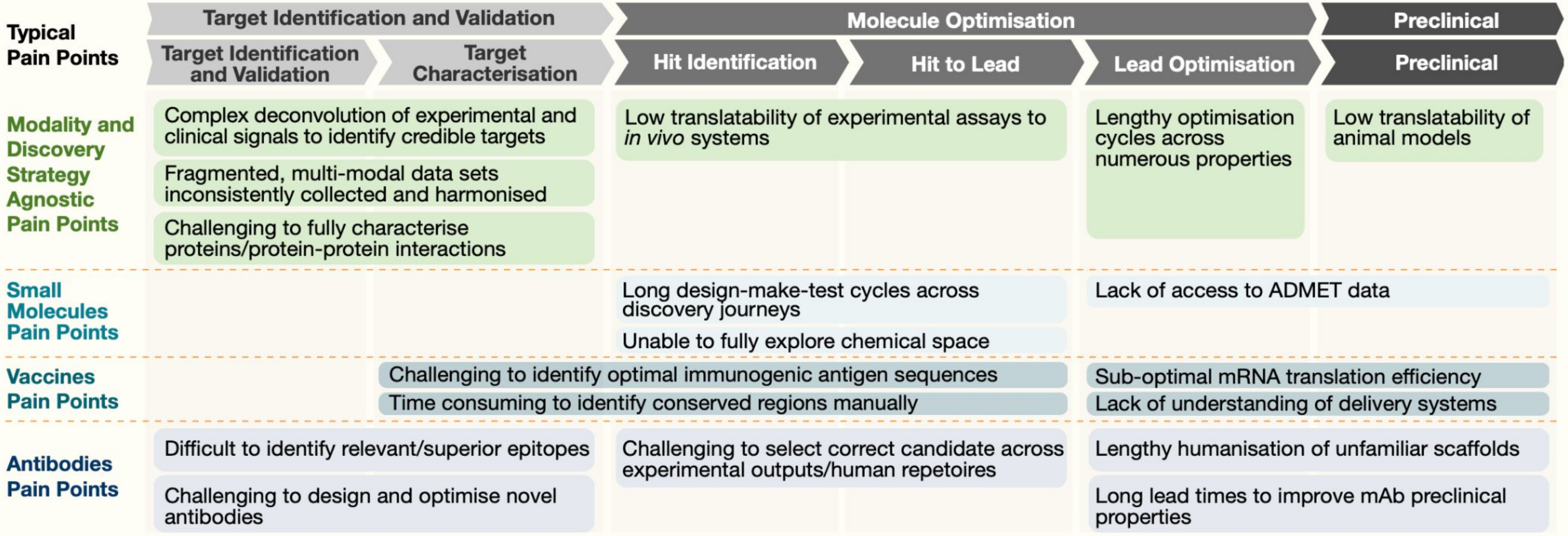
- Five major use case families identified: disease understanding, small molecule design, vaccine design, antibody design, and safety/toxicity evaluation

- Field growing rapidly: 34% annual growth in publications, 17% in patents over last 5 years

- Investment heavily concentrated in:
  - Three therapeutic areas: oncology, neurology, and COVID-19 (~70% of funding)
  - High-income countries and China
  - Understanding disease and small molecule optimization (80% of publications)

BOSTON
CONSULTING
GROUP

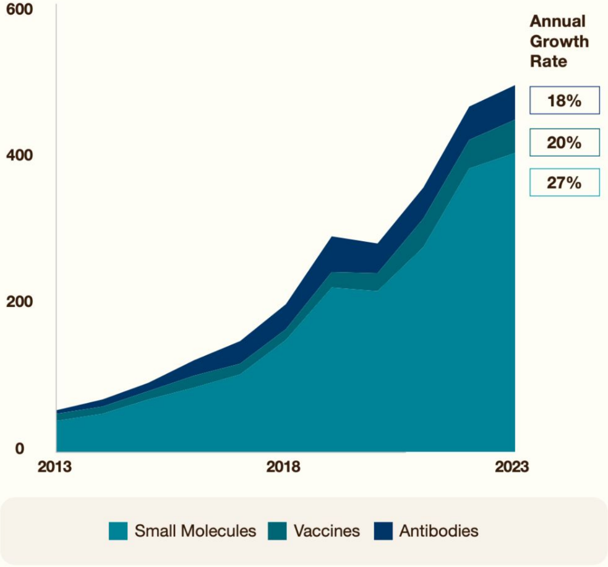
Figure 4 – Key pain points along the drug and vaccines discovery process today



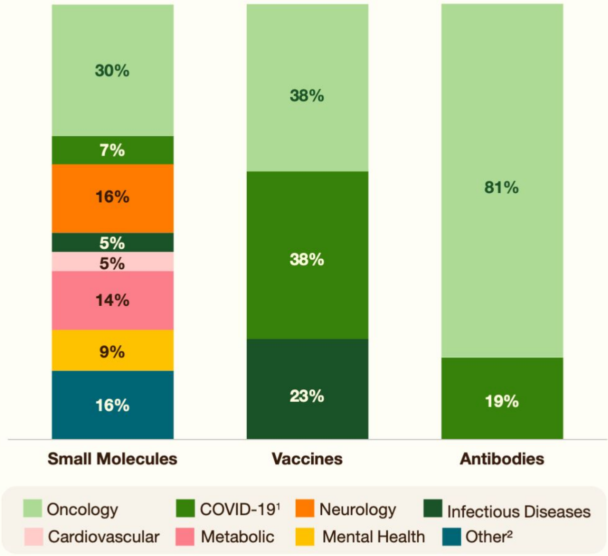
Typical Pain Points	Target Identification and Validation		Molecule Optimisation			Preclinical
	Target Identification and Validation	Target Characterisation	Hit Identification	Hit to Lead	Lead Optimisation	Preclinical
Modality and Discovery Strategy Agnostic Pain Points	Complex deconvolution of experimental and clinical signals to identify credible targets Fragmented, multi-modal data sets inconsistently collected and harmonised Challenging to fully characterise proteins/protein-protein interactions		Low translatability of experimental assays to *in vivo* systems		Lengthy optimisation cycles across numerous properties	Low translatability of animal models
Small Molecules Pain Points			Long design-make-test cycles across discovery journeys Unable to fully explore chemical space		Lack of access to ADMET data	
Vaccines Pain Points		Challenging to identify optimal immunogenic antigen sequences Time consuming to identify conserved regions manually			Sub-optimal mRNA translation efficiency Lack of understanding of delivery systems	
Antibodies Pain Points	Difficult to identify relevant/superior epitopes Challenging to design and optimise novel antibodies		Challenging to select correct candidate across experimental outputs/human repetoires		Lengthy humanisation of unfamiliar scaffolds Long lead times to improve mAb preclinical properties	

Figure 8 – Pipeline and therapeutic area focus of 'AI-first' biotechs

Rapidly growing 'AI-first' biotech pipelines mainly consist of small molecules…



… and focus on oncology and COVID-19[1], leaving other diseases underserved



[1] Numbers are one-off and not an ongoing trend

[2] Gastrointestinal, Immunology, Respiratory

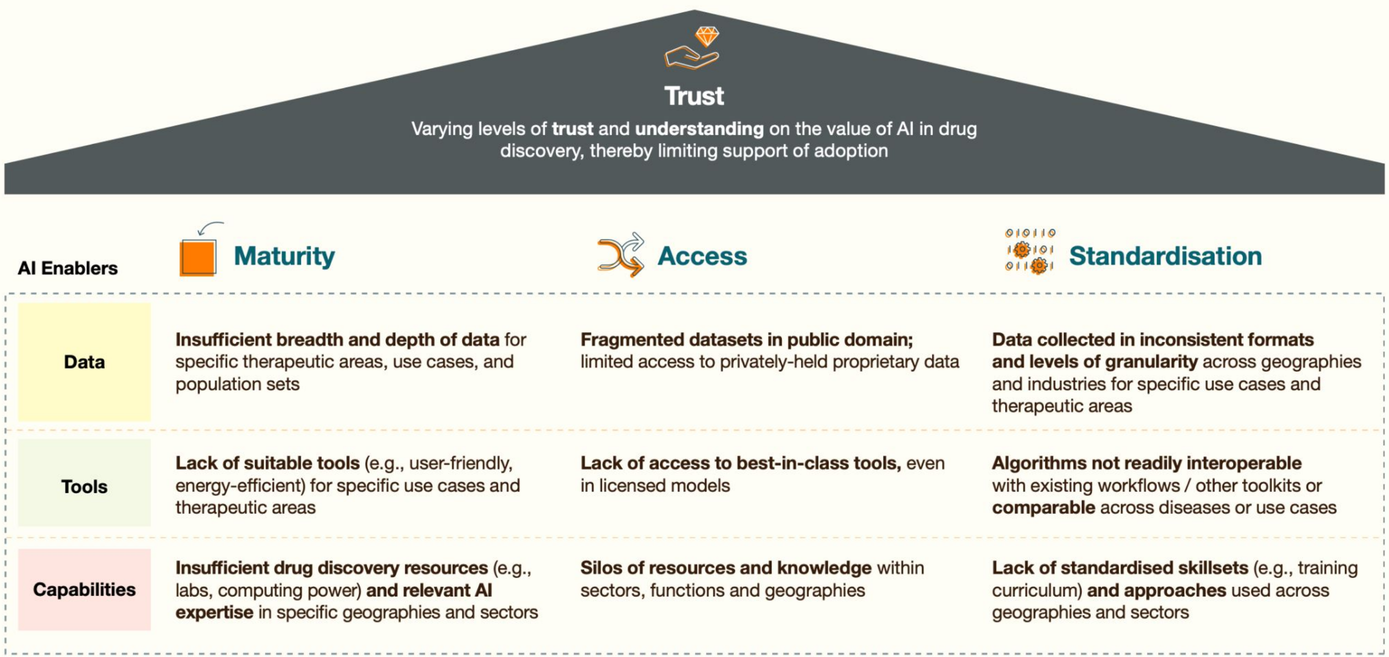
Current Challenges & Adoption:

- **Adoption varies significantly**:
  - Less than 1/3 of organizations use AI tools routinely
  - Higher adoption in industry vs academia
  - Higher in high-income countries (42%) vs low/middle-income countries (19%)

- **Key barriers to overcome**:
  - Trust and lack of proven value in drug discovery
  - Limited access to high-quality datasets
  - Shortage of interdisciplinary capabilities
  - Risk of amplifying health equity disparities in underserved therapeutic areas

- **Early results show promise**:
  - Potential for 25-50% time and cost savings in R&D through preclinical
  - Clinical trials will be crucial test for AI-developed drugs
  - Success in clinical trials could transform R&D economics

Figure 14 – Overview of barriers limiting the adoption of AI in drug discovery

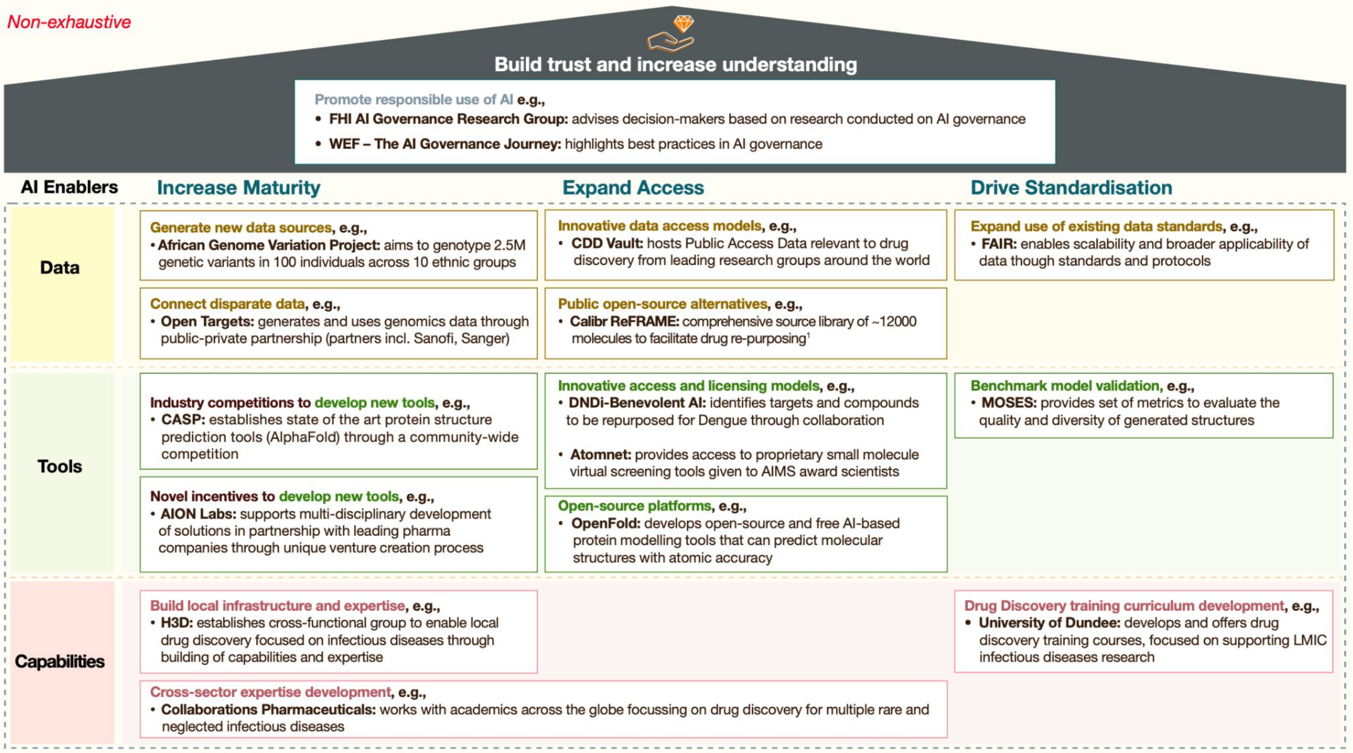


Trust

Varying levels of **trust** and **understanding** on the value of AI in drug discovery, thereby limiting support of adoption

AI Enablers	Maturity	Access	Standardisation
Data	**Insufficient breadth and depth of data** for specific therapeutic areas, use cases, and population sets	**Fragmented datasets in public domain;** limited access to privately-held proprietary data	**Data collected in inconsistent formats and levels of granularity** across geographies and industries for specific use cases and therapeutic areas
Tools	**Lack of suitable tools** (e.g., user-friendly, energy-efficient) for specific use cases and therapeutic areas	**Lack of access to best-in-class tools,** even in licensed models	**Algorithms not readily interoperable** with existing workflows / other toolkits or **comparable** across diseases or use cases
Capabilities	**Insufficient drug discovery resources** (e.g., labs, computing power) **and relevant AI expertise** in specific geographies and sectors	**Silos of resources and knowledge** within sectors, functions and geographies	**Lack of standardised skillsets** (e.g., training curriculum) **and approaches** used across geographies and sectors

Figure 19 – Examples of solutions already underway today to address barriers to adopting AI in drug discovery



Non-exhaustive

Build trust and increase understanding

Promote responsible use of AI **e.g.,**
- **FHI AI Governance Research Group:** advises decision-makers based on research conducted on AI governance
- **WEF – The AI Governance Journey:** highlights best practices in AI governance

AI Enablers	Increase Maturity	Expand Access	Drive Standardisation
Data	**Generate new data sources, e.g.,** • **African Genome Variation Project:** aims to genotype 2.5M genetic variants in 100 individuals across 10 ethnic groups **Connect disparate data, e.g.,** • **Open Targets:** generates and uses genomics data through public-private partnership (partners incl. Sanofi, Sanger)	**Innovative data access models, e.g.,** • **CDD Vault:** hosts Public Access Data relevant to drug discovery from leading research groups around the world **Public open-source alternatives, e.g.,** • **Calibr ReFRAME:** comprehensive source library of ~12000 molecules to facilitate drug re-purposing[1]	**Expand use of existing data standards, e.g.,** • **FAIR:** enables scalability and broader applicability of data though standards and protocols
Tools	**Industry competitions to develop new tools, e.g.,** • **CASP:** establishes state of the art protein structure prediction tools (AlphaFold) through a community-wide competition **Novel incentives to develop new tools, e.g.,** • **AION Labs:** supports multi-disciplinary development of solutions in partnership with leading pharma companies through unique venture creation process	**Innovative access and licensing models, e.g.,** • **DNDi-Benevolent AI:** identifies targets and compounds to be repurposed for Dengue through collaboration • **Atomnet:** provides access to proprietary small molecule virtual screening tools given to AIMS award scientists **Open-source platforms, e.g.,** • **OpenFold:** develops open-source and free AI-based protein modelling tools that can predict molecular structures with atomic accuracy	**Benchmark model validation, e.g.,** • **MOSES:** provides set of metrics to evaluate the quality and diversity of generated structures
Capabilities	**Build local infrastructure and expertise, e.g.,** • **H3D:** establishes cross-functional group to enable local drug discovery focused on infectious diseases through building of capabilities and expertise **Cross-sector expertise development, e.g.,** • **Collaborations Pharmaceuticals:** works with academics across the globe focussing on drug discovery for multiple rare and neglected infectious diseases		**Drug Discovery training curriculum development, e.g.,** • **University of Dundee:** develops and offers drug discovery training courses, focused on supporting LMIC infectious diseases research

How successful are AI-discovered drugs in clinical trials?

* AI-discovered molecules show potential to nearly double the success rate of clinical trials, potentially increasing end-to-end success rates from 5-10% to 9-18%, which would represent a significant improvement in pharmaceutical R&D productivity.

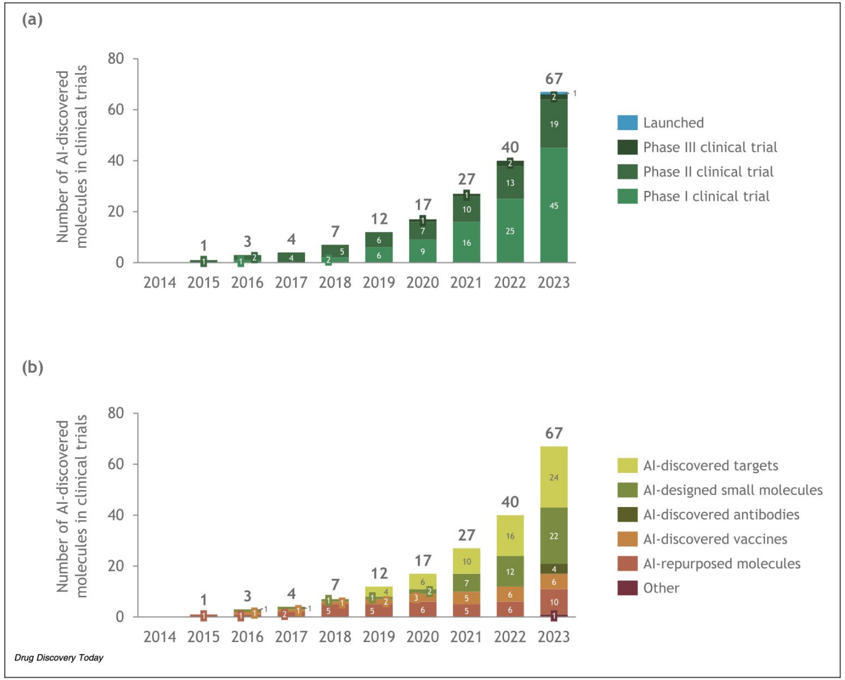
* This increased efficiency could allow pharmaceutical companies to either maintain current output levels with reduced resources and costs, or develop more new drugs using existing resources.

* Beyond current observations, AI techniques show promise for further improving clinical performance in Phase II and III trials, particularly through better disease understanding and target identification.

* Many biotech companies, pharmaceutical firms, and academic institutions are investing in AI applications for drug discovery, focusing on areas like OMICs data, reverse translation, patient-derived models, and using large language models to analyze disease data.

* The research shows AI is already demonstrating benefits in preclinical workflows by making the process faster and more cost-effective, with these advantages now beginning to appear in clinical trials as well.

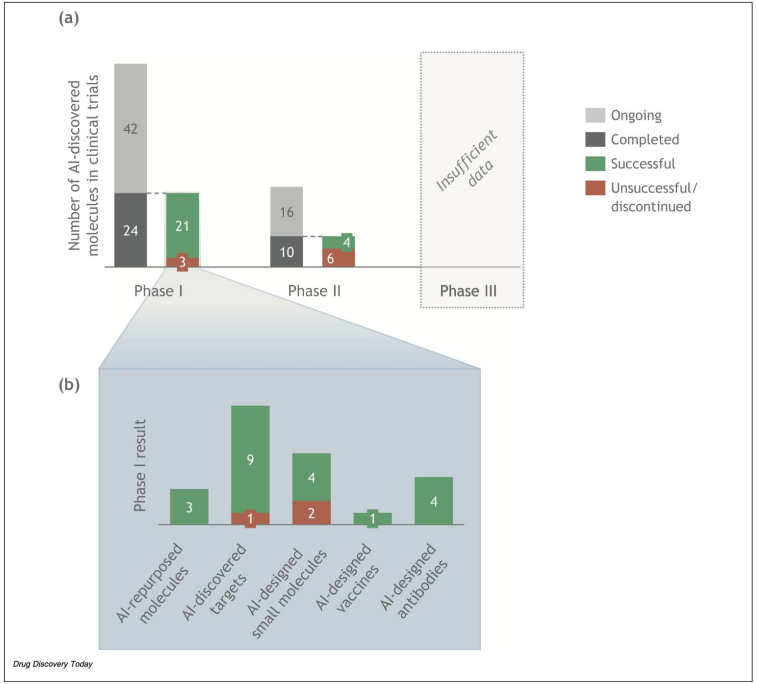
* While acknowledging statistical limitations, the researchers suggest these early results indicate a promising future for AI in pharmaceutical R&D, with more clinical results from AI-discovered molecules expected to emerge in coming years.



Drug Discovery Today

FIGURE 1

Number of molecules discovered by AI-first Biotechs that have entered clinical trials. The analysis includes molecules that were partnered with pharmaceutical companies and excludes COVID-19-related molecules. **(a)** AI-discovered molecules by clinical Phase. **(b)** AI-discovered molecules by mode-of-discovery.



Drug Discovery Today

FIGURE 2

The success of AI-discovered molecules in clinical trials so far. The analysis includes molecules that were partnered with pharmaceutical companies and excludes COVID-19-related molecules. **(a)** Clinical success of AI-discovered molecules by clinical Phase. **(b)** AI-discovered molecules that have completed Phase I trial, by mode-of-discovery.

That's the hype … now the counter view on the previous slides' data

What you will see is that in almost every case, **these targets were *already known* to be implicated in the disease under investigation**. In some of these examples, in fact there are several drugs already in the clinic targeting the same proteins, or even therapies that are already on the market working through the same mechanisms (*C. diff* toxin B, e.g.) I don't think any of these are bad targets, let me make that clear. **There are some really interesting things on the list, but I do not see how any of them can be classified as "target discovered by AI". I really don't.**

They say that of the 24 therapies that have reported Phase I, 21 were successful. The industry standard success rate for Phase I is 66% for all indications and 76% for lead indications, so while the AI-based examples here might be getting through Phase I at a higher rate, **both the sample size and outright differences are too small, in my opinion, to make that claim.** I would also add that few (if any) of these compounds have had any laying-on-of-hands for Phase I optimization, because for the most part no one knows any AI techniques to do that specifically (to the best of my knowledge).

Meanwhile, **four out of ten have succeeded in Phase II, which for what it's worth (an even smaller sample!) is exactly the same as for non-AI compounds**. And frankly, this is where I'd expect the numbers to be the same, because Phase II success is all about picking the right targets, and (as mentioned above) **these were all already targets that people go interested in the old-fashioned way, not because the AI picked them out of the ether and said to go for them. Why shouldn't they fail at the same rate as everybody else's stuff?**

https://www.science.org/content/blog-post/ai-drugs-so-far

Setting the right expectations …

If you take the hype and PR at face value over the last 10 years, you would think it goes from five percent to 90 percent. But if you know how these models work, it goes from five percent to maybe six or seven percent.

Patrick Malone, KdT Ventures

Group discussion: What are the implications of this?

Real world adoption challenges

Navigating scepticism and embracing AI

- AI has the ability to accelerate processes and improve efficiency that could offer immense benefits for patients, healthcare, and economies, but there is **widespread nervousness in the industry**, particularly among those who are not familiar with the technical side of AI.
- The resistance often **stems from a lack of understanding, compounded by the complexity of interdisciplinary communication**. People know an area very well—whether it's biology, chemistry, or pharmacology—but venturing outside that comfort zone into tech can be daunting.
- ACTION: To overcome this scepticism, **a stepwise approach is recommended**, starting with small pilot projects that demonstrate tangible benefits, which are communicated in a suitable language across the organisation, and which are paired with facts-based and honest education. Quick wins can help build trust in new technologies, easing nervousness and encouraging broader adoption.

Communication: the key to overcoming barriers

- Communication between scientific and technical teams is paramount to adopting new technologies.
- The convergence of disciplines, such as biology and AI, often leads to breakthroughs. However, **the difficulty lies in ensuring both sides understand each other**, with miscommunication increasing nervousness.
- To bridge this gap, **agreeing on a common language and breaking down complex concepts are essential steps** - the modeller needs to understand the variables that represent input and output into the model sufficiently well, while the wet lab scientist needs to grasp essentials of 'AI-ready data', hence forming a productive interface across the team.
- ACTION: A long-term relationship and understanding of the company's specific needs is necessary, with **technical vendors needing to take a big-picture view of how they can offer value beyond immediate solutions and not offer generic solutions that may not align with a company's actual milestones or challenges**.

Real world adoption challenges (con't)

Understanding biotech's board commitments

- From a business perspective, **biotech companies are faced with the pressure of meeting board-driven milestones**.
- These milestones—such as hit identification, lead optimization, and candidate nomination—are critical for securing investment and increasing company valuation.
- In practice, they can be both a constructive driver, or become destructive and a distraction, depending on the particular situation.

Managing scientific data complexity in small biotechs

- Many early-stage biotech companies are highly data-driven, in particular those focusing on a particular technology enabler to generate said data, producing vast amounts of experimental data through cell-based assays, phenotypic screening, imaging analyses and various biological assays.
- However, **storing, managing, analysing, and leveraging this data is no easy task, especially for a small company with limited resources**.
- This challenge is compounded by the increasing complexity of biological data, particularly in imaging, which often contains hidden insights that are difficult to extract without advanced computational tools.
- In addition to data management, small companies may not have the expertise or resources to fully utilise the rich datasets they generate.
- **This reflects a broader industry trend in which smaller biotechs struggle to fully capitalise on their experimental results**.

Real world adoption challenges (con't)

The untapped potential of automation and AI

- Automation has the potential to make basic tasks like measuring key parameters faster.
- Furthermore, AI is intrinsically able to provide insights into complex data patterns that humans may overlook, offering a force multiplier effect for companies with limited manpower. However, t**he practical implementation of these technologies remains challenging.**
- Many small biotechs do not have the resources to invest in sophisticated AI tools or extensive automation systems.
- PARADOX: while AI and automation could revolutionise their operations, they often require significant upfront investment that small companies cannot easily afford. **Some companies who do have automation as their main focus often tend to not have sufficient capability to translate their capabilities into value in projects, due to lack of capabilities in more translational areas.**

Data as Intellectual Property

- Data generated by biotechs holds significant intellectual property (IP) value, in particular data that is unique to an organisation that can be used to generate unique competitive insights. However, which data should be generated, and how? How is translation into later (such as clinical) stages performed? How does one know which data is needed in the future, and how are compatibilities in experimental setups addressed? (etc.)
- DILEMMA: **how to generate, store and manage data so that it remains useful in the long term, even as the company's resources grow and new technologies become available.**
- In big pharma these challenges are often addressed through large-scale projects with significant (but often still insufficient) funding.
- But smaller biotechs must navigate the same complex data issues with far fewer resources, making strategic decisions about what to prioritise and how to maximise the long-term value of their data and its utilisation.

TOPIC 2

ML, Knowledge Graph, LLM, …

Concepts:

Rule-driven vs stochastic systems

Prediction vs Inference systems

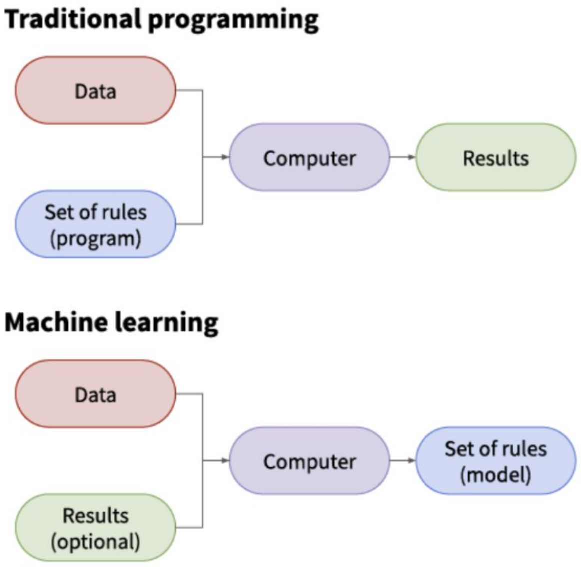
Relationships between entities

Reasoning systems

Feature selection

Provenance

What is Machine Learning



- Linear regression: an algorithm, but not a model
  - Input: A dataset that contains examples of a potential relationship between a dependent variable and one or more independent variables
  - Output: The coefficients of the linear equation that best approximates a relationship between the inputs and outputs (aka a linear regression model)
- A linear regression model: an algorithm and a model
  - Input: One or more independent variables
  - Output: A prediction of a dependent variable

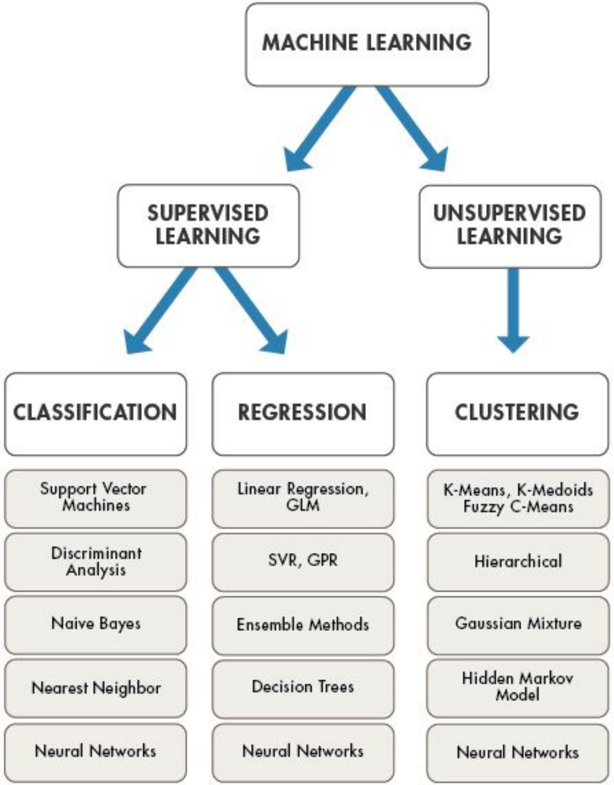
In practice, it's common to use the same term to refer to the learning algorithm and the machine learning model, i.e. the term "linear regression" may refer to the technique of finding the aforementioned optimal line, or to the line (the actual model) itself.

High-level breakdown of ML methods

There are three main machine learning paradigms: supervised learning, unsupervised learning, and reinforcement learning.

- Supervised learning: The process of using labeled data to learn relationships between features of the data, in order to make predictions about unseen or future data.
- Unsupervised learning: The process of using unlabeled data to learn relationships between features of the data, in order to extract meaningful information from the data.
- Reinforcement learning: The process of using actions within an environment to learn properties of the environment, in order to determine the best actions to take.

Other types of learning include paradigms that are a combination of supervised and unsupervised learning, like semi-supervised learning and self-supervised learning, and broad concepts, like transfer learning and online learning.



MACHINE LEARNING

SUPERVISED LEARNING

UNSUPERVISED LEARNING

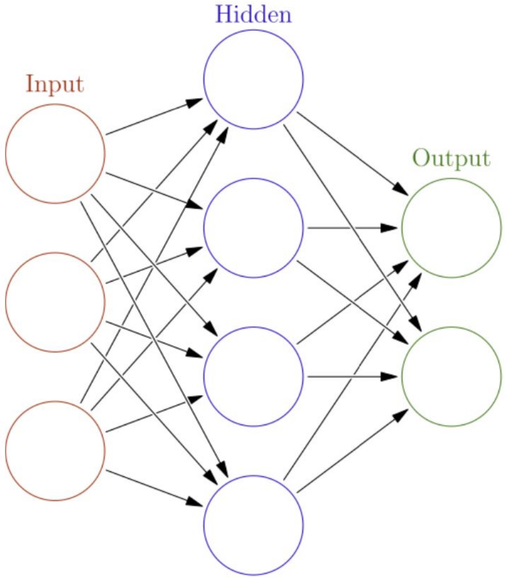
CLASSIFICATION	REGRESSION	CLUSTERING
Support Vector Machines	Linear Regression, GLM	K-Means, K-Medoids Fuzzy C-Means
Discriminant Analysis	SVR, GPR	Hierarchical
Naive Bayes	Ensemble Methods	Gaussian Mixture
Nearest Neighbor	Decision Trees	Hidden Markov Model
Neural Networks	Neural Networks	Neural Networks

What is Deep Learning

The term “deep learning”, however, refers to algorithms that are structured and implemented as a neural network with many layers. Deep learning models can perform supervised, unsupervised, and reinforcement learning tasks.

A neural network is organized into layers of artificial neurons (also called “perceptrons” or simply “neurons”), where the neurons of a layer are connected to the neurons of another layer, in the sense that the output(s) of a neuron in one layer become (part of) the input to a neuron in another layer.

These connections go in one direction, from the first to the last layer of the network via some number of “hidden” layers in between. The first and last layers are known as the input and output layers of the network, respectively.



This graphic from Wikipedia’s article on artificial neural networks illustrates the structure of a neural network, with each circle representing a neuron, each circle color representing a different layer, and each arrow representing a connection from one neuron to another.

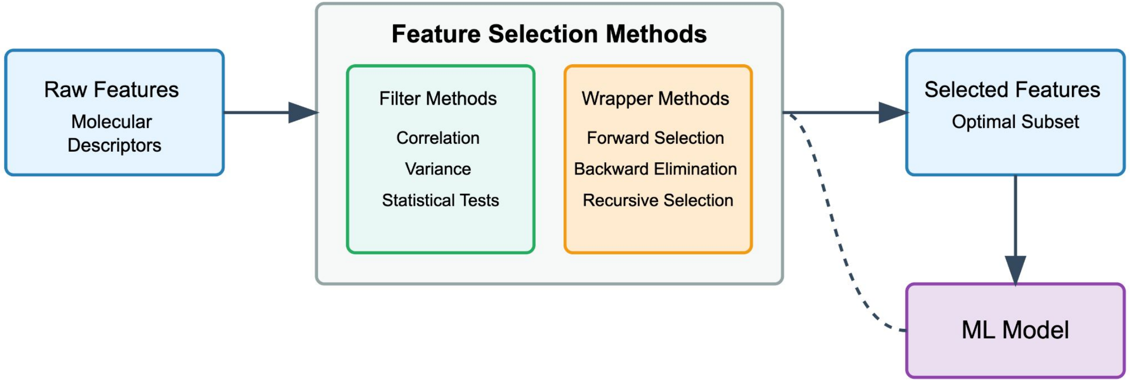
What is Feature Selection

Feature selection systematically reduces high-dimensional molecular descriptor data (often 1000s of features) to the most informative subset, critical for drug discovery where descriptor spaces are vast. This improves model interpretability and reduces computational complexity.

Filter methods evaluate features independently using statistical measures. In drug discovery, this helps identify molecular descriptors that strongly correlate with biological activity while removing redundant or noisy features that could mislead the model.

Feature selection helps prevent overfitting in drug discovery models by removing spurious correlations in molecular descriptors, particularly important when working with limited experimental data typical in early-stage drug development.

The process enables better model interpretability by identifying key molecular properties that drive biological activity, providing actionable insights for medicinal chemists in lead optimization.



Why Feature Selection is needed

Curse of dimensionality - Drug molecules often have thousands of descriptors (features), but limited experimental data points. This imbalance causes models to fit noise rather than true structure-activity relationships, leading to poor generalization on new compounds.

Computational inefficiency - Training ML models with irrelevant molecular descriptors wastes computational resources and increases training time. For large-scale virtual screening, this inefficiency becomes a significant bottleneck.

Reduced interpretability - When models use too many features, identifying which molecular properties actually drive biological activity becomes nearly impossible. This hinders medicinal chemists from making informed decisions in lead optimization.

Data redundancy - Many molecular descriptors are highly correlated (e.g., different ways of measuring molecular size). Without feature selection, these redundancies can bias model predictions and obscure important structure-activity patterns.

Higher validation requirements - Models with excessive features need larger validation sets to ensure reliability, which is particularly problematic in early-stage drug discovery where experimental data is scarce and expensive to obtain.

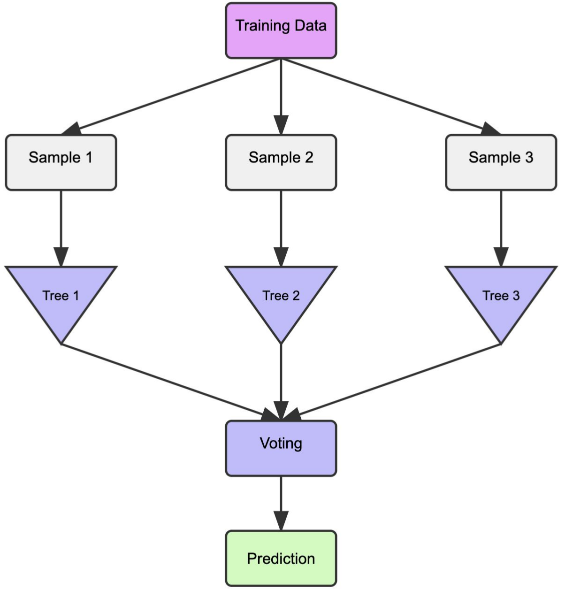
What is Random Forest

Random Forest is an ensemble machine learning method that combines multiple decision trees to create a more robust and accurate model.

1. **Multiple Trees**: The algorithm creates many decision trees, each trained on a random subset of the training data (this is called "bagging" or bootstrap aggregating).

2. **Feature Randomization:** For each tree, only a random subset of features is considered at each split point. This ensures that the trees are different from each other and reduces overfitting.

3. **Voting/Averaging**: For prediction, all trees in the forest "vote" on the outcome. For classification, it uses majority voting; for regression, it averages the predictions.



What is a Knowledge Graph

Knowledge graphs are structured representations of information that show how different entities (like people, places, things, or concepts) are connected through various relationships. They're particularly powerful in scientific data analysis because they can capture complex relationships and allow for sophisticated querying and inference.

Knowledge graphs are particularly valuable in scientific data analysis:

Integration of Heterogeneous Data

- Knowledge graphs excel at integrating data from multiple sources and domains. In scientific research, this means connecting information from publications, experimental data, clinical trials, and various databases into a unified structure.

Discovery of Hidden Relationships

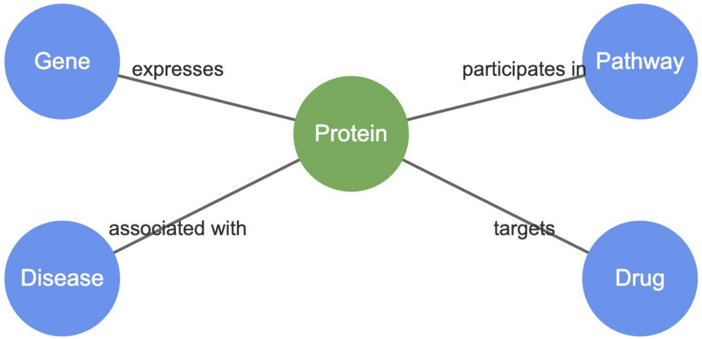
- They enable researchers to discover non-obvious connections between entities. For example, in drug discovery, a knowledge graph might reveal that a drug developed for one condition could potentially treat another disease due to shared biological pathways.

Contextual Understanding

- Knowledge graphs preserve the context of relationships, which is crucial in scientific analysis. For instance, a protein's role might change depending on its cellular location or the presence of other molecules.

Inference and Prediction

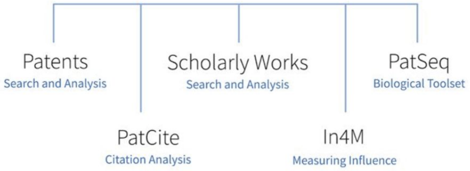
- Using the connected nature of the graph, researchers can make predictions about unknown relationships based on existing patterns. This is particularly useful in areas like:
  - Drug repurposing
  - Disease mechanism understanding
  - Protein function prediction
  - Gene-disease associations



Example of Knowledgegraph technology: Patent discovery

LENS.ORG

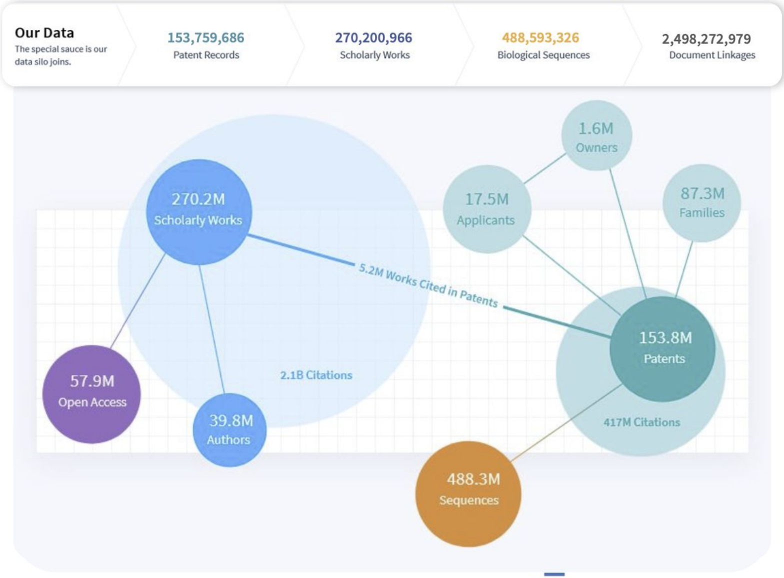
Solving The Problem Of Problem Solving ™



DATA SILO LINKAGES

Our **Special Sauce**

153M	Patent Records
87M	Patent Families
17M	Applicants
1.6M	Owners
270M	Scholarly Works
39M	Authors
25k	Research Institutions
5.2M	Works cited in patents
2.1B	Scholarly citations
488M	Bio Sequences
2.5B	Document Linkages



Our Data
The special sauce is our data silo joins.

153,759,686	270,200,966	488,593,326	2,498,272,979
Patent Records	Scholarly Works	Biological Sequences	Document Linkages

LENS.ORG

Tips for
Using free ChatGPT & Claude

Avoiding usage limits with free ChatGPT (or Claude) and PDF papers:

Preparation

1. **Divide the Paper into Sections**:
   - Split the paper into manageable sections (e.g., Abstract, Introduction, Methods, Results, Discussion, Conclusion, etc.).
   - Save each section as a separate text file or clearly label them for sequential upload.
2. **Identify Key Questions**:
   - Determine the specific questions or insights you want to extract from the paper (e.g., summary of methods, key findings, limitations).
   - Craft a few targeted prompts for each section to guide the analysis.
3. **Compress the Text** (Optional):
   - Focus on relevant parts of the paper by removing unnecessary details (e.g., appendices, extended data tables, or references).
   - Use summarization tools to reduce the length of very detailed sections.

Uploading and Prompting

4. **Upload in Segments**:
   - Paste or upload one section at a time.
   - Use clear instructions like: "Please summarize the **Methods** section of this paper and highlight the main techniques used."
5. **Use Specific Prompts**:
   - Provide context for the analysis:
     - Example: *"This is the **Introduction** of a paper on protein engineering. Please summarize the research objectives and the rationale for the study."*
6. **Request Follow-up Summaries**:
   - After analyzing individual sections, ask for a summary or synthesis of all sections combined.
     - Example: *"Based on the previously analyzed sections, summarize the paper's key findings and implications."*
7. **Deep Dive (if needed)**:
   - For particularly complex or dense sections, ask for further clarification or detail:
     - Example: *"Can you explain the significance of the findings in the Results section in simpler terms?"*

Avoiding usage limits with free ChatGPT (or Claude) and web pages:

Why Web Pages Save Tokens

1. **Selective Extraction**:
   - When providing a link to a web page, you can ask ChatGPT to analyze only specific parts of the page. This avoids including irrelevant content like headers, footers, navigation menus, or references.
   - Example: *"Please summarize only the Abstract and Results sections from this page."*
2. **URL-Based Analysis** (with tools):
   - If tools like `web.search` or other plugins are available, they often fetch and process the core content more efficiently by removing extraneous page elements.
3. **Avoiding Pasted Overhead**:
   - When copying and pasting the text, additional formatting artifacts (like extra line breaks, symbols, or headings) may inadvertently inflate token usage. Using a clean web page avoids this.

Steps for Efficient Web Page Analysis

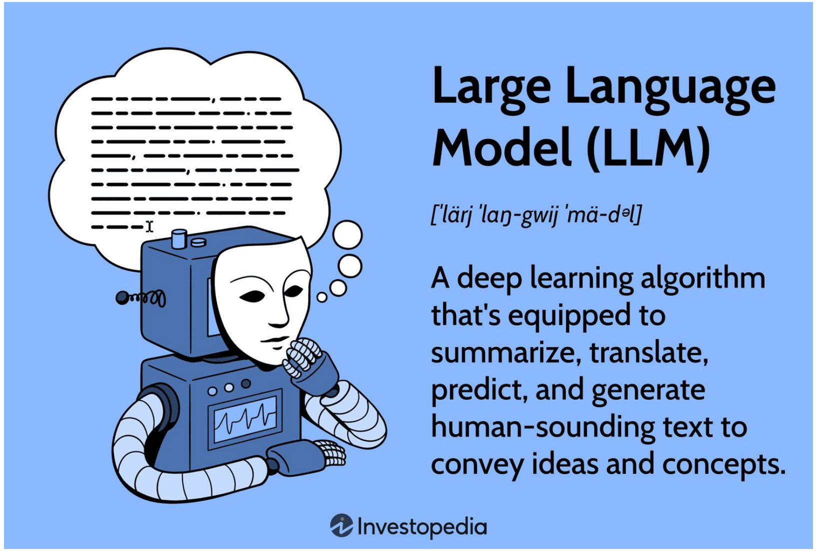
1. **Check Content Availability**:
   - Ensure the full text is accessible on the web page and not behind a paywall or login.
2. **Provide Specific Prompts**:
   - Example: *"Analyze the research paper at this link and summarize the Abstract, Methods, and Results sections."*
3. **Request a Content Filter**:
   - Example: *"Please skip references, tables, and figure captions while summarizing the content from this page."*
4. **Iterative Analysis**:
   - If the paper is long, break the task into parts:
     - Example: *"From this link, summarize the Abstract and Introduction first."*
     - Then, follow up with: *"Now summarize the Results and Discussion sections."*

Quick Activity

Using ChatGPT, explore how the cancer paper is using the Decision Tree method.

https://www.mdpi.com/2077-0383/13/8/2177

LLM in Drug Discovery??

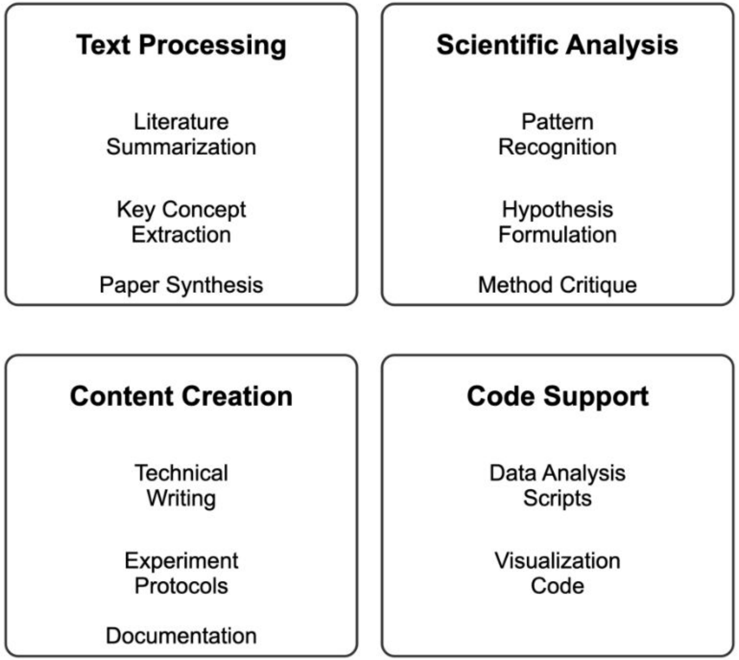


Large Language Model (LLM)

['lärj 'laŋ-gwij 'mä-dᵊl]

A deep learning algorithm that's equipped to summarize, translate, predict, and generate human-sounding text to convey ideas and concepts.

Investopedia



Text Processing

Literature Summarization

Key Concept Extraction

Paper Synthesis

Scientific Analysis

Pattern Recognition

Hypothesis Formulation

Method Critique

Content Creation

Technical Writing

Experiment Protocols

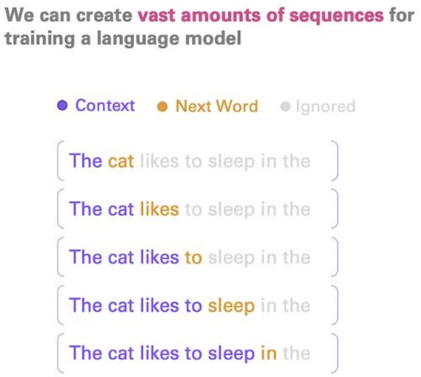
Documentation

Code Support

Data Analysis Scripts

Visualization Code

Training an LLM … to predict the next word in sequence



We can create **vast amounts of sequences** for training a language model

● Context ● Next Word ● Ignored

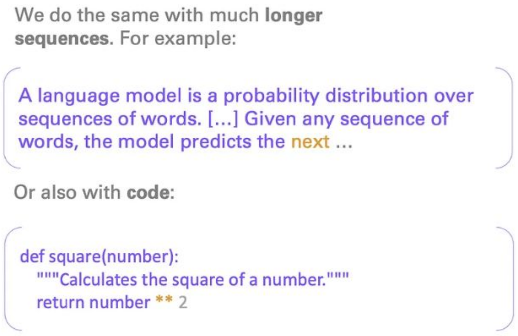
The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the

The cat likes to sleep in the



We do the same with much **longer sequences**. For example:

A language model is a probability distribution over sequences of words. [...] Given any sequence of words, the model predicts the next ...

Or also with **code**:

```
def square(number):
    """Calculates the square of a number."""
    return number ** 2
```

And as a result – the model becomes incredibly good at predicting the next word in any sequence.

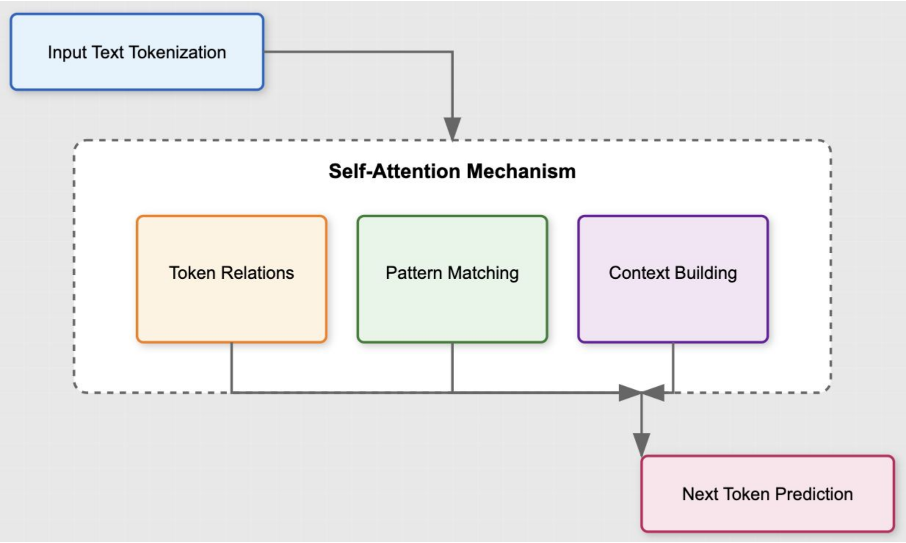
Massive amounts of traning data can be created relatively easily.

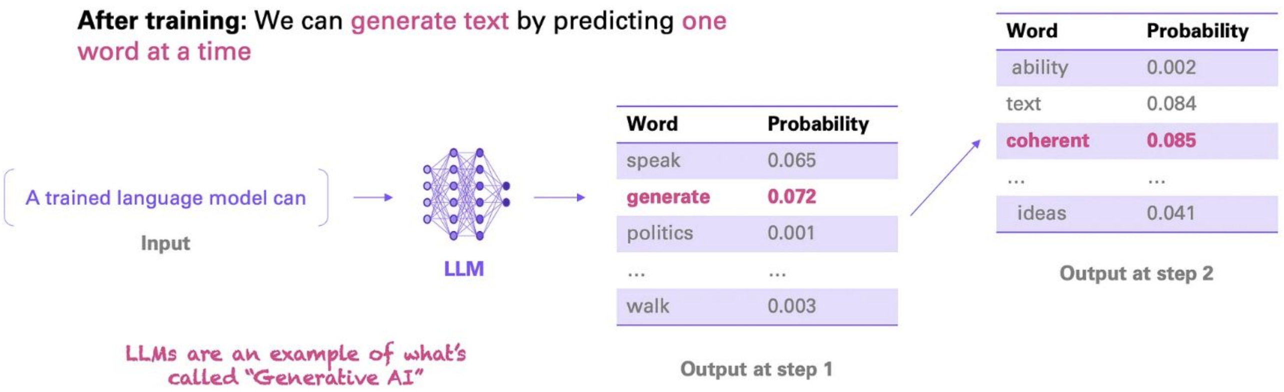
All we are doing here is to train a neural network (the LLM) to predict the next word in a given sequence of words, no matter if that sequence is long or short, in German or in English or in any other language, whether it’s a tweet or a mathematical formula, a poem or a snippet of code. All of those are sequences that we will find in the training data.

If we have a large enough neural network as well as enough data, the LLM becomes really good at predicting the next word.

Generating Natural Language

We can feed the extended sequence back into the LLM and predict another word, and so on. In other words, using our trained LLM, we can now generate text, not just a single word. This is why LLMs are an example of what we call Generative AI. We have just taught the LLM to speak, so to say, one word at a time.





After training: We can generate text by predicting one word at a time

Word	Probability
speak	0.065
generate	**0.072**
politics	0.001
...	...
walk	0.003

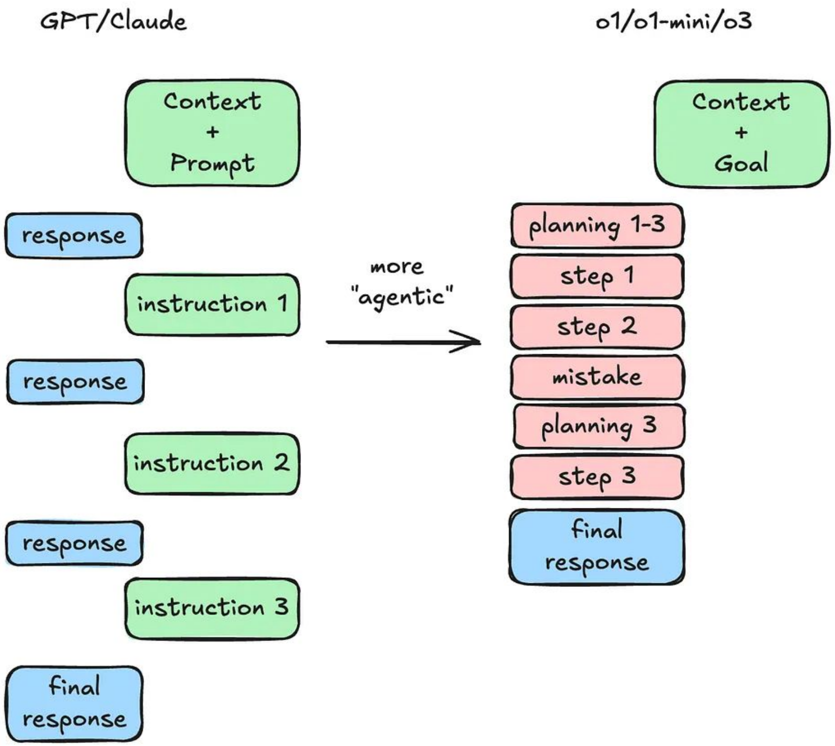
Output at step 1

Word	Probability
ability	0.002
text	0.084
coherent	**0.085**
...	...
ideas	0.041

Output at step 2

LLMs are an example of what's called "Generative AI"

Evolution of LLMs



https://www.latent.space/p/o1-skill-issue

Pragmatic Guidelines to using LLMs

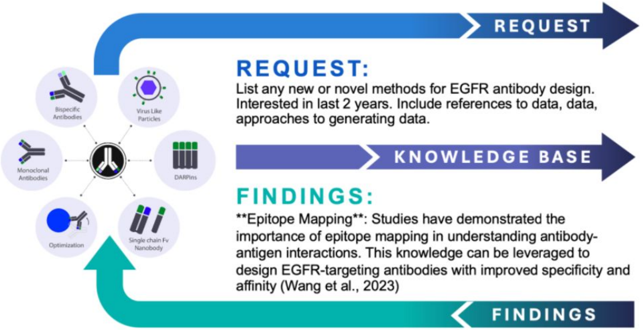
These recommendations reflect the need to upload your own .pdf and .csv files, when using ChatGPT:

- Where measurement data is used in your uploaded documents, label them as variables in your queries.
- In your queries, do not embed statements within statements – bring everything out in simple single purpose statements.
- Where ChatGPT stops because of long answers, break up the prompts and ask it to export to .csv or .pdf instead of printing on the screen. If it still stops part way through, type “Please continue from where you left off and finish the answer.”
- When the results are inaccurate and you need to re-run the query, try switching to a new ChatGPT window/chat. Another trick is to clear short-term memory within a prompt workflow, using this command “From now on, assume [new context] without considering our previous conversation.” For example, “From now on, assume we are talking about human-only clinical trials without considering our previous conversation.”

Pragmatic guidelines when deciding what LLM system to use

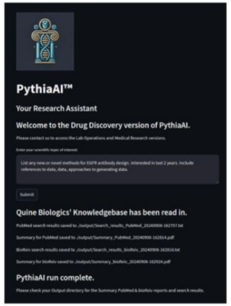
- LLM systems are made up of LLM models (for example, OpenAI and Claude) and software frameworks (for example, Autogen Studio and Open Web UI). Both these technology types are evolving rapidly, with multiple offerings. Any system you select needs to be adaptable to technology changes, for example if an effective drug discovery LLM model is offered as open-source next year.
- In addition to a rapidly evolving technology landscape, it is important to understand the changing balance between commercial and open-source offerings. As with other industries, these words of wisdom stand (relatively) true: “Wait a year, and someone will offer it for free!” It is important to be agile and respond to changes in this balance.
- Your LLM system may need to support both research-intensive activities, such as extracting new insights from large amounts of (.pdf) research publications, and workflow-driven tasks, such as designing a lab experiment. These are different problems that need to be solved with different approaches. A useful tip here is to design your system with a human performing the proposed LLM functions, and then swap the LLM technology in for the human once the design is ready.

Example use in Antibody Engineering



REQUEST:
List any new or novel methods for EGFR antibody design. Interested in last 2 years. Include references to data, data, approaches to generating data.

FINDINGS:
Epitope Mapping: Studies have demonstrated the importance of epitope mapping in understanding antibody-antigen interactions. This knowledge can be leveraged to design EGFR-targeting antibodies with improved specificity and affinity (Wang et al., 2023)



PythiaAI™

Your Research Assistant

Welcome to the Drug Discovery version of PythiaAI.

Please contact us to access the Lab Operations and Medical Research versions.

Enter your scientific topic of interest:

List any new or novel methods for EGFR antibody design. Interested in last 2 years. Include references to data, data, approaches to generating data.

Quine Biologics' Knowledgebase has been read in.

PubMed search results saved to ./output/Search_results_PubMed_20240908-162757.txt

Summary for PubMed saved to ./output/Summary_PubMed_20240908-162814.pdf

BioRxiv search results saved to ./output/Search_results_bioRxiv_20240908-162816.txt

Summary for bioRxiv saved to ./output/Summary_bioRxiv_20240908-162924.pdf

PythiaAI run complete.

Please check your Output directory for the Summary PubMed & bioRxiv reports and search results.

Figure 2: Description of high-level workflow using PythiaAI™ (right) with the Quine computational system (left), including the scientific question (Request) and output (Findings)

PythiaAI™ operates as follows:

1. Hypothesis-Driven Search: Scientists can input domain-specific knowledge and ask targeted questions, such as identifying novel methods for EGFR antibody design. The LLM refines the search space by curating relevant data from sources like PubMed, bioRxiv, and proprietary research archives.
2. Business Logic Integration: By combining scientific queries with business considerations, such as determining which diseases to target based on market size or strategic opportunity, the LLM helps researchers make informed decisions. For example, should the scientist focus on rare diseases with high unmet needs, or pursue broader targets that are already under investigation by larger pharmaceutical companies?
3. Curation of Unstructured Data: LLMs convert unstructured data into actionable insights. For instance, in the context of EGFR antibody design, the system might provide an output such as **epitope mapping** — identifying novel protein regions that haven't been explored or offering new perspectives on existing ones.
4. Augmenting Scientific Expertise: The LLM acts as a virtual research assistant, assisting scientists in areas where they may not have deep expertise. For example, if the researcher is focused on antibody engineering but lacks the medical knowledge to identify the most promising disease targets, the platform can fill that gap by sifting through clinical data and providing suggestions on lesser-known protein targets with therapeutic potential.

COMPARISION: Core Characteristics and Relationships

Aspect	Artificial Intelligence (AI)	Natural Language Processing (NLP)	Large Language Models (LLMs)
Definition	The broader field of creating systems that can simulate intelligent behavior	A subfield of AI focused on enabling computers to understand and process human language	Neural network architectures specifically designed to process and generate human language at scale
Scope	Encompasses all forms of machine intelligence, including vision, robotics, expert systems, and decision-making	Focuses specifically on computational linguistics and language-related tasks	Specializes in language understanding and generation using transformer-based architectures
Historical Development	Emerged in the 1950s with early computing; evolved through symbolic AI, expert systems, and machine learning	Developed in the 1960s, initially rule-based, then statistical, now neural-based	Emerged in 2017 with the transformer architecture; rapid evolution through BERT, GPT series, etc.
Key Technologies	Includes machine learning, neural networks, expert systems, genetic algorithms, and robotics	Includes tokenization, parsing, word embeddings, and various ML algorithms	Primarily based on transformer architecture with self-attention mechanisms
Primary Applications	Autonomous systems, game playing, pattern recognition, problem solving, and general intelligence tasks	Language translation, sentiment analysis, text classification, information extraction	Text generation, conversation, code writing, creative tasks, and complex reasoning
Data Requirements	Varies by application; can use structured or unstructured data	Typically requires annotated linguistic data and text corpora	Requires massive amounts of text data for pre-training
Processing Approach	Can be rule-based, statistical, or neural, depending on the application	Combines linguistic rules with statistical and neural methods	Primarily neural, using attention mechanisms and deep learning
Strengths	Versatile problem-solving capabilities across domains	Specialized language understanding and task-specific performance	Powerful language generation and transfer learning abilities
Limitations	May struggle with common-sense reasoning and generalization	Can be brittle and domain-specific	Resource-intensive, potential for bias, lack of true understanding
Relationship to Others	Parent field encompassing both NLP and LLMs	Broader field that includes LLMs as one implementation approach	Specialized implementation that advances both AI and NLP goals

COMPARISON: Key Differentiating Features

Feature	AI	NLP	LLMs
Learning Approach	Multiple paradigms (supervised, unsupervised, reinforcement)	Often task-specific supervised learning	Primarily self-supervised pre-training with optional fine-tuning
Scale	Varies by application	Typically moderate-scale models	Extremely large-scale models with billions/trillions of parameters
Resource Requirements	Varies widely	Moderate computing resources	Substantial computing resources for training and inference
Interpretability	Varies by method; some highly interpretable	Often provides linguistic interpretability	Generally black-box with limited interpretability
Real-world Integration	Widely deployed in specific applications	Common in specialized language tasks	Increasingly deployed as general-purpose tools

COMPARISON: LLM vs Knowledge Graph

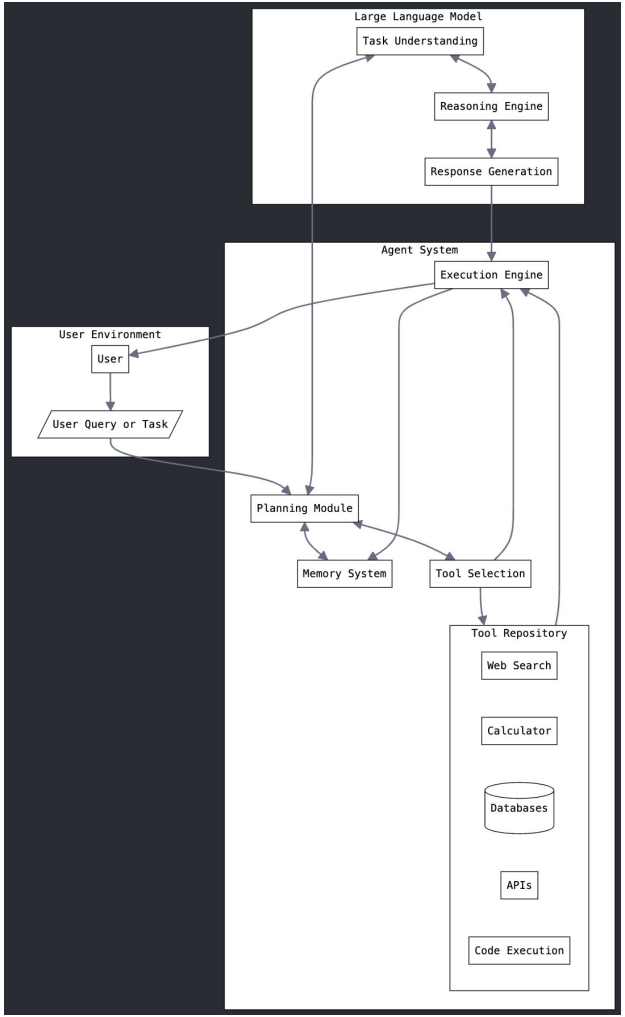
	Semantic RAG) LLM System	Knowledge Graph (with Ontology)
Ease of App Development	Easier to implement initially	More complex initial setup, requires ontology design
Data Integration	Good for diverse data types, including unstructured (PDFs) and structured (Excel)	Excellent for integrating diverse data types, provides unified perspective
Information Retrieval	Efficient retrieval based on semantic similarity	Powerful retrieval based on relationships and context
Accuracy	Can be prone to hallucinations	Reduces hallucinations due to well-defined structure
Scalability	May face challenges with large-scale data	Highly scalable, efficient for complex queries
Reasoning Capabilities	Limited to pattern matching and similarity	Strong logical reasoning and inference capabilities
Maintenance	Requires regular updates to vector database	Easier to maintain and update with new information
Cost Efficiency	May have higher costs for vector database scaling	More cost-efficient in the long run
User Experience	Provides relevant information quickly	Offers more context-aware and interconnected results
Explainability	Limited explainability of results	Higher explainability due to transparent relationships

LLM Agents

LLM agents are about autonomous decisions and tools.

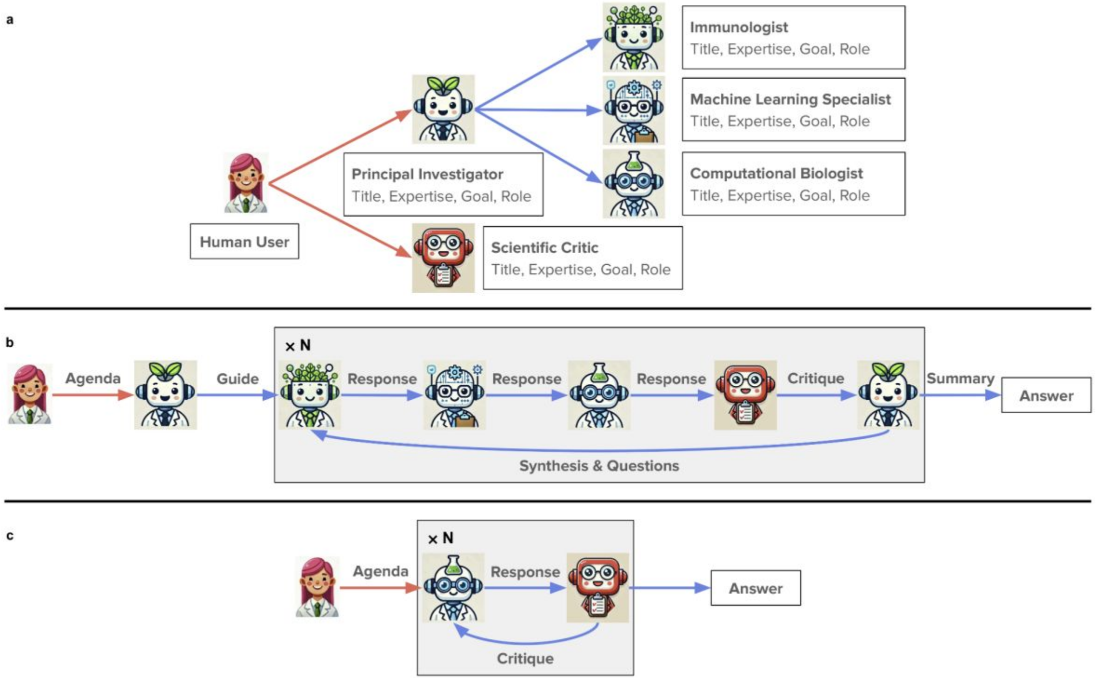
This acts as a force-multiplier on the use of LLMs.

An LLMs are force-multipliers for leveraging unstructured knowledge.



LLM Agents example … this is cool

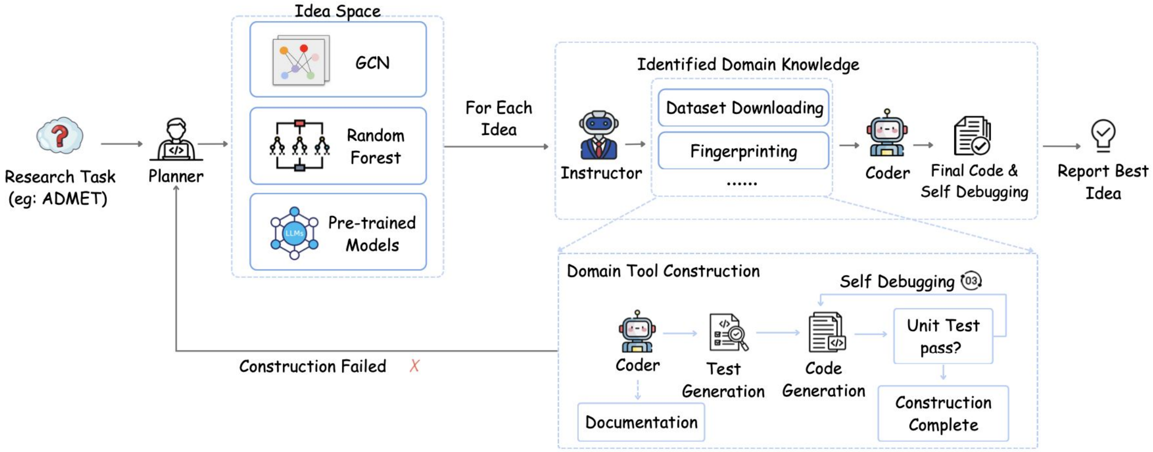
The Virtual Lab: AI Agents Design New SARS-CoV-2 Nanobodies with Experimental Validation



https://www.biorxiv.org/content/10.1101/2024.11.11.623004v1

Another example …

DrugAgent: Automating AI-aided Drug Discovery Programming through LLM Multi-Agent Collaboration



Framework overview of DrugAgent. Given an AI-based drug discovery task described in natural language (i.e., user's input, e.g., design an AI model to predict Absorption (one of the ADMET properties) using the PAMPA dataset, the LLM Planner first produces a couple of potential ideas (e.g., GCN (graph convolutional network), random forest, pretrained model (such as ChemBERTa)).

Then, for each idea, the LLM Instructor transfers the idea into code based on domain knowledge (e.g., dataset acquisition and molecular fingerprinting).

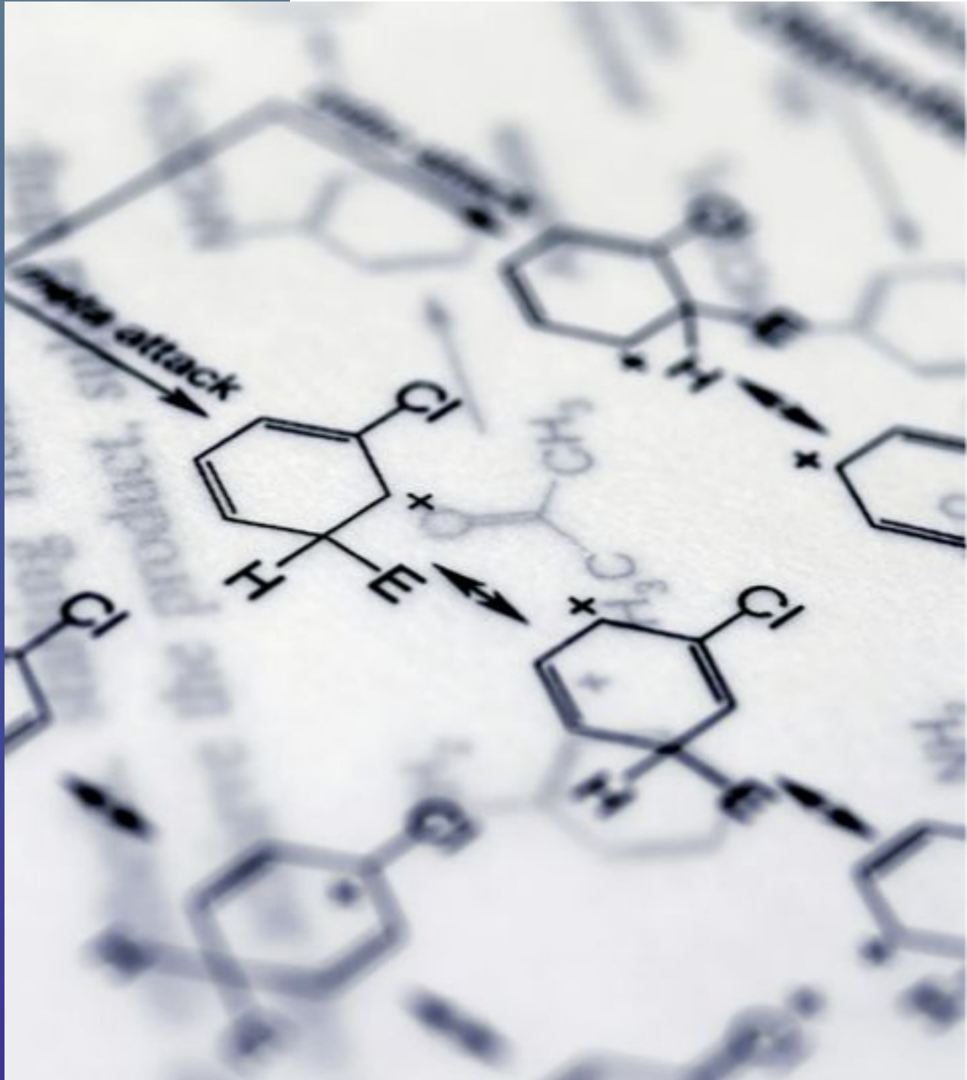
Then, the Coder debugs and implements the code and evaluates the performance. Finally, all the results are collected and the best idea is reported (e.g., random forest achieves the best performance in predicting absorption).

https://arxiv.org/pdf/2411.15692

A “secret” project I am working on …

Accelerating data understanding & efficient decision making

During translation from late preclinical to clinical

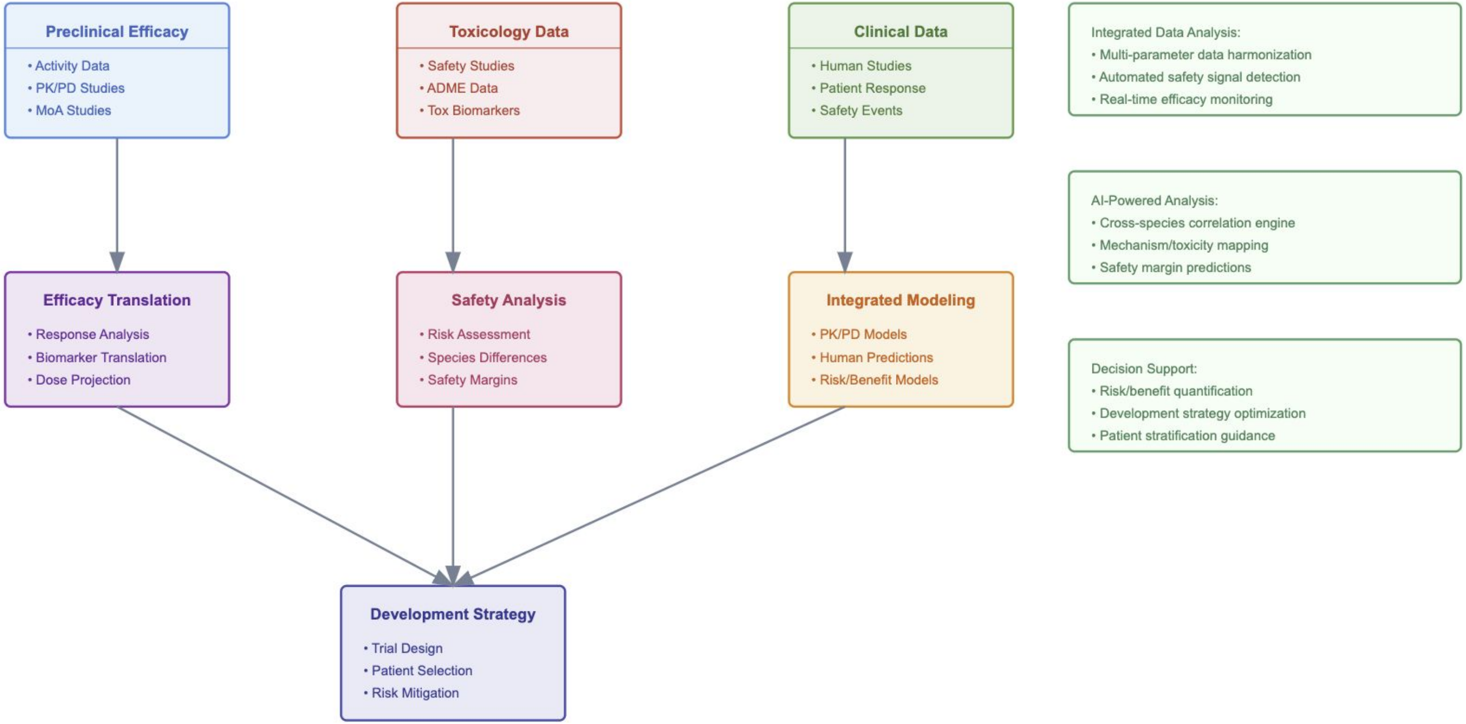


Problem / Opportunity

Scientists need a way to quickly analyze / understand findings in their data, connect these findings to existing scientific literature, and generate evidence-based hypotheses within hours rather than weeks.

If you are interested in participating as an early user, please let me know.

Translational Research Workflow: Efficacy and Safety Analysis



Preclinical Efficacy
- Activity Data
- PK/PD Studies
- MoA Studies

Toxicology Data
- Safety Studies
- ADME Data
- Tox Biomarkers

Clinical Data
- Human Studies
- Patient Response
- Safety Events

Integrated Data Analysis:
- Multi-parameter data harmonization
- Automated safety signal detection
- Real-time efficacy monitoring

AI-Powered Analysis:
- Cross-species correlation engine
- Mechanism/toxicity mapping
- Safety margin predictions

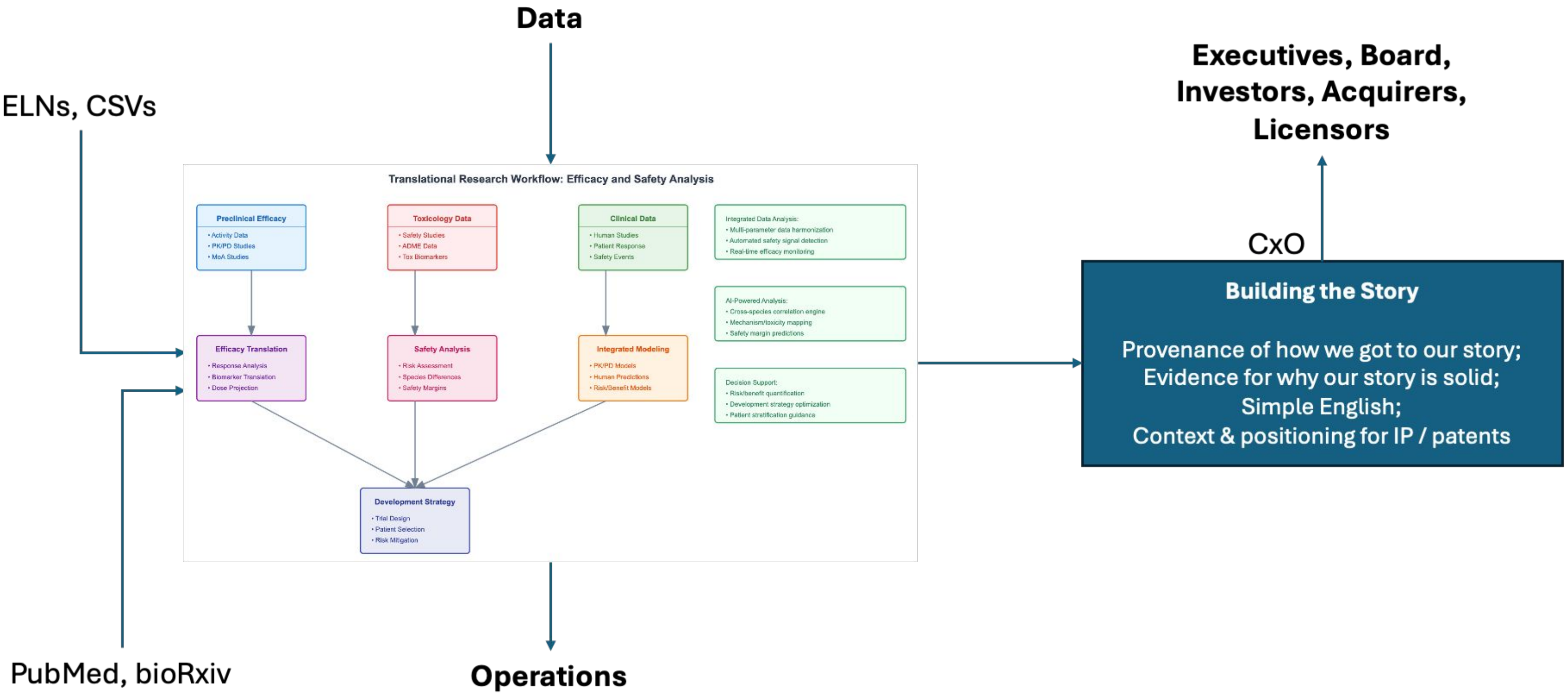
Efficacy Translation
- Response Analysis
- Biomarker Translation
- Dose Projection

Safety Analysis
- Risk Assessment
- Species Differences
- Safety Margins

Integrated Modeling
- PK/PD Models
- Human Predictions
- Risk/Benefit Models

Decision Support:
- Risk/benefit quantification
- Development strategy optimization
- Patient stratification guidance

Development Strategy
- Trial Design
- Patient Selection
- Risk Mitigation



Data

ELNs, CSVs

PubMed, bioRxiv

Translational Research Workflow: Efficacy and Safety Analysis

Preclinical Efficacy
- Activity Data
- PK/PD Studies
- MoA Studies

Toxicology Data
- Safety Studies
- ADME Data
- Tox Biomarkers

Clinical Data
- Human Studies
- Patient Response
- Safety Events

Integrated Data Analysis:
- Multi-parameter data harmonization
- Automated safety signal detection
- Real-time efficacy monitoring

Efficacy Translation
- Response Analysis
- Biomarker Translation
- Dose Projection

Safety Analysis
- Risk Assessment
- Species Differences
- Safety Margins

Integrated Modeling
- PK/PD Models
- Human Predictions
- Risk/Benefit Models

AI-Powered Analysis:
- Cross-species correlation engine
- Mechanism/toxicity mapping
- Safety margin predictions

Decision Support:
- Risk/benefit quantification
- Development strategy optimization
- Patient stratification guidance

Development Strategy
- Trial Design
- Patient Selection
- Risk Mitigation

Operations

Building the Story

Provenance of how we got to our story;
Evidence for why our story is solid;
Simple English;
Context & positioning for IP / patents

CxO

Executives, Board, Investors, Acquirers, Licensors

Setting the right expectations …

The pressure comes from improving efficiencies to deliver on these milestones as quickly as possible. For vendors and tech partners working with biotech companies, understanding these milestones is key to aligning their services and helping accelerate the drug discovery process.

Dr Wilkie, CEO Mironid Ltd

Group discussion: What are the implications of this?

Quick Activity

Using ChatGPT, to explore the ACEL paper for the effect of acarbose on lifespan.
What prompt would you use?

https://pubmed.ncbi.nlm.nih.gov/30688027/

TOPIC 3

Data Strategy, Quality, and Cleaning

Concepts:

Junk-in junk-out

Culture

FAIR

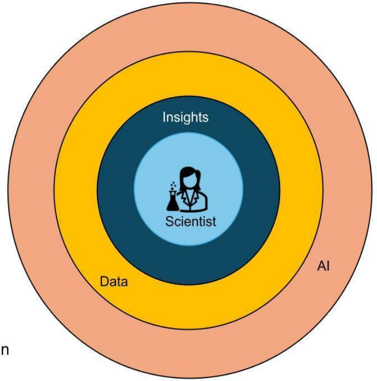
Data Strategy

Whichever approach you take, you may spend a lot of time reaching the conclusion that your data is not good enough for even simple analyses. Your first investment must be in your data strategy – understanding and developing the quality and utility of your wet lab (and/or animal) data. The following are initial recommended steps:

1. **Assess your data quality and utility and develop a data build vs buy strategy**. Some limited knowledge of AI/statistical techniques will be needed for this, as it affects the utility calculation; this knowledge can be gained through experienced advisors.

2. **Develop a both short-term and longer-term computational strategy** that leverages current data and planned new data. Be open to paying for datasets and tools, and even look at open-source software if you have the in-house software skills. Each phase of the strategy should be targeted at helping answer key scientific questions that your investors are expecting you to answer.

3. Importantly, **take some risk in going after these steps**; you are already taking (and one could argue: much more) risk in your wet lab analyses! Use this opportunity to develop relationships with data partners and tool vendors, as well as the life-sciences data science community. As a final point, it is important to remember with all the AI and data available, that the scientist must stay at the centre of the universe (Figure 1). Tools and data are only as good as the insights they generate; the creativity and invention will come from the scientists.



Enabling the scientist should be the goal of an AI-focused strategy

TIME ALLOWING

Setting the right expectations …

There is a disconnect between individual scientific achievement and collective organisational goals. High-ranking scientists are often celebrated for their groundbreaking research but are not held accountable for ensuring their work is reproducible or that it benefits the broader scientific community.

Mr Conway, CEO, 20/15 Visioneers

Group discussion: What are the implications of this?

Culture matters

1. **Institutionalising data stewardship training**: From secondary education through to graduate school, there is a pressing need to incorporate data stewardship into the curriculum. Scientists should be trained not just in their specific disciplines, such as chemistry or biology, but also in the principles of managing data as a valuable and reusable asset. This training would help inculcate a culture of meticulous data handling from the very beginning of a scientist's career.

2. **Implementing rigorous accountability mechanisms**: Organisations need to develop and enforce protocols that hold all scientists accountable for the reproducibility and quality of their work. This could involve regular audits of research data and methodologies, ensuring that senior scientists adhere to the same standards as their junior colleagues.

3. **Promoting a humble, collaborative culture**: Drawing from examples like Sweden, where humility and collaboration are ingrained cultural values, organisations should strive to promote a work environment where self-promotion is secondary to the quality and impact of the research. Encouraging scientists to work collaboratively and share credit can lead to more reliable and comprehensive research outcomes.

4. **Leadership by example**: Cultural change must be driven from the top down. Leaders in research organisations should model the behaviour they expect from their teams, demonstrating a commitment to data quality, collaboration, and continuous learning. When leadership prioritises these values, it sets a precedent that permeates the entire organisation.

5. **Rewarding reproducibility and data quality**: Shifting the focus of recognition and rewards from individual achievements to reproducibility and data quality can help change the underlying cultural dynamics. By celebrating efforts that contribute to robust and reusable data, organisations can align incentives with long-term research success.

Impact of poor data quality

1. **Irrelevant data for a given purpose -** leading to the generation of machine learning models that predict other endpoints than those needed for decision making in practice (hence in the worst case even being misleading as a result).

2. **Incomplete data can cause ML models to miss patterns or relationships**. For example, if crucial bioactivity data for certain compounds is missing, the model may not fully account for the structure-activity relationships, leading to inaccurate predictions.

3. **Inconsistencies, such as variations in how data is recorded** (eg, different units of measurement or naming conventions), can confuse models and lead to erroneous predictions. For instance, if the same compound is labelled differently in different datasets (due to a different tautomer, salt form, there are many reasons why this can happen), the model might treat it as different entities, skewing the results.

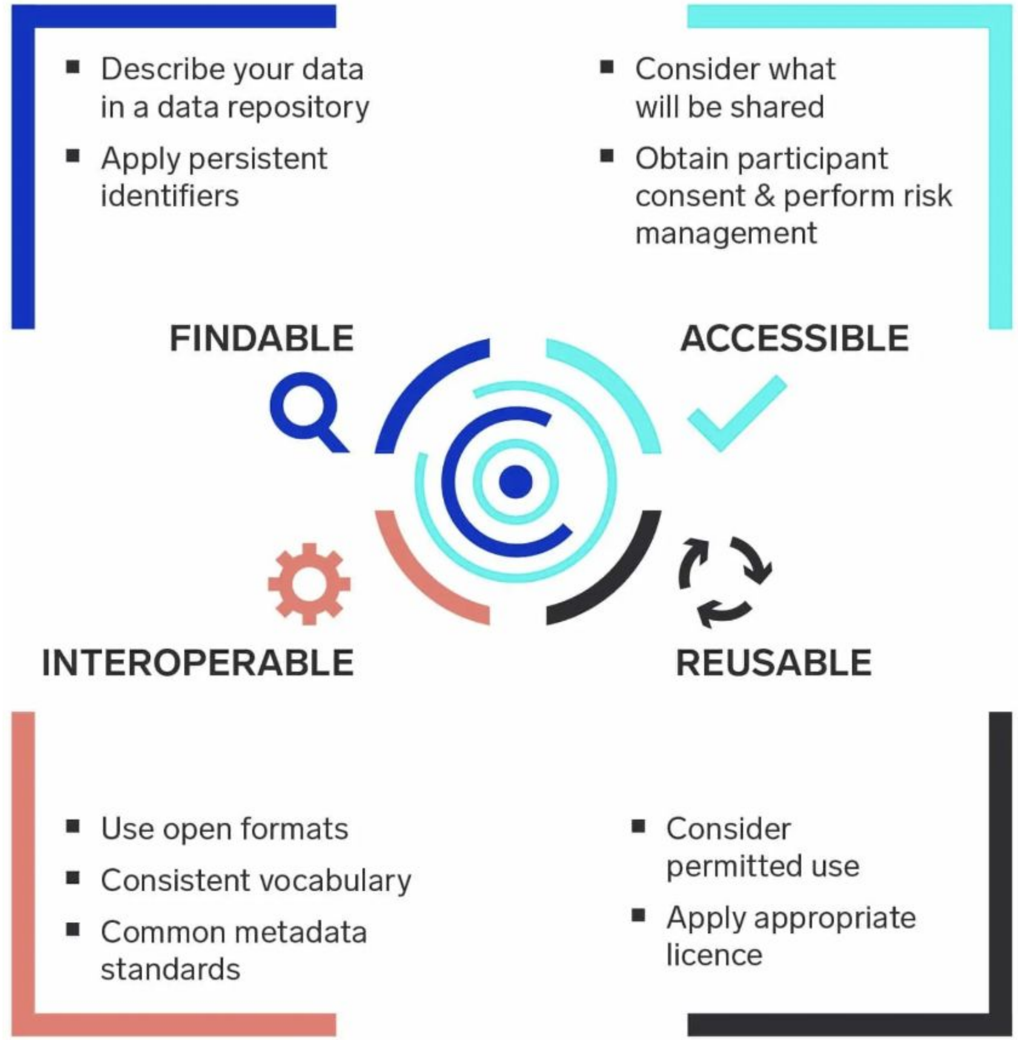
4. **Bias in data can lead to models that are not generalisable** or that perform poorly on certain subsets of data. For instance, if your training data is biased toward a particular chemical scaffold or a specific set of biological targets, the model will be less effective in predicting the activity of compounds outside these categories.

5. **Noise in data**, which may arise from experimental errors, variability in biological assays, or inconsistent conditions, **can obscure true signals** and reduce the model's ability to learn relevant patterns. This can result in a higher rate of false positives or negatives.

6. **Duplicate records and redundant features** or data points **can inflate the dimensionality of the data** without adding new information, leading to overfitting and reduced model performance.

7. **Scientific reproducibility is compromised when data quality is poor**, as other researchers or systems might not replicate the findings.

FAIR data



FINDABLE

- Describe your data in a data repository
- Apply persistent identifiers

ACCESSIBLE

- Consider what will be shared
- Obtain participant consent & perform risk management

INTEROPERABLE

- Use open formats
- Consistent vocabulary
- Common metadata standards

REUSABLE

- Consider permitted use
- Apply appropriate licence

https://www.idbs.com/the-fair-principles-a-quick-introduction/

FAIR data benefits

Enhancing Research Reproducibility and Validation

- Your experimental results and computational models become independently verifiable by other researchers, which is crucial for drug development where small variations can have significant impacts. When data is properly documented and accessible, other teams can validate findings and build upon your work with confidence.

Accelerating Discovery Timelines

- Well-organized, findable data prevents duplicate experiments and allows researchers to quickly build on previous findings. When data from different stages of discovery (target identification, screening, lead optimization, etc.) is interoperable, you can more efficiently move compounds through your pipeline.

Improving Collaboration Opportunities

- Standardized data formats and clear metadata make it easier to share information between different teams, departments, and external partners. This is especially valuable in drug discovery where you often need to combine expertise across chemistry, biology, and computational disciplines.

FAIR data benefits (con't)

Protecting Intellectual Property

- Having well-documented, traceable data helps establish clear proof of discovery timelines and supports patent applications. FAIR principles ensure that your valuable research data remains accessible and understandable even as team members change over time.

Maximizing Data Value Over Time

- Drug discovery data often has value beyond its initial use case. When data is reusable and well-documented, you can more easily apply it to new therapeutic targets or combine it with new datasets for machine learning applications. This extends the return on your research investment.

Meeting Regulatory Requirements

- FAIR principles align with regulatory expectations for data integrity and traceability. Having standardized, accessible data makes it easier to compile regulatory submissions and respond to regulatory queries throughout the drug development process.

TIME ALLOWING

Discussion

How can you adjust for outliers and biases in scientific data that you regularly deal with?

LUNCH

TOPIC 4

Reading Material 4, Downloadable papers & data

No-code Method; Lab

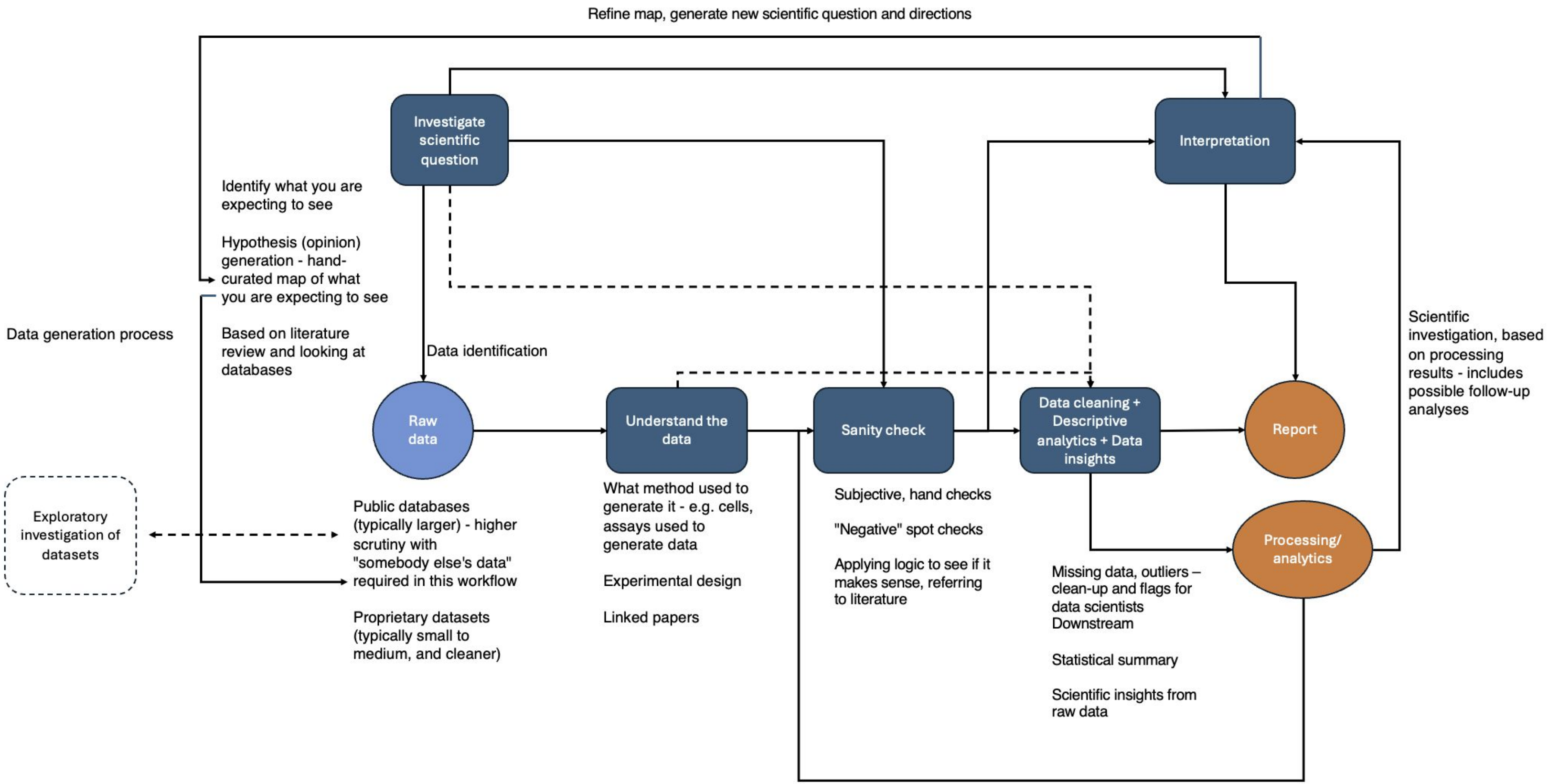
Acknowledgement: Nina Truter

Nina Truter is a translational scientist with a deep focus on understanding mechanisms of action in drug development and leveraging disparate datasets in biotech. Based in South Africa, she has worked extensively with international biotech companies, specialising in therapeutic development for aging-related diseases and complex conditions such as glioblastoma and Autosomal Dominant Polycystic Kidney Disease (ADPKD).

Her recent work includes consulting for UK-based biotech firms and leading initiatives in HitchhikersAI.org to advance the translation of AI and data science into practical biotech solutions such as identifying combination therapy opportunities and enhancing patient selection. In her work, she uses a systems approach to integrate insights from diverse datasets across *in vitro, in vivo*, and human models—to answer critical scientific questions, and translates biological mechanisms into models that are used by advanced analytical methods such as Pearlian causal inference.

For more: https://njtruter.wixsite.com/ninatruter

Scientific Workflow



Refine map, generate new scientific question and directions
Investigate scientific question
Interpretation
Identify what you are expecting to see
Hypothesis (opinion) generation - hand-curated map of what you are expecting to see
Based on literature review and looking at databases
Data generation process
Data identification
Scientific investigation, based on processing results - includes possible follow-up analyses
Raw data
Understand the data
Sanity check
Data cleaning + Descriptive analytics + Data insights
Report
Exploratory investigation of datasets
Public databases (typically larger) - higher scrutiny with "somebody else's data" required in this workflow
Proprietary datasets (typically small to medium, and cleaner)
What method used to generate it - e.g. cells, assays used to generate data
Experimental design
Linked papers
Subjective, hand checks
"Negative" spot checks
Applying logic to see if it makes sense, referring to literature
Processing/ analytics
Missing data, outliers – clean-up and flags for data scientists Downstream
Statistical summary
Scientific insights from raw data

Important considerations

Defining the research question
A well-defined research question is the cornerstone of an effective scientific workflow in drug discovery. The more specific your question, the easier it becomes to identify relevant data and design subsequent steps in your workflow. This initial phase often involves an iterative process: refining your question, conducting a literature review, and assessing available data to ensure the right level of specificity and relevance of your research question. AI tools like ChatGPT can help refine your question and provide an overview of the research landscape before you dive into a full literature review.

Hypothesis generation
The hypothesis generation process is equally important. Before diving into data analysis, a hypothesis must be developed based on literature reviews and public datasets. The scientific question guides the entire investigation, and without a clear hypothesis, the research could become unfocused and exploratory. Having a well-defined hypothesis allows researchers to assess datasets critically and ensures that their analysis remains grounded in the biological context. Creating a rough map containing the relevant variables that influence the outcome of the scientific question based on literature review and logic can help structure the hypothesis. This map can be used as a “checklist” when assessing whether a dataset contains the necessary variables to answer the research question.

Data identification
When searching for public data, tools like Perplexity.ai can aid in the process of identifying relevant databases by for example, asking “Which database should I use to search for data on the effects of longevity drugs in rodents?”. While ChatGPT and Claude.ai are useful for general information to questions, Perplexity.ai tends to provide more accurate, “fact based” answers. Google Dataset Search or PubMed’s “Associated Data” feature can uncover datasets linked to publications. After identifying a potentially useful dataset, Claude.ai can summarize experimental methods to determine if the dataset is the right fit for your research question. Creating a descriptive spreadsheet to catalog potential datasets, along with a broad description of their contents, helps streamline the selection process. In some cases, combining multiple datasets may be necessary to comprehensively address your research question.

Important considerations

Understanding Data

Before diving into analysis, ample time should be spent reviewing the raw data. Browsing through datasets, often in Excel format, can clarify how the data were generated, helping you choose appropriate analytic methods and establish sanity checks. For data types that are less familiar, ChatGPT can be helpful in explaining the experimental method and for establishing potential validation steps. Alternatively, search for review papers or papers using a similar method and understand how it was applied in that context.

Visualization is another powerful tool for data understanding—experimenting with different methods can provide varied perspectives. ChatGPT can also aid in deciding which visualisation options are available and what information each will provide, based on the data and your research question. Additionally, running analyses on both the raw/”uncleaned” and “cleaned” versions of the dataset helps assess the impact of outliers and can guide decisions on whether to include or exclude them.

Analyzing and Interpreting Results

When it comes to data analysis, Claude.ai has analytics tools that offer specific methods which can improve the data analytic process. Although ChatGPT is helpful as an initial step to understand results, it should be used as a tool for creating literature review ideas and hypothesis generation, not as a fact-based system. The scientific question should stay the anchor of the interpretive process, together with your understanding of the raw data and output from analytics. Here, it is helpful to toggle between two mindsets - one of a creative scientist, which is useful for creating avenues of exploration and one of a critic when assessing the merit of these avenues.

Exploratory investigations and missed opportunities

Often, datasets are generated for a specific research question, but they may contain additional information that could be useful for answering new or unrelated questions. This is particularly true for large public datasets, where the breadth of data available can sometimes be overwhelming. Researchers may miss opportunities to generate new insights simply because they are focused on their initial question and do not have the resources to explore other possibilities.

Additionally, exploratory analyses can be valuable for identifying new biological markers or hypotheses. For instance, a dataset generated to study protein expression in one context might also reveal valuable information about other biological pathways or processes. However, exploratory investigations can be resource-intensive, both in terms of time and computational power. Researchers need to balance their focused analysis with the potential for broader discoveries.

Tools used

1. KEGG Pathway Database - The KEGG (Kyoto Encyclopedia of Genes and Genomes) Pathway database provides information on molecular interaction and reaction networks for various biological pathways. https://www.kegg.jp/kegg/pathway.html

2. STRING Database - STRING is a database of known and predicted protein-protein interactions, integrating both physical and functional associations, https://string-db.org

3. Reactome - Reactome is an open-source, curated pathway database that provides insights into biological processes and molecular interactions, https://reactome.org

4. GeneCards - GeneCards is a comprehensive database that provides detailed information on all known and predicted human genes, including functions, pathways, and related diseases, https://www.genecards.org

5. Cytoscape - Cytoscape is a software platform for visualising molecular interaction networks and integrating these networks with gene expression profiles and other data, https://cytoscape.org

6. Mendeley - Mendeley is a reference manager and academic social network that helps researchers organise research papers, collaborate online, and discover the latest scientific research, https://www.mendeley.com

7. Zotero - Zotero is a free, easy-to-use tool to help researchers collect, organize, cite, and share research, https://www.zotero.org

8. TensorFlow - TensorFlow is an open-source platform for machine learning, commonly used for deep learning applications and large dataset analysis, https://www.tensorflow.org

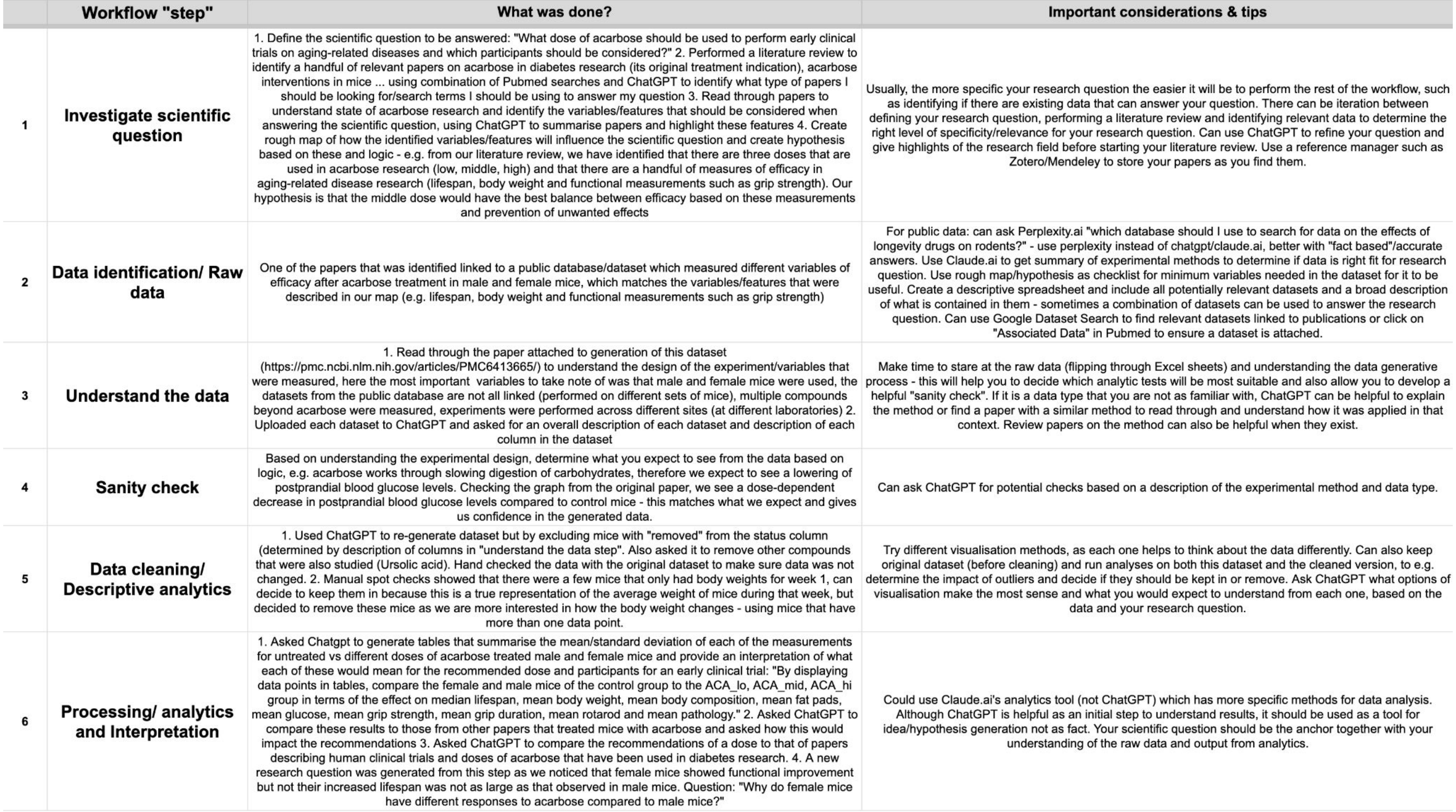
9. PyTorch - PyTorch is an open-source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, https://pytorch.org

Setting the right expectations …

You've got this mass amount of data, and how do you move it around? How do you even access it easily? And then, what do you do with it? This challenge is compounded by the increasing complexity of biological data, particularly in imaging, which often contains hidden insights that are difficult to extract without advanced computational tools.

Dr Thomas, CEO, Five Alarm Biosciences

Group discussion: What are the implications of this?



	Workflow "step"	What was done?	Important considerations & tips
1	**Investigate scientific question**	1. Define the scientific question to be answered: "What dose of acarbose should be used to perform early clinical trials on aging-related diseases and which participants should be considered?" 2. Performed a literature review to identify a handful of relevant papers on acarbose in diabetes research (its original treatment indication), acarbose interventions in mice ... using combination of Pubmed searches and ChatGPT to identify what type of papers I should be looking for/search terms I should be using to answer my question 3. Read through papers to understand state of acarbose research and identify the variables/features that should be considered when answering the scientific question, using ChatGPT to summarise papers and highlight these features 4. Create rough map of how the identified variables/features will influence the scientific question and create hypothesis based on these and logic - e.g. from our literature review, we have identified that there are three doses that are used in acarbose research (low, middle, high) and that there are a handful of measures of efficacy in aging-related disease research (lifespan, body weight and functional measurements such as grip strength). Our hypothesis is that the middle dose would have the best balance between efficacy based on these measurements and prevention of unwanted effects	Usually, the more specific your research question the easier it will be to perform the rest of the workflow, such as identifying if there are existing data that can answer your question. There can be iteration between defining your research question, performing a literature review and identifying relevant data to determine the right level of specificity/relevance for your research question. Can use ChatGPT to refine your question and give highlights of the research field before starting your literature review. Use a reference manager such as Zotero/Mendeley to store your papers as you find them.
2	**Data identification/ Raw data**	One of the papers that was identified linked to a public database/dataset which measured different variables of efficacy after acarbose treatment in male and female mice, which matches the variables/features that were described in our map (e.g. lifespan, body weight and functional measurements such as grip strength)	For public data: can ask Perplexity.ai "which database should I use to search for data on the effects of longevity drugs on rodents?" - use perplexity instead of chatgpt/claude.ai, better with "fact based"/accurate answers. Use Claude.ai to get summary of experimental methods to determine if data is right fit for research question. Use rough map/hypothesis as checklist for minimum variables needed in the dataset for it to be useful. Create a descriptive spreadsheet and include all potentially relevant datasets and a broad description of what is contained in them - sometimes a combination of datasets can be used to answer the research question. Can use Google Dataset Search to find relevant datasets linked to publications or click on "Associated Data" in Pubmed to ensure a dataset is attached.
3	**Understand the data**	1. Read through the paper attached to generation of this dataset (https://pmc.ncbi.nlm.nih.gov/articles/PMC6413665/) to understand the design of the experiment/variables that were measured, here the most important variables to take note of was that male and female mice were used, the datasets from the public database are not all linked (performed on different sets of mice), multiple compounds beyond acarbose were measured, experiments were performed across different sites (at different laboratories) 2. Uploaded each dataset to ChatGPT and asked for an overall description of each dataset and description of each column in the dataset	Make time to stare at the raw data (flipping through Excel sheets) and understanding the data generative process - this will help you to decide which analytic tests will be most suitable and also allow you to develop a helpful "sanity check". If it is a data type that you are not as familiar with, ChatGPT can be helpful to explain the method or find a paper with a similar method to read through and understand how it was applied in that context. Review papers on the method can also be helpful when they exist.
4	**Sanity check**	Based on understanding the experimental design, determine what you expect to see from the data based on logic, e.g. acarbose works through slowing digestion of carbohydrates, therefore we expect to see a lowering of postprandial blood glucose levels. Checking the graph from the original paper, we see a dose-dependent decrease in postprandial blood glucose levels compared to control mice - this matches what we expect and gives us confidence in the generated data.	Can ask ChatGPT for potential checks based on a description of the experimental method and data type.
5	**Data cleaning/ Descriptive analytics**	1. Used ChatGPT to re-generate dataset but by excluding mice with "removed" from the status column (determined by description of columns in "understand the data step". Also asked it to remove other compounds that were also studied (Ursolic acid). Hand checked the data with the original dataset to make sure data was not changed. 2. Manual spot checks showed that there were a few mice that only had body weights for week 1, can decide to keep them in because this is a true representation of the average weight of mice during that week, but decided to remove these mice as we are more interested in how the body weight changes - using mice that have more than one data point.	Try different visualisation methods, as each one helps to think about the data differently. Can also keep original dataset (before cleaning) and run analyses on both this dataset and the cleaned version, to e.g. determine the impact of outliers and decide if they should be kept in or remove. Ask ChatGPT what options of visualisation make the most sense and what you would expect to understand from each one, based on the data and your research question.
6	**Processing/ analytics and Interpretation**	1. Asked Chatgpt to generate tables that summarise the mean/standard deviation of each of the measurements for untreated vs different doses of acarbose treated male and female mice and provide an interpretation of what each of these would mean for the recommended dose and participants for an early clinical trial: "By displaying data points in tables, compare the female and male mice of the control group to the ACA_lo, ACA_mid, ACA_hi group in terms of the effect on median lifespan, mean body weight, mean body composition, mean fat pads, mean glucose, mean grip strength, mean grip duration, mean rotarod and mean pathology." 2. Asked ChatGPT to compare these results to those from other papers that treated mice with acarbose and asked how this would impact the recommendations 3. Asked ChatGPT to compare the recommendations of a dose to that of papers describing human clinical trials and doses of acarbose that have been used in diabetes research. 4. A new research question was generated from this step as we noticed that female mice showed functional improvement but not their increased lifespan was not as large as that observed in male mice. Question: "Why do female mice have different responses to acarbose compared to male mice?"	Could use Claude.ai's analytics tool (not ChatGPT) which has more specific methods for data analysis. Although ChatGPT is helpful as an initial step to understand results, it should be used as a tool for idea/hypothesis generation not as fact. Your scientific question should be the anchor together with your understanding of the raw data and output from analytics.

For the example data:
https://www.hitchhikersai.org/reading

https://docs.google.com/spreadsheets/d/1nyYnrSrMiA0SFzIzZ2VQ0iz7DtYqk59GETeT05cyLeU/edit?pli=1&gid=0#gid=0

Hands-on Lab

“Investigating a Scientific Question”

The goal here is to learn how to use the tools to implement the Scientific Workflow, starting with an example paper.

Struggle a little … there is no correct answer, this is a creative exercise.

The experience will give you confidence with these tools in your own work environment.

Challenge - Investigating a Scientific Question

Scientific que: What dose of acarbose should be used to perform early human clinical trials on aging-related diseases and which participants should be considered?

Perform a literature review to identify a handful of relevant papers on acarbose in diabetes research (its original treatment indication), acarbose interventions in mice … using combination of **Pubmed** searches and **ChatGPT** prompts to identify what type of papers I should be looking for/search terms I should be using to answer my question.

Scrape the papers to understand state of acarbose research and identify the variables/features that should be considered when answering the scientific question, using ChatGPT and/or Claude (**hint**: Claude can do descriptive analytics) to summarise papers and highlight these features. How do the identified variables/features influence the scientific question and create hypotheses based on these and logic? See if you can show some relevant visuals.

Investigate scientific question
Data identification/ Raw data
Understand the data
Sanity check
Data cleaning/ Descriptive analytics
Processing/ analytics and Interpretation

Starting Paper: ACEL Paper Abstract*

To follow-up on our previous report that acarbose (ACA), a drug that blocks postprandial glucose spikes, increases mouse lifespan, **we studied ACA at three doses: 400, 1,000 (the original dose), and 2,500 ppm, using genetically heterogeneous mice** at three sites.

Each dose led to a significant change (by log-rank test) in both sexes, with larger effects in males, consistent with the original report. There were no significant differences among the three doses.

The two higher doses produced 16% or 17% increases in median longevity of males, but only 4% or 5% increases in females. Age at the 90th percentile was increased significantly (8%–11%) in males at each dose, but was significantly increased (3%) in females only at 1,000 ppm. The sex effect on longevity is not explained simply by weight or fat mass, which were reduced by ACA more in females than in males. ACA at 1,000 ppm reduced lung tumors in males, diminished liver degeneration in both sexes and glomerulosclerosis in females, reduced blood glucose responses to refeeding in males, and improved rotarod performance in aging females, but not males.

Three other interventions were also tested: ursolic acid, 2-(2-hydroxyphenyl) benzothiazole (HBX), and INT-767; none of these affected lifespan at the doses tested. The acarbose results confirm and extend our original report, prompt further attention to the effects of transient periods of high blood glucose on aging and the diseases of aging, including cancer, and should motivate studies of acarbose and other glucose-control drugs in humans.

https://pmc.ncbi.nlm.nih.gov/articles/PMC6413665/

*same paper we looked at in the morning

Lab Readouts

End of day, so enjoy the discussion.

We are done. 🎉
Questions?

Please remember to check out the reading material in the eBooklet.

An please attend my NexusXP panel session, my Podium talk, and my DDW session talk at SLAS2025. I have details, if you can’t find them in the program.

raminderpal@20visioneers15.com
raminderpal@hitchhikersai.org

THANK YOU!

Guideline	Implementation example 1	Implementation example 2
1. Data Collection and Entry	Standardisation: Implement standardised operating procedures (SOPs) for data collection and entry, including consistent use of units, naming conventions, and data formats.	Training: Train the team involved in data collection on the importance of data quality and the specific protocols to follow to minimise errors.
2. Data Validation	Automated Checks: Use automated validation scripts to check for common issues such as missing values, duplicates, outliers, and inconsistencies in units or formats.	Manual Review: Periodically perform manual reviews of a subset of the data to identify any issues that automated checks might miss.
3. Data Cleaning	Missing Data Handling: Develop a strategy for handling missing data, such as deciding when to use imputation, exclude data points, or flag datasets for further investigation.	Outlier Detection: Implement methods to identify and investigate outliers, determining whether they represent true variability or errors.
4. Data Integration	Harmonisation: Ensure that data from different sources or experiments are harmonised before integration. This includes reconciling different naming conventions, units, and formats.	Cross-Validation: Use cross-referencing methods to validate integrated datasets, checking for consistency and correctness.
5. Data Documentation	Metadata: Maintain detailed metadata for each dataset, including information about the origin, collection method, and any preprocessing steps. This helps in tracking data provenance and understanding the context.	Version Control: Use version control systems for datasets to track changes and ensure that any modifications are well-documented and reversible.
6. Data Monitoring	Continuous Monitoring: Implement ongoing monitoring of data quality metrics, such as completeness, accuracy, and consistency, throughout the data lifecycle.	Alerts: Set up automated alerts to notify relevant personnel if data quality metrics fall below predefined thresholds.
7. Data Auditing	Regular Audits: Conduct regular data audits to assess the overall quality of your datasets. This involves checking for adherence to data quality standards and identifying any systemic issues.	Audit Trails: Maintain audit trails that log all data processing steps, transformations, and any changes made to the data. This ensures traceability and accountability.

Guideline	Implementation example 1	Implementation example 2
8. Bias and Variability Checks	Bias Analysis: Regularly assess your datasets for potential biases, such as over-representation of certain chemical scaffolds or biological targets. Use statistical techniques to quantify bias and take corrective actions. This involves a change in mindset in particular, from project-based data generation, to process-based data generation (where said process involves the use of AI models by default)	Variance Analysis: Analyse the variability in your data, especially in biological assays, to understand the level of noise and its impact on model performance.
9. Data Redundancy and Duplication Checks	Duplicate Detection: Implement robust mechanisms to detect and remove duplicate records to prevent skewing of the data.	Feature Redundancy Check: Use techniques like correlation analysis to identify and eliminate redundant features that do not contribute new information.
10. Data Imbalance Handling	Balance Check: Continuously monitor the balance of different classes in your data (e.g., active vs. inactive compounds). Address imbalances through methods like oversampling, undersampling, or synthetic data generation.	Model Adaptation: If data imbalance is unavoidable, consider using model algorithms that are better suited to handle imbalanced data.
11. Data Security and Access Control	Access Control: Restrict access to data based on roles and responsibilities to prevent unauthorised modifications or data entry errors.	Data Security: Implement security measures to protect data from corruption, loss, or unauthorised access, ensuring that data integrity is maintained.
12. Communication and Collaboration	Interdisciplinary Collaboration: Foster collaboration between data scientists, domain experts, and IT professionals to ensure that data quality requirements are clearly understood and addressed.	Feedback Loop: Establish a feedback loop where issues identified by data scientists or model results are communicated back to the experimental team to refine data collection processes.

Guideline	Implementation example 1	Implementation example 2
13. Use of Quality Control Samples	Control Samples: Include quality control samples (e.g., known standards or replicates) in experimental runs to monitor and ensure consistency in assay performance.	QC Analysis: Regularly analyse the results of quality control samples to identify any drifts or deviations in experimental conditions that could impact data quality.
14. Data Quality Metrics and Reporting	Define Metrics: Define specific data quality metrics such as predictivity for a downstream endpoint, accuracy, completeness, consistency, timeliness, and uniqueness. Use these metrics to evaluate and report on the quality of your data regularly.	Reporting: Regularly report on data quality metrics to stakeholders, ensuring transparency and facilitating continuous improvement.
15. Continuous Improvement	Root Cause Analysis: When data quality issues are identified, perform root cause analysis to understand the underlying reasons and implement corrective actions.	Iterative Process: Treat data quality improvement as an iterative process, continuously refining and enhancing your strategies as new challenges and technologies emerge.