lucidata

White Paper

ChatGPT in Drug Discovery: Rise of Large Language Models

Contents

Abstract

While large language models have been around for a while, Open AI's ChatGPT has sparked unprecedented interest in their application in diverse domains, including problems in drug discovery.

In these early days, ChatGPT and similar models are showing promising signs of applications in information extraction, authoring scientific text, prediction of hypothesis and predicting chemical entities.

In this whitepaper, we discuss the potential and possibilities the model can bring to make drug research a less laborious process. We demystify it's underlying technology, hypothesize on it's applications in various steps of discovery and discuss limitations.

Authors - Anjali Gupta, *Ph.D., Scientific Writer*; Gaurang Mahajan, *Ph.D., ML-Researcher*; Himanshi Allahabadi, *M.S., Data Scientist*; Mukund Chaudhry, *B.Sc., Lead Engineer (Bioinformatics)*; Vishal Samal, *B.Tech., Data Scientist*

Editors - Aqsa Aleem*, *M.A., Associate Marketing Manager*; Shruti Malavade, *MBA, Product Marketing Manager*; Swetabh Pathak, *M. Tech., CTO and Co-Founder*

Design - Sohom Mandal, *B.Des., Senior Visual and Information Designer*

All contributors are affiliated with **Elucidata Corporation**

*Corresponding author email address: **aqsa.aleem@elucidata.io**

ChatGPT has captured the public imagination in a way that few things have. **Large language models** (LLMs), the class of Machine Learning models that ChatGPT belongs to, are maturing at an astonishing pace, showcasing remarkable versatility across diverse domains. ML community has been taken by surprise by the quantum jump from the previous benchmark - **BERT** models released in 2018.

The Life Sciences community is curious about the impact of ChatGPT on their work. Can ChatGPT write papers/reports? Can LLMs generate hypothesis? More generally, can these models help speed-up Drug R&D?

These are early days for ChatGPT and other LLMs. (We will use ChatGPT and LLMs interchangeably in this whitepaper. Strictly speaking that's not correct) We are in uncharted territories. The best we can do is take educated guesses and look up to early adopters.

Early adopters of ChatGPT(or LLMs) are trying a number of applications in drug R&D:

- finding similar compounds to those that researchers are studying;
- extracting information from unstructured text (the use case we will show in depth below),
- proposing re-engineered compounds and identifying mutations that alter pathogenicity; and
- determining whether novel compounds are patented.

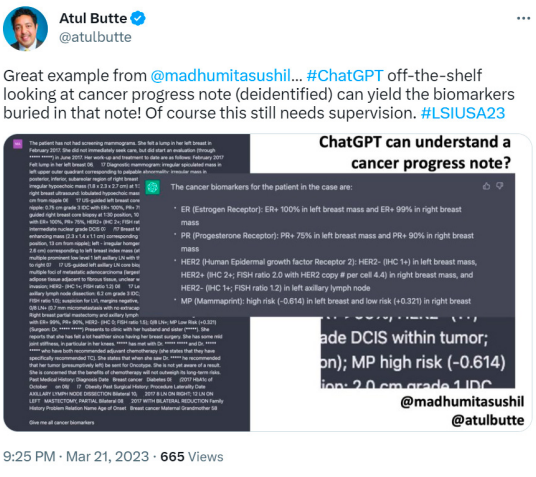
In this white-paper, we present a simple explanation of what GPT (Generative Pre-trained Transformer) models are, track the evolution from pre-GPT models to the powerful ChatGPT, highlight novel applications in drug discovery, share thoughts from leaders in the field and discuss limitations. We also provide an exclusive peek at a proof of concept study on biocuration, conducted by Elucidata's ML scientists.

Before we delve in into the details, here's a sneak peek of what generalists and experts have to say.

Word Around the Industry

The world is abuzz with excitement and speculation about the potential revolution ChatGPT could bring. Opinions within the biopharma industry itself are varied. Some hail it as a revolutionary technology that has the potential to transform drug discovery and development, while others express concerns about its limitations and ethical implications. Since the release of ChatGPT and other LLMs, there has been a shift in imagination and perception of what the future holds.

Twitter is abuzz with chatter.



We also reached out to some thought leaders to see what they thought.

"ChatGPT and other forms of Generative AI are a groundbreaking technology that will shape up the world in unprecedented ways.

This technology can be leveraged to accelerate the drug discovery process. There are examples of generating novel chemical entities using such models already. Closed loop systems that produce small molecules, testing them, and then deciding the next experiments without human involvement might be transformative.

The technology is definitely promising. There is evidence that such models can be creative. Coming up with ideas that human researchers overlooked. But it surely has it limitations- the lack of understanding of how the model arrived at the prediction is one concern. The growing gap in access to certain medications and knowledge is another concern. But let's wait, and we'll see."

-**Avi Ma'ayan**
Professor, Department of Pharmacological Sciences; Director, Mount Sinai Center for Bioinformatics

"I believe the release of GPT and other LLM models has changed my imagination of what is possible in the life sciences industry over the next 10 years. We are living through an inflection point in history, and I hope the life sciences industry does not lag behind in adopting this technology. This is just the beginning, and the impact of AI in life sciences will be profound, changing every stage of drug discovery and patient care.

While there will still be room for human expertise, companies will need to adapt to meet the expectations in a post-GPT world. Limitations of the technology can be mitigated with proper processes and human intervention will still be required, but the extent of it will be worked out as we go. At least for the next few months, using GPT or LLMs in a context-specific way will create value for businesses, and building a team will need to evolve to include experts in LLMs and domain knowledge for effective utilization of the technology."

-**Abhishek Jha**
CEO & Co-Founder, Elucidata Corp.

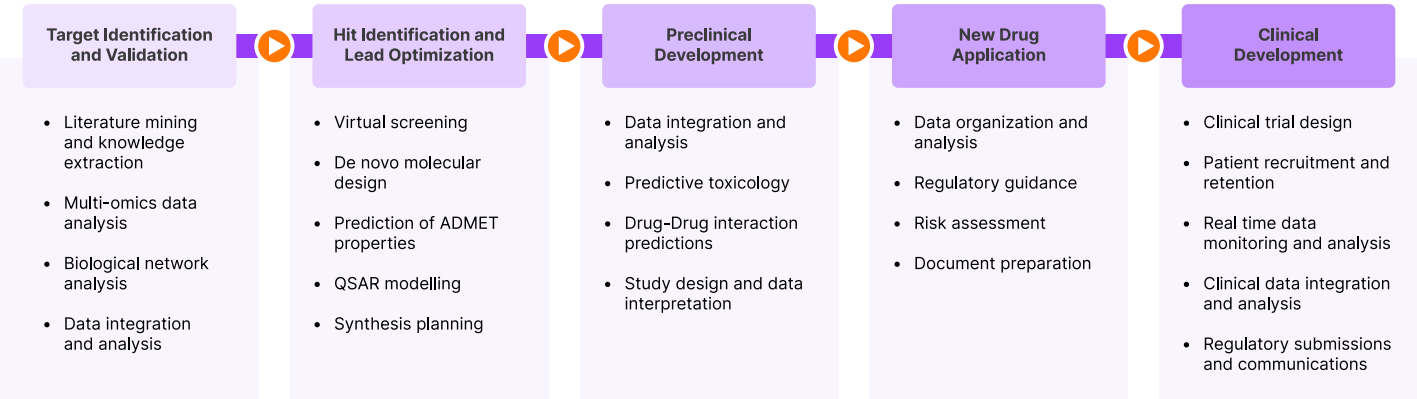
How Can ChatGPT and LLMs Revolutionize Drug Discovery? Taking It Step by Step

TL;DR: In this section, we look at some potential applications of ChatGPT/LLMs in the various steps of drug discovery. As seen from the examples below, this technology can make it easier to extract information from large amounts of text, augment data when needed, author reports and propose hypothesis.

LLMs use statistical models to analyze vast amounts of data, learning the patterns and connections between words and phrases. They mimic human intelligence for languages. LLMs are trained to distill the semantic and syntactic "structure" in large corpora of data. Once trained, they can be prompted to leverage that 'learning' to create novel entities/sequences which follow implicit rules of language. Unlike older models, LLMs don't need these rules to specifically be programmed into the models.

LLMs are built to work with human language. Hence making them appropriate out of the box to do tasks like answer queries or author text. Just as LLMs are built to work with native language (plain English or computer code), we can imagine those that are trained with chemical or biological language (SMILES strings, protein and DNA sequences). If ChatGPT can write code and debug programs, similar models can potentially generate sequences of interest. Indeed they have.

Take another example to consider the possibilities. The space of possible drugs consistent with the rules of chemistry and therapeutic guidelines is very vast (10^30 - 10^60). Only a tiny fraction of this chemical space is sampled by compound libraries used for drug screening (in the range of millions). LLM like models (sometimes called generative models) can help extrapolate into the under-explored regions of this space and come up with novel plausible candidates.

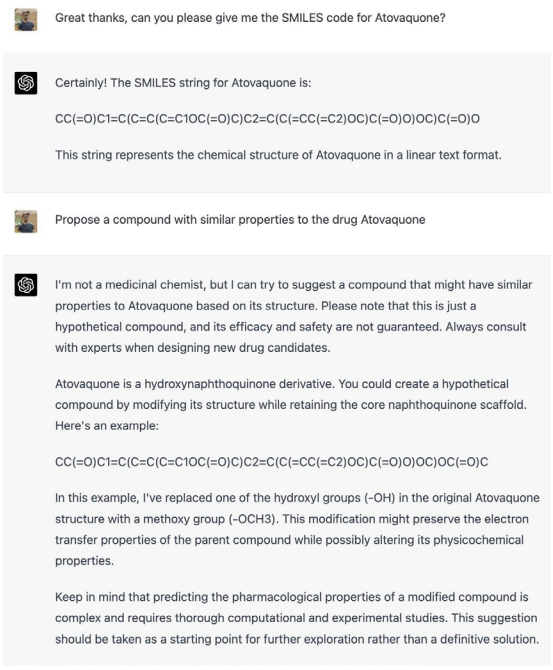


ChatGPT and models like it can potentially impact multiple stages of the drug discovery process. Here we'll look at the potential impact at each stage.

1. Target Identification and Validation

Target ID and validation requires scientists to be on top of vast amounts of foundational scientific literature. Researchers spend months gathering information from published and in-house sources. ChatGPT can significantly accelerate this process by consuming textual data from scientific literature, public databases, patents, clinical trials and more.

For example, to find a novel target for a specific condition, GPT can mine information from these sources and answer specific queries. The big difference from earlier models, like BERT referred earlier, is that it does not have to be trained for every new kind of information. Earlier, ML scientists would train one model to extract drug information. Then another to extract patient IDs and so on. This was a linear process where each step was (almost) independent of the other. These new age LLMs can learn about new entities without needing more training - they just need good prompts.



LLMs can significantly reduce the amount of work required to assemble knowledge graphs like entities. Like Knowledge Graphs have been used to extract hitherto unknown relationships, LLMs can extract relationships from millions of documents. The exciting thing about ChatGPT is that it could possibly predict new causative hypotheses as well - for example linking genes to diseases or compounds to treatments.

In a real world example, AstraZeneca used ML (not ChatGPT) to crawl through tons of biomedical text to find drug targets that they would have not considered otherwise. ML revealed a connection between things that might seem unrelated at first. Researchers can then take a closer look and see if the connection makes sense.

2. Hit Identification and Lead Optimization

ChatGPT could be used for performing virtual screening of large compound libraries, identifying potential ‘hit' molecules, or designing molecules de novo based on desired properties and structural features. For examples, we look at the work that similar models have done. These other models are trained not necessarily on natural language but other corpora.

AbSci recently published their work on the generative design of antibodies to target three different disease-associated molecules (HER2 receptor, VEGF growth factor, and spike protein of SARS-CoV-2). These antibody candidates were generated in a zero-shot manner, i.e. without the model being shown any existing binders during training and without any further optimization. About a million AI-designed binders were screened in a high-throughput wet lab platform and this process yielded a few altogether novel and plausible antibody candidates which, for the case of HER2, were found more effective than the current therapeutic trastazumab. This is claimed to be the first study that follows up generative antibody design with experimental validation and may be an important stepping stone towards rapid design of antibodies (antibody-based therapeutics accounted for 30% of FDA-approved biologics in 2022).

In another example, models have been developed for protein design. These are so called diffusion models. Diffusion models are capable of generating images from text descriptions. Most notable

recent developments in this area were OpenAI's DALL-E 2 and Stability AI's Stable Diffusion. Two recent studies – Chroma by Generate Biomedicines and RoseTTAFold Diffusion by an academic lab at the University of Washington – adapted diffusion models to generate protein designs with specified constraints on shape, size and function. The potential of such algorithms to direct protein generation was illustrated with a proof-of-concept design of a new protein which binds to the parathyroid hormone (a regulator of calcium levels in the blood) more strongly than known hypocalcemia drugs.

3. Preclinical Development

Recently IBM Research used deep generative autoencoders followed by virtual screening to demonstrate the development of broad-spectrum and safe antimicrobial compounds. The whole process from design to experimental testing was completed in only around 7 weeks, highlighting the potential of the computational pipeline to accelerate candidate drug discovery (which might take a few years with the existing methods). In recent work (under review), the team has applied their foundational generative model to discover inhibitors against two SARS-CoV-2 protein targets, and claim that 2 out of only 4 predicted candidates which were synthesized show desired activity in in vitro experiments (that's 50% initial success rate).

4. Clinical Development

ChatGPT by it's nature can generate new text. It can write Shakespearean poetry or write prose like Hemmingway. It only follows that a recent application of similar models is data augmentation. In biomedical settings data generation is costly and time-consuming. So artificially generated data can potentially mitigate small data problems, provide better coverage of edge cases, reduce data biases, and enable the development of reliable models. Pre-ChatGPT models were good at extracting known relationships but not at producing something new.

Genetic data is abundant for vanilla use-cases. But for niche populations it can be very expensive to generate or acquire. A proof-of-concept for artificial genomic data produced by generative neural networks was demonstrated by Gretel.ai in partnership with Illumina in 2021. The study replicated a genome-wide association study with synthetic mouse genotype data.

Generative models can also enrich data as complex as single cell data. CscGAN is a generative adversarial neural network that can augment sparse single-cell RNA-sequencing datasets with artificial cells. Augmented data was shown to improve downstream analyses like marker gene detection and the performance of classifiers trained on the single-cell profiles, and is suggested to reduce human and animal experimentation costs of transcriptomics data generation for disease research.

LLM like generative models can also help in designing clinical trials by suggesting populations, endpoints, and dosing regimens. It could aid in patient recruitment by studying health records and clinical data to identify eligibility for clinical trials. They could monitor real-time clinical trial data to identify trends, detect safety signals, and predict potential issues that can impact the trial's success.

In a nutshell, ChatGPT and similar technologies can be used to solve a variety of problems. Perhaps even more significantly, models trained using similar techniques - but on different kind of corpora - can produce novel information, bridge data gaps and make it easier to come up with new hypothesis.

Journey from PreGPT to ChatGPT

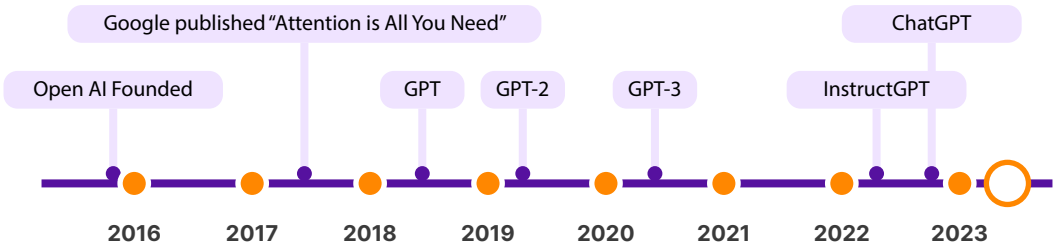
Natural Language Processing (NLP) has come a long way in the past few decades, with several key developments leading to the creation of advanced language models like ChatGPT.

During the early years of NLP up until the 2010s, researchers predominantly used rule-based and simple statistical methods to tackle language problems. They tried to solve issues like text summarization, machine translation, and sentiment analysis using techniques such as keyword extraction, n-grams, and regular expressions. The systems that relied on these methods required manual creation of rules and patterns, making them difficult to scale as the complexity of language increased. This was a labor - intensive process and restricted the usefulness of NLP systems.

TL;DR: The journey of Large Language Models (LLMs) from pre-GPT to GPT-4 has been marked by significant advancements. Early methods relied on rule-based and statistical techniques - making them not scalable. The breakthrough came with the release of GPT-2 in 2019, a large-scale language model with 1.5 billion parameters, which showcased the potential of deep learning in language generation tasks. GPT-3, released in 2020 with a staggering 175 billion parameters, set a new benchmark for NLP models, achieving state-of-the-art performance without fine-tuning.

The resurgence of deep neural networks (DNN) in the early 2010s paved the way for the rapid advancement of natural language processing (NLP) models, including the development of the first generation of GPT models. These early GPT models demonstrated the potential of deep learning in language generation tasks, but were limited in terms of model size and training data.

However, the breakthrough came with the release of GPT-2 in 2019, which pushed the boundaries of what was previously possible in language modeling. GPT-2 introduced a massive scale language model with 1.5 billion parameters, leading to remarkable improvements in language generation quality and accuracy. This model showcased the potential of large-scale language models for a wide range of applications, including text completion, question answering, and text generation. Moreover, GPT-2 also raised ethical concerns due to its potential for generating fake news and misinformation.



Building on the success of GPT-2, OpenAI released GPT-3 in 2020, which represented a significant leap in the capabilities of language models. With a staggering 175 billion parameters, GPT-3 set a new benchmark for large-scale language models, achieving state-of-the-art performance on a wide range of NLP tasks without task-specific fine-tuning.

The unprecedented size of GPT-3 allowed for more creative and versatile text generation, making it a powerful tool for content creation, language translation, and code generation, among others.

Demystifying GPT: GPT for Dummies

GPT (Generative Pre-trained Transformer) at its core is a large artificial neural network. Its a series of models released by OpenAI starting with GPT-1, GPT2, GPT3 and very recently, GPT4.

Dataset	# tokens	Proportion within training
Common Crawl	410 billion	60%
WebText2	19 billion	22%
Books1	12 billion	8%
Books2	55 billion	8%
Wikipedia	3 billion	3%

To get a clearer understanding of it, let's unpack each word in "Generative Pre-trained Transformer" one by one.

"Pretrained" in the context of Generative Pre-trained Transformers (GPT) refers to the initial training phase that the model undergoes on a large corpus of text data before it is fine-tuned for specific tasks. During training, GPT is tasked to predict the next word in a given sequence of words (sort of like a fill in the blank task). It is penalized if it gets these words wrong and rewarded if it gets them right. GPT is "Generative" in that after training, it can be used to generate coherent and contextually relevant text by sampling from these predictions.

E.g.

Input: "In a faraway land, there was a magical creature called the Glitterwing. The Glitterwing had the unique ability to..."

Output of GPT: "... change colors depending on its mood. When it was happy, its wings would shimmer in bright shades of gold and silver, casting dazzling reflections on the surrounding landscape."

The "Transformer" in GPT refers to the underlying neural network architecture used to build the model. The Transformer architecture, introduced by Vaswani et al. in a 2017 paper titled "Attention is All You Need," was especially designed handle sequence-based data, such as natural language text.

Unleashing ChatGPT: The Next Generation of LLMs

TL;DR: Key breakthroughs that enabled the success of ChatGPT, built on GPT3, include the scalability of transformer networks, the ability of large language models to perform tasks without task-specific training, and the use of human feedback to improve model usability.

OpenAI has not revealed the details of ChatGPT which is built on GPT4. GPT3 had 175 billion parameters and was trained on billions of words/tokens of text taken from the internet. Some have **speculated** that it has more than 1 trillion parameters - making it 6 times as big as GPT3. Size of the model isn't the only important thing though. The success of ChatGPT can be attributed to a few key breakthroughs that allowed researchers to overcome critical limitations:

1. Transformer Networks Scale Incredibly Well

Transformers were designed to address the computational limitations of previous sequence models like RNNs (Recurrent Neural Networks) and LSTMs (Long Short-Term Memory Networks). Sequence models are the machine learning models that input or output sequences of data. In this context, they are models used for producing natural language.

A key innovation of the transformer architecture is what is called 'self-attention mechanism'. Self-attention allows the model to weigh the importance of each word in a given sequence relative to the others, helping it capture context and long-range dependencies more effectively and computationally cheaply. Self-attention mechanism is highly parallelizable. This makes transformers computationally efficient during training compared to RNNs and LSTMs. Hence allowing researchers to train ever larger language models.

This efficiency meant that the you could train transformer models with billions of parameters. This wouldn't have come to much if the quality of output didn't increase. But as expected - by intuition - the quality of the model output also improved with the size of the parameters.

Language Model	Release Year	Parameters	Architecture
Word2Vec	2013	~16M-300M*	Shallow Feed-Forward Neural network
ELMo	2018	~94M	BiLSTM
BERT (BERT-Base)	2018	~110M	Transformer
GPT-2 (GPT-2 Large)	2019	~774M	Transformer
T5 (T5-11B)	2019	~11B	Transformer
GPT-3 (GPT-3 175B)	2020	~175B	Transformer

Note the significant increase in parameters after the introduction of transformer models.

**Note: The parameter size of Word2Vec models depends on the size of the vocabulary and the chosen embedding dimension. The range provided here is an estimate based on typical configurations.*

2. Ability To Work Without Task-specific Training

For the longest time, if you wanted a model to perform a task, you had to train the model on data that was specifically created for that task. GPTs have done away with that limitation. They are

what are called - Few Shot Learners. As the name suggests, GPTs can learn with few examples instead of needing a large training corpus.

The breakthrough came with GPT-3. The seminal GPT-3 paper, released in 2020 was titled "Language Models are **Few-Shot Learners**". The paper demonstrated that if you could train a sufficiently large language model, that model would get the ability to perform other tasks. **Even tasks it wasn't explicitly trained for** - with few shots or in other words a finite set of examples.

3. Human Feedback Can Improve Language Model Usability

ChatGPT is the same as GPT3 in terms of architecture and training. However, the difference in usability of the models is huge. GPT3 is harder to use because it tries to predict the next word in whatever input you give it. This means, that if you want it to write an email, you have to write the first section of the email and then hope that it will write the remaining. For users, this is an awkward way of getting it to do what you want it to do.

To overcome this limitation, OpenAI tuned the GPT3 model to follow instructions using a reinforcement learning procedure in which GPT3 would generate some output, that output would be rated by a human, and GPT3 would be rewarded based on that rating. This step, called Reinforcement Learning on Human Feedback was how ChatGPT was created starting from GPT3.

In short, GPTs are a significant improvement over state-of-the-art language models.

- They are not built on rule based training. Hence allowing them to generalize for multiple tasks.
- Advances in architecture have allowed for building models with 100s of billions of parameters, versus 10s of millions.
- Larger models have been shown to produce much higher quality outputs.
- GPTs can be trained on few examples for new tasks versus needing a large training set.
- ChatGPT can be used for new tasks with 0 examples just using prompts.

Given all the promise of ChatGPT we put it to task for what we do best - curation. The results are exciting - if not mind boggling. This is a narrow use case which goes to show just how wide ranging the possibilities are.

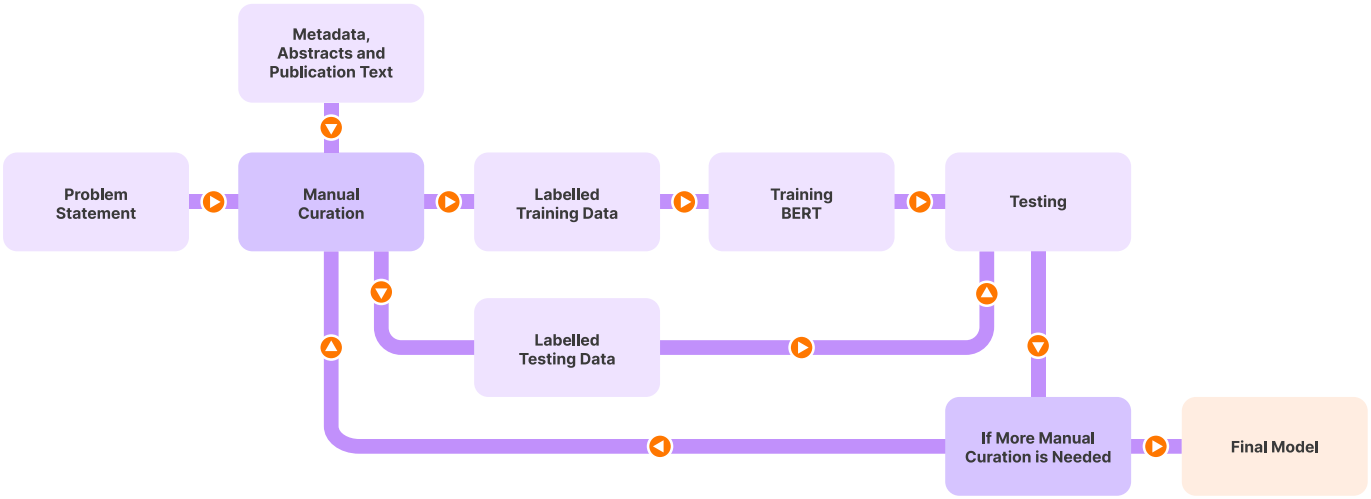
Elucidata + ChatGPT: A Curation Case-Study

Biomedical Data Curation at Scale and Potential Applications

TL;DR: At Elucidata, we use BERT-like models to identify biomedical entities from meta data abstracts and publications. While we've established a robust biocuration process, this poses limitations in terms of resources used and time spent in training BERT models. To address these limitations, we are experimenting with ChatGPT and prompt-based engineering to extract biomedical entities from publications. We have achieved an accuracy and F1 score close to 83%, for sample level disease extraction.

At **Elucidata**, our proprietary technology enables us to curate relevant biomedical data fields better than the state of the art. **Take a look at it here.** Our technology can extract metadata at the dataset and sample level, such as disease, drug, tissue, and others.

To automate and standardize the curation process, we leverage Bio-NLP to extract relevant entities from metadata, abstracts, and publications.



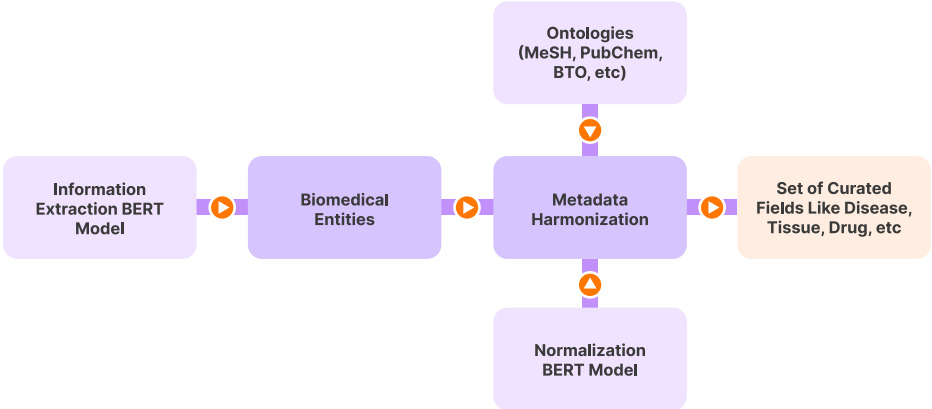
In the flow chart, one can see the high-level process of training a model:

- **Field definition and manual curation:** In this step, we define the field we will be extracting and create guidelines to generate the training data. The training data is labeled with a double-blinded review process. This gives us reliable data to train task-specific models.

- **Training a task-specific model:** After we have generated training data, we preprocess large corpus of texts (like publications) by breaking them down into smaller paragraphs. These are then used to train a tasks specific BERT models.

- **Testing and more manual curation:** With a different dataset - aka test data - we then evaluate the accuracy of models. If the models achieve desired performance, they can be used for curation in 'production'. More often than not though we have to run multiple training iterations or get more training data. A few iterations (5 to 15) are usually needed to get 'production ready models' - models ready to be thrown into the wild.

Lastly, we standardize the extracted information with specific ontologies (like MeSH, PubChem, BTO, etc). This gives us what are called 'normalized entities' - fancy geek-speak for terms picked from a regulated dictionary.



The flow chart above shows the process for getting normalized curated fields,

- **Information extraction using task-specific model:** During this step, we get relevant entities from the metadata, publications, and abstracts. This is done with the model we trained for the specific task.

- **Metadata harmonization:** We map extracted entities to standard ontologies using a model trained for just this purpose. The 'normalization' model maps extracted entities to ontologies.

This process has limitations:

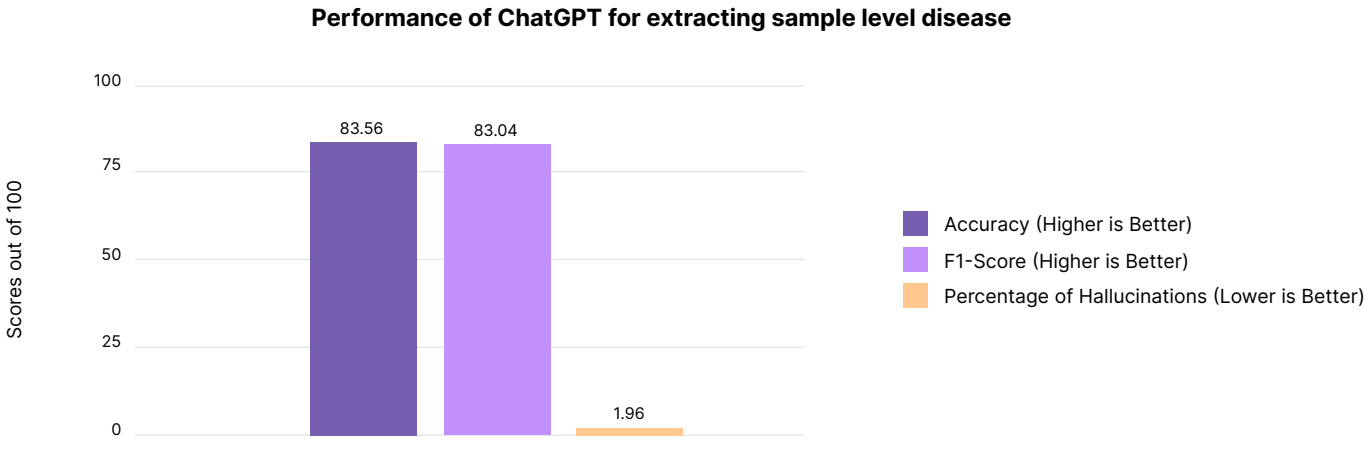
1. As the discerning reader might have noticed, the curation of each new field (like cell line, tissue etc.) requires new work to be done. In other words, more time and resources.

2. Moreover, pre-GPT models have limitations. They can not extract information from different parts of a paragraph and join them together as the answer. In ML-speak, they can't retain contexts.

What Has ChatGPT Got to Do with It?

Given ChatGPT's obviously advanced capabilities, we were curious whether OpenAI's ChatGPT could overcome some of these challenges. These are early days but we are happy to report that our initial work shows a lot of promise. The model training step is now replaced by prompt engineering, which will be supported by the curation guidelines created by experts. Prompt engineering in simple terms is the questions that we ask ChatGPT. Good prompts get better answers.

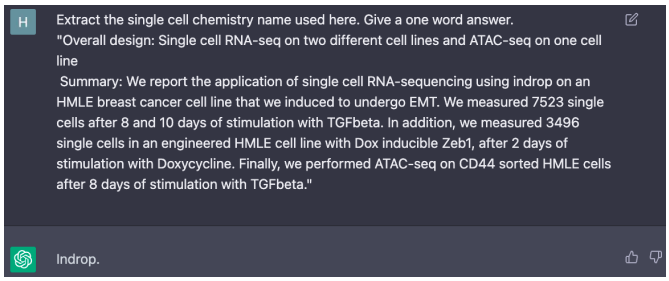
Some additional benefits of ChatGPT over existing models are that it can potentially be used to extract multiple fields in one go. Thus making the process computationally much cheaper.



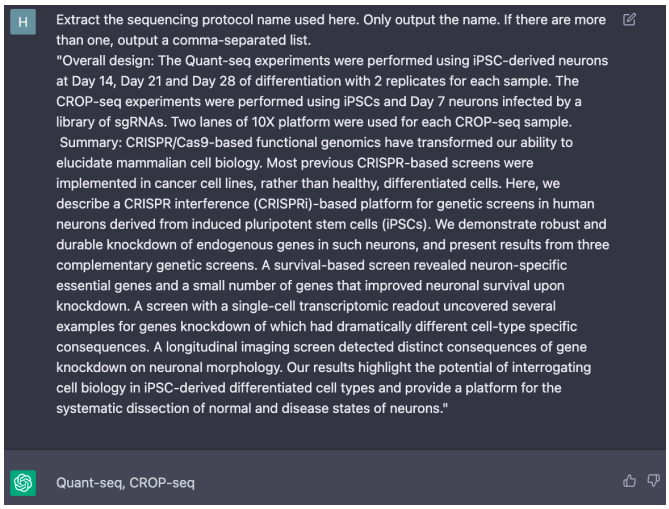
In short, with the advent of ChatGPT the metadata curation process at Elucidata might start looking very different. It could significantly decrease the barrier for extracting new information. Our experts' can spend their time doing high-level review rather than low-level labeling. These are early days still. That said, ChatGPT could, most likely will, help us save significant time and resources.

Some Examples of ChatGPT in Action for Curation

Metadata	ChatGPT	BERT
*We present **sc-FTDseq**, a microchip platform for single-cell freeze-thaw lysis directly toward 3' mRNA sequencing. It offers format flexibility with a much simplified, ...*	*'sc-FTDseq'*	*''*
*....separation. fluidigm c1 single-cell auto prep array for mrna-seq (10–17 µm) and **fluidigm c1** system were used for cell capture following the fluidigm protocol, ...*	*'Fluidigm C1'*	*''*
*Single-cell suspensions were prepared from 10-12 weeks C57Bl/6 mice and single cell RNA-seq libraries were generated using **3' V2 chemistry kit** on **Chromium Single cell** controller (**10x Genomics**).*	*'10X Genomics Chromium Single Cell 3' V2 chemistry kit'*	*'10x'*



ChatGPT successfully extracts single-cell chemistry used in a study



ChatGPT is able to extract multiple labels correctly

Limitations and Risks

As with any new technology, we are in the hype phase of ChatGPT. Expectations are through the roof. That said, there are certain limitations and risks associated with the use of ChatGPT and other LLMs. We'll leave you with a few pointers to think about:

1. **Data Quality and Accessibility**
   The performance of ChatGPT relies heavily on the quality and accessibility of data. If the data used for training is incomplete, biased, or inaccurate, the model's predictions may not be trustworthy.

2. **Lack of Experimental Validation**
   While ChatGPT excels at generating predictions and hypotheses, it will not replace the need to to conduct actual experiments. Experimental validation will be essential - as always - to confirm the predictions.

3. **Limited Biological Comprehension**
   Although ChatGPT can produce text that resembles human language, it lacks a deep understanding of the underlying biology of systems. It can't, at least not yet, replace the deep context that trained humans bring. It still does not have the capability to capture the intricate nature of biological systems.

4. **Model Hallucinations**
   ChatGPT, although creative most of the time, can generate output that is plausible-sounding but incorrect. This can happen due to various reasons such as the quality of training data, ambiguous user input, bias in the data or other reasons that we just don't understand just yet. It is essential to be cautious of the content generated by ChatGPT.

In Conclusion

As we look towards the future, ChatGPT is poised to play a pivotal role in transforming many industries - including biopharma and life sciences. Its ability to extract meaningful information, produce largely accurate text and generate novel ideas holds great potential for accelerating the pace of drug discovery and development.

However, these are very early days. It's not the end game! When we look back 10 years from now, this will be a very primitive version of what ML and LLMs can do. This is barely scratching the surface.

Let the games begin. :)

References

Journal articles/Reviews

1. Chen, W., Liu, X., Zhang, S., & Chen, S. (2023). Artificial intelligence for drug discovery: Resources, methods, and applications. *Molecular therapy. Nucleic acids, 31, 691–702*. **Artificial intelligence for drug discovery: Resources, methods, and applications**

2. Blanco-Gonzalez, A., Cabezon, A., Seco-Gonzalez, A., Conde-Torres, D., Antelo-Riveiro, P., Pineiro, A., & Garcia-Fandino, R. (2022). The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies. *arXiv preprint*. **The Role of AI in Drug Discovery: Challenges, Opportunities, and Strategies**

3. Sharma, G., & Thakur, A. (2023). ChatGPT in Drug Discovery. **10.26434/chemrxiv-2023-qgs3k**

4. Xue, V. W., Lei, P., & Cho, W. C. (2023). The potential impact of ChatGPT in clinical and translational medicine. *Clinical and translational medicine*, 13(3), e1216. **https://doi.org/10.1002/ctm2.1216**

5. Tong, Y., & Zhang, L. (2023). Discovering the next decade's synthetic biology research trends with ChatGPT. *Synthetic and systems biotechnology, 8(2)*, 220–223. **Discovering the next decade's synthetic biology research trends with ChatGPT**

6. Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., Dos Santos, C., Chen, P. Y., Yang, Y. Y., Tan, J. P. K., Hedrick, J., Crain, J., & Mojsilovic, A. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nature biomedical engineering, 5(6)*, 613–623. **https://doi.org/10.1038/s41551-021-00689-x**

7. Hie, B., Candido, S., Lin, Z., Kabeli, O., Rao, R., Smetanin, N., ... & Rives, A. (2022). A high-level programming language for generative protein design. *bioRxiv*, 2022-12. **A high-level programming language for generative protein design**

8. Shanehsazzadeh, A., Bachas, S., McPartlon, M., Kasun, G., Sutton, J. M., Steiger, A. K., ... & Meier, J. (2023). Unlocking de novo antibody design with generative artificial intelligence. *bioRxiv*, 2023-01. **Unlocking de novo antibody design with generative artificial intelligence**

Blogs

1. **The Role of GPT-4 in Drug Discovery by Andrew White**, March 2023

2. **Unleashing the Power of ChatGPT in Pharma** by Alberto Francesconi, March 2023

3. **ChatGPT knows how it could shape the future of healthcare** by Rob Verhuel, February 2023

4. **AI is dreaming up drugs that no one has ever seen. Now we've got to see if they work** by Will Douglas Heaven, February 2023

5. **Gretel.ai + Illumina - Using AI to create safe, synthetic datasets for genomics** by Alex Watson, March 2023

6. **How AI That Powers Chatbots and Search Queries Could Discover New Drugs** by Karen Hao, December 2022

7. **Biotech labs are using AI inspired by DALL-E to invent new drugs** by Will Douglas Heaven, December 2022

Videos

1. **Can ChatGPT do single-cell bioinformatic analysis?**

About Elucidata

Elucidata transforms biological discovery by providing high quality bulk RNA-seq and single-cell data, among other data types. They support discovery programs at top pharma companies and have 35+ research partners from premier biopharma companies and research labs.

Their FAIR biomedical data platform, Polly, makes data easily findable and more reusable. Elucidata has helped R&D teams scale up and has enabled 10x faster identification of therapeutic assets with high odds of success in the clinic. Having aided the detection of multiple validated drug targets across immunology, oncology, and metabolic disorders, Elucidata looks forward to helping more teams reach their R&D goals quicker!

For more information, visit **Elucidata | Home** or reach out to us at **info@elucidata.io**.

Locations/Offices

San Francisco
(Headquarters)

Cambridge

Delhi
(Tech Hub)

Bangalore
(Tech Hub)