Paper CM05

GPTs and LLMs in the context of Life Sciences and Clinical Trials

Jérôme Vétillard, Microsoft, Paris, France
Mark Lambrecht, SAS, Brussels, Belgium

SUMMARY

Generative artificial intelligence (AI) shows transformative potential in life sciences. Models like generative adversarial networks, variational autoencoders, and transformers synthesize data to expand scientific frontiers. They generate medical images, molecules, proteins, and text at scale. Key applications include accelerating drug discovery, personalizing medicine, powering clinical trials, and automating data analysis. GANs develop novel molecules efficiently. VAEs augment healthcare datasets. Transformers contextualize genomic and clinical data. Data scarcity, regulations, and ethics present challenges and safeguards. Collaborative solutions address quality and representation issues. Documentation and oversight ensure privacy, transparency, and fairness. Harmonizing global standards balances innovation and safety. Opportunities include pharmacogenomics revolutionizing care through genetic profiles. COVID vaccines exemplify AI roles in structure prediction and supply optimization. Decentralized trials improve through remote data collection and predictive modeling.

Clinical data automation enhances through data synthesis, augmentation, and predictive models. SAS works on a generative AI programming aid that generates workflows while preserving expertise. Solutions uphold evidence-based decisions through rigorous validation and human review. Responsible progress requires vigilance. Regular evaluation and stakeholder involvement identify unintended impacts for course correction. Cooperation cultivates understanding to realize technology's benefits judiciously, guided by priorities of trust, ethics and communities' self-determination. With persevering commitment to diversity and justice, generative AI's application nurtures healthcare serving all. Continual learning promotes aligning innovation and humanity for society's shared well-being. The future remains luminous through collaboration.

INTRODUCTION

The convergence of artificial intelligence (AI) and life sciences holds tremendous potential for revolutionizing healthcare and scientific research. While AI's roots date back to the mid-20th century, applying it to life sciences more recently ushered in groundbreaking advancements. Machine learning algorithms now lend unprecedented insights by parsing immense biological datasets, while neural networks accelerate drug discovery at an exponential pace.

Large language models and self-supervised transformers have also emerged, reshaping natural language capabilities across industries. Their contextual understanding powers myriad applications from summarization to question answering. As interdisciplinary collaboration grows, so too does our grasp of complex living systems and their intricate molecular underpinnings.

This article explores the dynamic applications of generative AI - a subset focused on synthesis - within life sciences. Special attention is given to clinical data automation, a critical area for optimizing patient care based on empirical evidence. The following sections outline generative techniques, use cases, opportunities, and responsible path forward for this promising yet nuanced field.

FUNDAMENTALS OF GENERATIVE AI

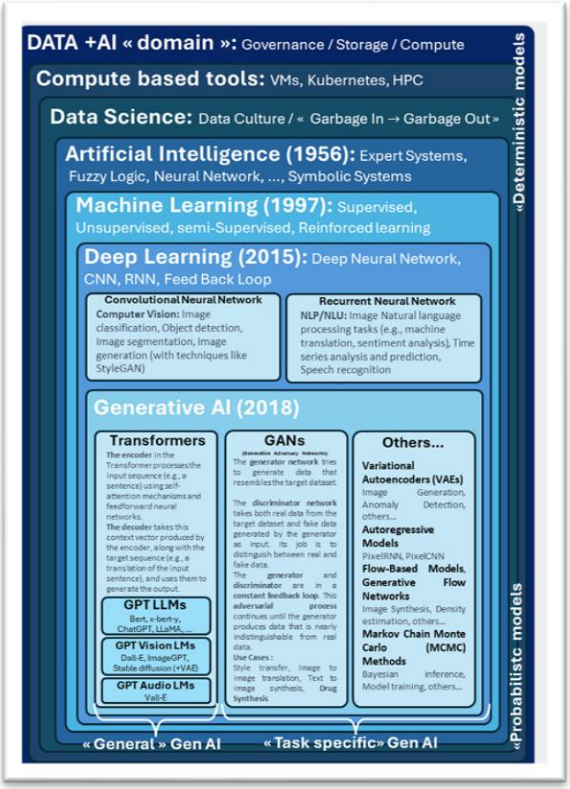
Generative AI algorithms emulate complex patterns obscured in nature's chaos. Trained on comprehensive (very large) datasets, these models generate synthetic yet statistically indistinguishable alternatives. Though limited by training constraints, their synthetic outputs expand scientific frontiers.

Variational autoencoders probabilistically encode inputs, empowering molecular design with endless structural combinations. Generative adversarial networks craft photorealistic simulations to supplement scarce real-world samples. Transformers process language contextually through self-attention, revolutionizing diverse natural language tasks.

These algorithms unlock imaginative possibilities. But responsible use demands understanding limitations and biases inherent to synthetic creations. With judicious application and domain expertise, generative AI's potential to further knowledge and improve lives knows no bounds.

DIFFERENT KINDS OF GENERATIVE AI ALGORITHMS

Figure 1 : AI evolution through the years
(Image from Jérôme Vetillard 2023)



Generative Adversarial Networks (GANs): GANs consist of two neural networks—a generator and a discriminator—that work in tandem. The generator produces synthetic data, while the discriminator evaluates it. The two networks are trained together in a sort of "cat-and-mouse" game until the generator produces data that the discriminator can hardly distinguish from real data.

Variational Autoencoders (VAEs): VAEs are probabilistic generative models that encode input data into a latent variable space and then decode it to reconstruct the data. They are particularly useful for tasks that require a generative process but also need the model to learn the data distribution explicitly.

Transformers: Initially designed for NLP tasks, transformers have become a cornerstone in the field of generative models. They employ a self-attention mechanism that allows them to consider other parts of the input when processing an individual data point.

This has led to significant advancements in NLU and NLP, enabling more coherent and contextually relevant text generation. Transformers are the backbone of many Large Language Models (LLMs) that are capable of generating human-like text.

Others: There are other types of generative models like Restricted Boltzmann Machines (RBMs), Long Short-Term Memory Networks (LSTMs) for sequence generation, and more recent innovations in generative modeling.

APPLICATION DOMAINS

Life sciences now benefit abundantly from generative techniques. Variational autoencoders enhance healthcare imaging while exploring molecular diversity in silico. Generative adversarial networks augment rare medical cases and simulate personalized treatment responses. Transformers capture proteomic intricacies through sequence modeling.

VAEs:

- Image Generation: Used in tasks like denoising and super-resolution.
- Data Augmentation: Enhances bioinformatics and healthcare datasets.
- Anomaly Detection: Flags deviations in industrial time-series data.
- Drug Discovery: Generates new molecular structures.

GANs:

- Image Synthesis: Handles tasks from image translation to art creation.
- Data Augmentation: Boosts medical imaging datasets.
- NLP: Applied to text generation, often with other architectures.
- Simulation: Models environmental conditions in engineering and robotics.

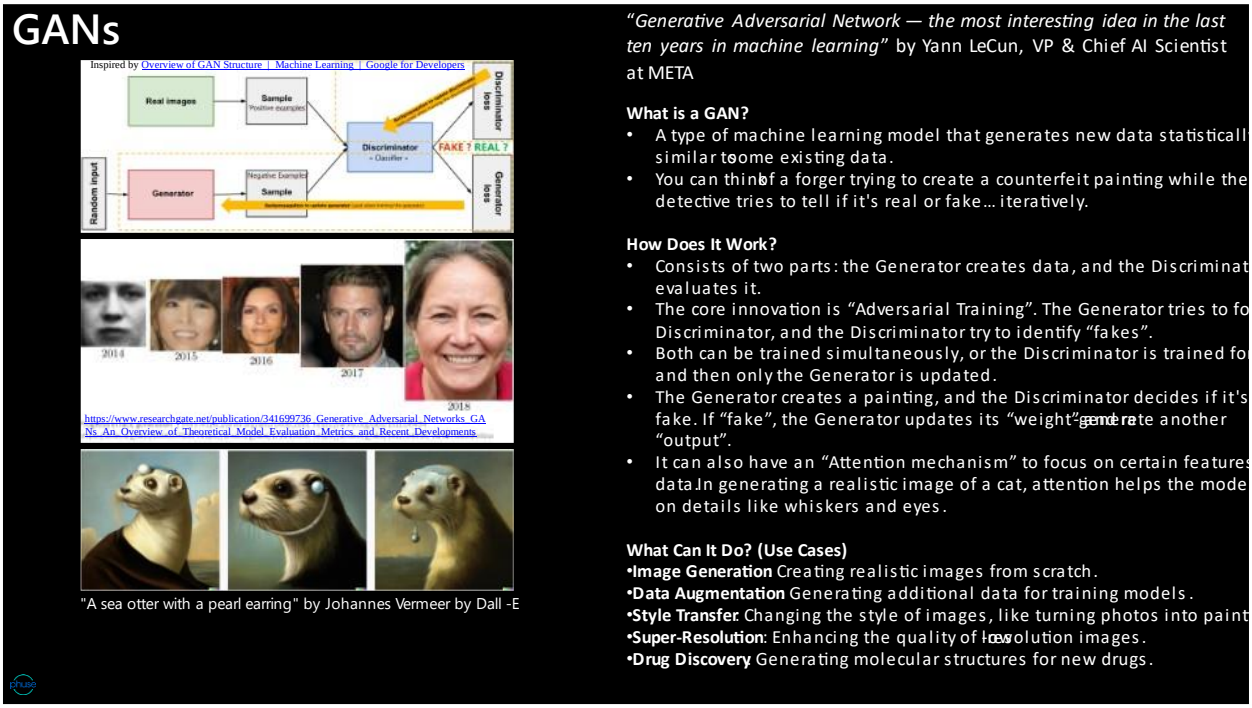


Figure 2 : Into GAN architecture and capabilities

Transformers:

- NLU: Powers tasks like sentiment analysis and summarization.
- Text Generation: Underpins Large Language Models.
- Multimodal Learning: Handles diverse input types, like images and text.
- Bioinformatics: Applied to sequence alignment and protein folding.

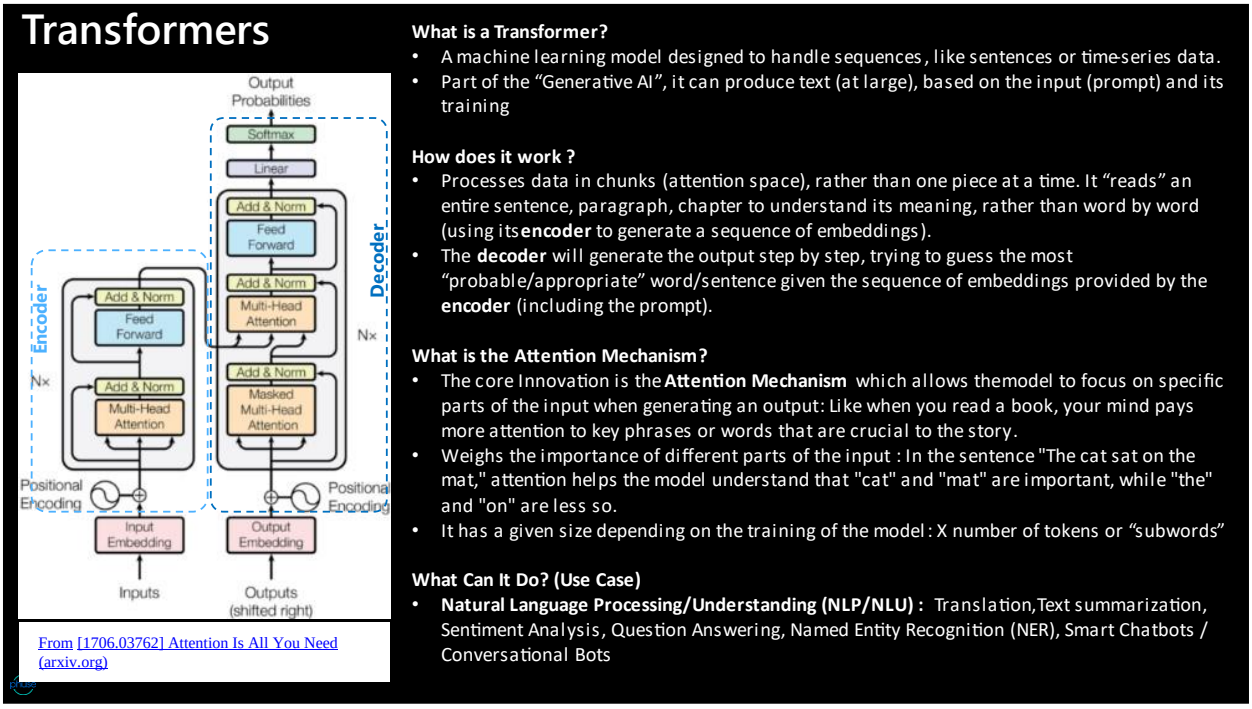
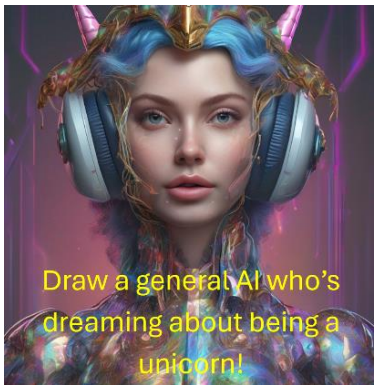


Figure 3 : Into transformers' architecture and capabilities

Generative AI has also been widely and historically applied to domains that are external to Health Care & Life Sciences.



Art & Design: AI like OpenAI's Dall-E crafts artwork and music.

NLP: Transformers revolutionize translation and summarization tasks.

Finance: Generative models aid in risk assessment and pattern recognition for trading.

Autonomous Vehicles: GANs generate synthetic data to simulate diverse driving conditions for training, though certification for synthetic data remains a challenge.

CURRENT STATE OF LIFE SCIENCES

While opportunities abound, certain challenges also impede progress. Data quality and accessibility remain problematic, necessitating novel collaborative solutions. Ethical applications of AI demand privacy, transparency, and fairness to uphold patient trust.

Yet interdisciplinary teamwork opens new paths. Big data insights now guide personalized therapeutics through pharmacogenomics. Accessible clinical trials embrace diverse populations. Data-driven decision-making transforms healthcare with rigor and care. Responsibly navigating these shifts, technology and human insight fuse to elevate society.

CHALLENGES

DATA AND ITS IMPACT

Scarcity of quality data presents a key challenge, especially in specialized fields like genomics. This lack of data impacts machine learning model generalizability and reliability, hampering emerging areas like personalized medicine. Developing regions face exacerbated issues due to less robust collection infrastructure. Scarce biomedical and genetic data has a big impact on patient privacy, a major ethical concern.

However, collaborative data sharing offers solutions. Initiatives building comprehensive public resources benefit all populations. Sensitivity to cultural representation enhances datasets' validity for diverse communities.

EDUCATION AND ETHICS

An ethical approach to generative AI in healthcare demands vigilance with sensitive patient information. Generative A can exacerbate many of the issues traditionally addressed within clinical bioethics. Issues like informed consent, privacy and potential misuse require ongoing dialogue between innovators, regulators and the public.

Cross-disciplinary training cultivates awareness. Curricula blending fields from data science to healthcare equip partners navigating complex realities and envisioning equitable opportunities.

RESPONSIBLE EVOLUTION

From "Ethical AI" the discourse matured to "Trustworthy AI" through documentation, transparency and accountability. Vigilant self-monitoring and willingness to engage diverse voices detects potential harms. Ensuring technologies augment rather than replace human expertise upholds care quality. Together, addressing data access, education, regulation and responsible design cultivates understanding between communities and innovation. Partnership overcomes challenges through cooperative spirit attuned to humanity.

OPPORTUNITIES

ADVANCING PRECISION HEALTHCARE

Personalized medicine revolutionizes care through targeted therapies informed by an individual's unique genomic and clinical profiles. Pharmacogenomics reduces adverse reactions through genetic testing, as seen with DPD deficiency treatment. However, marginalized groups encounter obstacles accessing these advances. Partnerships educate underserved communities and establish inclusive research participation models overcoming socioeconomic barriers to predictive prevention and intervention.

ACCELERATING DISCOVERY

Generative models streamline drug development by focusing libraries on optimized candidates versus exhaustive screening. This accelerated COVID-19 vaccine development, demonstrating AI's vital roles.
Yet data scarcity hinders some disease areas. Open innovation incentivizes data sharing between diverse partners representing all populations and contexts for equitable model generalization. Standards balance proprietary and public interests.

OPTIMIZING EVIDENCE

Real-world data augments clinical trials through AI-driven insights into diverse treatment responses over time. Decentralized methodologies make participation convenient through remote technologies.
However, algorithms risk perpetuating biases if underrepresented groups are not considered. Vigilant evaluation and feedback ensure fairness while preserving informed consent and privacy. Empathy guides technology as a tool, not replacement, for human expertise.

APPLICATIONS FOR GENERATIVE AI IN LIFE SCIENCES

The following sections outline key domains where generative techniques augment scientific processes:

DRUG DISCOVERY

The conventional methods of drug discovery, which primarily rely on combinatorial chemistry and high-throughput screening, are increasingly being viewed as laborious, expensive, and time-consuming. Generative AI technologies are emerging as transformative tools that offer more efficient, targeted, and cost-effective approaches to drug discovery.

Digitization of Molecular Structures: One of the first steps in integrating AI into drug discovery is the digitization of molecular entities. Algorithms like the Simplified Molecular Input Line Entry System (SMILES) have become instrumental in this regard. They convert intricate molecular structures into a more straightforward line notation, thereby facilitating easier manipulation by AI algorithms. This digitization acts as a crucial bridge between the realms of traditional chemistry and modern computational methods.

AI-Driven Molecular Design: Traditional combinatorial chemistry often results in a large library of compounds that necessitate extensive and costly screening. In contrast, AI models like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs) are designed to generate molecular structures with a higher probability of exhibiting desired pharmacological properties. By learning from existing molecular data, these AI models can focus on generating more promising candidates, thereby reducing the search space and conserving resources.

Virtual vs. High-Throughput Screening: The conventional high-throughput screening methods involve testing thousands of compounds for desired biological activity, a process that is both resource-intensive and often yields a low hit rate. Generative AI models, particularly transformers with attention mechanisms, can analyze existing biological data to predict how different compounds will interact with specific biological targets. This targeted approach significantly reduces the number of compounds that need to be synthesized and tested in the lab.

Challenges and Limitations: While AI-driven methods offer a more focused approach, they are not without limitations. For instance, the AI model might not generate a groundbreaking new entity that has no existing equivalent in the training data, emphasizing the need for diverse and comprehensive training data.

Biologics and Screening: In the field of biologics, the generation of new entities is often less challenging than screening them for specific functionalities and potential toxicity. Generative AI models can automate this screening process, which is particularly valuable in the development of monoclonal antibodies where functional specificity and safety are of utmost importance.

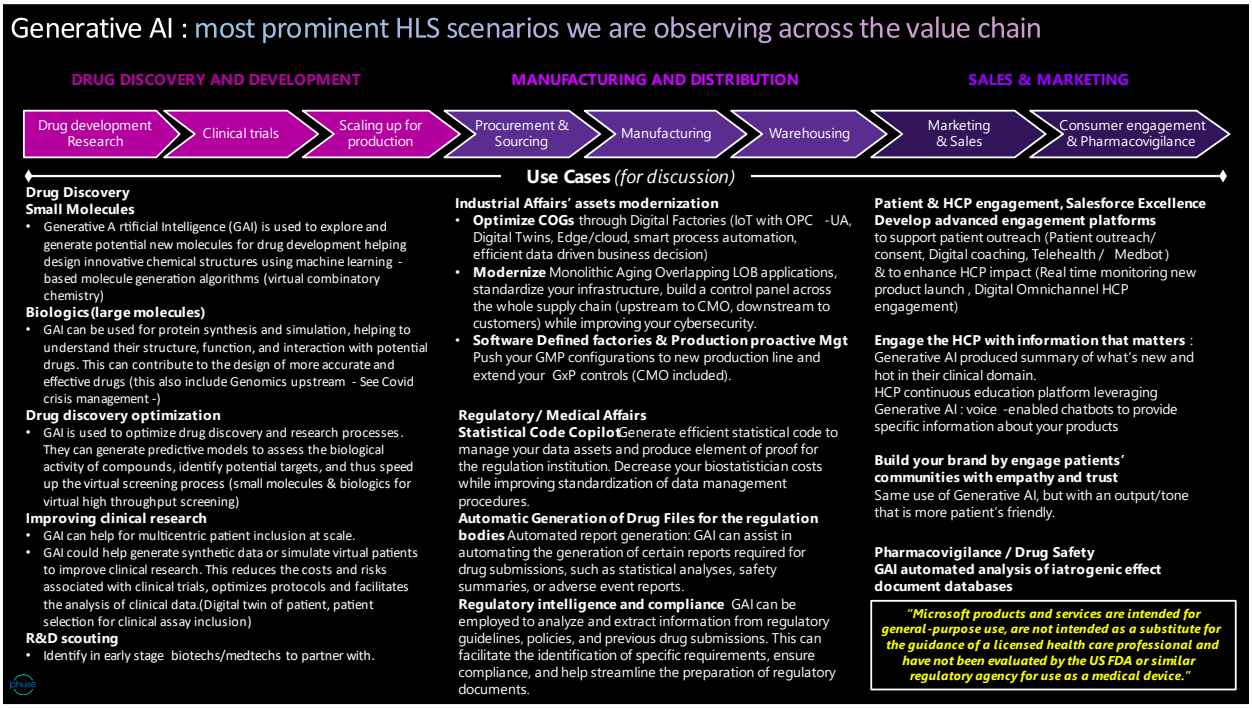
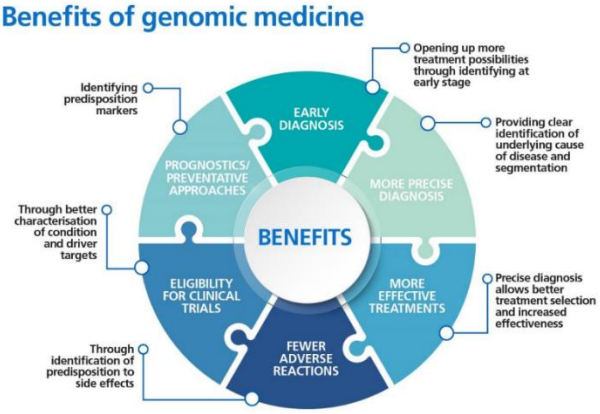


Figure 4 : Most prominent HLS scenarios we are observing across the value chain in Life Sciences

Computational Assessment: Whether employing traditional methods or AI-driven approaches, the ability to assess the biological activity, toxicity, and synthesizability of generated entities through computational methods is crucial. This 'in silico' evaluation is an essential step before these entities proceed to in vitro or in vivo testing, serving to accelerate the overall drug discovery process.

GENOMICS & PROTEOMICS



Source : NHS

Genomics has been a cornerstone in understanding the molecular mechanisms of diseases and has paved the way for personalized medicine. Traditional genomics research often relies on large-scale sequencing projects and manual annotation, which are both time-consuming and resource-intensive. Generative AI is emerging as a powerful tool to augment these traditional methods.

When sequence data is limited, generative techniques fill gaps. They model epigenetic factors to predict ancestral lineages or rare variant impacts. Sequence generation enhances clinical diagnostic searches.

Transformers contextualize relationships between non-coding and regulatory RNAs, protein coding regions, and environment to illuminate disease mechanisms. Modeling gene networks simulates functional impacts of mutations for precision medicine insights.

SEQUENCE GENERATION

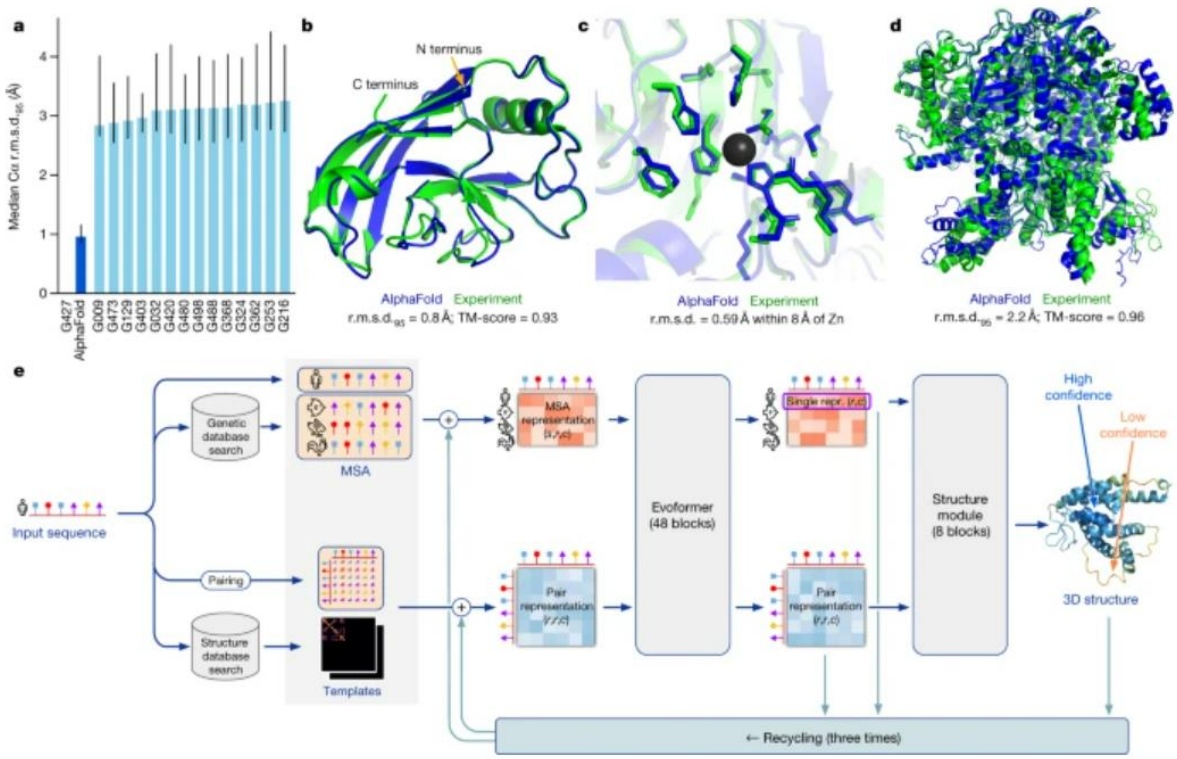
Generative AI models like GANs and VAEs offer a powerful tool for creating synthetic genomic sequences. While crafting sequences from A, T, G, C nucleotides may appear simple, the complexity arises in generating biologically relevant sequences or valid controls for experiments. These AI models excel in learning the statistical nuances of actual genomic data, enabling them to produce synthetic sequences virtually indistinguishable from real ones. For Proteomics, it is a matter of 20 amino acids to assemble.

Generative AI, especially transformers, can streamline the intricate task of understanding gene functions and interactions by leveraging their advanced attention mechanisms. These models excel in capturing complex gene relationships, thus enhancing functional genomics research. Additionally, they can integrate epigenetic factors like DNA methylation and histone modification into their analyses. This offers a comprehensive perspective on gene function and regulation, which is crucial in studying diseases like cancer where epigenetic shifts are functionally significant.

Proteomics is key in cellular biology and cellular physiology, playing a pivotal role in both personalized medicine and drug discovery. Traditional approaches, such as mass spectrometry, have been the mainstay for protein analysis but are often labor-intensive and may have limitations in comprehensiveness. The advent of Generative AI technologies is poised to transform this landscape, offering more efficient and expansive methods for both protein analysis and design.

IN SILICO SCREENING

Fig. 1: AlphaFold produces highly accurate structures.



*Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature **596**, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2 (image distributed under a creative commons attribution 4.0 International License)*

RoseTTAFold is another neural network-based approach that employs three-track attention layers to predict protein structures. It's particularly useful for proteins that lack similar structures in existing databases.

Here are a few commonly used techniques for evaluating binding force and **how HPC (High Performance Computing)** can assist in these processes:

- **Molecular Docking**: This computational approach aims to predict the optimal orientation and affinity of a ligand (often a protein) within the binding site of a target receptor. It involves a systematic exploration of possible conformations and orientations of the ligand. High-Performance Computing (HPC) significantly accelerates these calculations, allowing for the utilization of more sophisticated docking algorithms. This results in extensive sampling and more accurate scoring of potential binding poses, thereby increasing the reliability of the predictions.
- **Molecular Dynamics (MD) Simulations**: MD simulations offer a dynamic view of protein-receptor interactions by modeling the motion and interactions of atoms within the system over time. These simulations provide valuable insights into binding stability, ligand diffusion, and conformational changes that may occur during the binding process. Due to the computational complexity involved in these simulations, HPC is indispensable. It enables large-scale parallel computing, which is crucial for obtaining accurate and efficient simulation results.
- **Free Energy Calculations**: These calculations are designed to estimate the binding affinity or strength between a protein and a receptor. They involve complex algorithms, such as molecular mechanics force fields and advanced sampling techniques like enhanced sampling or alchemical methods. HPC plays a crucial role in executing these computationally intensive calculations efficiently, thereby allowing for more accurate and reliable predictions of binding affinity.
- **Quantum Mechanics/Molecular Mechanics (QM/MM) Simulations**: This hybrid approach combines the accuracy of quantum mechanics calculations for the binding region with the computational efficiency of molecular mechanics for the surrounding environment. QM/MM simulations can provide detailed insights into electronic properties, bond formation or breaking, and reaction mechanisms involved in the binding process. Given the computational demands of quantum mechanical calculations, HPC is essential for running these simulations efficiently.
- **Microsoft Quantum Elements** serves as a comprehensive Software Development Kit (SDK) and platform, primarily designed to advance the field of quantum computing. However, its utility extends beyond that, offering valuable tools and resources for quantum chemistry simulations and quantum-inspired algorithms.
  While the platform may not provide out-of-the-box functionalities specifically tailored for Quantum Mechanics/Molecular Mechanics (QM/MM) simulations, it does offer a robust foundation upon which researchers and developers can build.
  The QM/MM simulations are inherently complex and computationally demanding, requiring a blend of expertise in quantum mechanics, molecular mechanics, and programming. Microsoft Quantum Elements addresses this challenge by providing the capability to run quantum-inspired algorithms and simulations on quantum computers, a service colloquially referred to as "Qbits as a Service." This feature is particularly beneficial for those engaged in high-level research that demands significant computational resources.

In the broader context of molecular biology and drug discovery, High-Performance Computing (HPC) has been instrumental in enabling researchers to handle large datasets and perform complex algorithms and calculations, particularly in the evaluation of protein-receptor binding forces. HPC has revolutionized this field by significantly accelerating the time required for these evaluations, thereby facilitating the discovery and optimization of potential drug candidates.

Microsoft Quantum Elements aims to take this a step further. By utilizing quantum algorithms that run on quantum computers, the platform seeks to go beyond the current capabilities of HPC. This opens up new possibilities for computational efficiency and accuracy, potentially revolutionizing how we approach complex problems in molecular biology, drug discovery, and other scientific domains.

High-performance computing (HPC) accelerates protein-receptor binding analysis by handling large datasets and complex algorithms. It enhances research accuracy and speeds up drug discovery. Microsoft Quantum Elements aims to surpass HPC capabilities by offering quantum algorithms for even faster, more precise evaluations.

TAILORED THERAPEUTICS THROUGH PRECISION PROFILING

Personalized medicine aims to match treatments to individuals' unique biological and clinical profiles. By analyzing genomic and "omics" data, providers can pinpoint genetic predispositions, mutation risk levels, and biomarker expressions specific to each person.

Generative modeling strengthens this precision approach. Generative adversarial networks simulate diverse patients' likely responses and side effects to different therapies based on their molecular and lifestyle profiles. This helps identify the options most optimally suited for each individual.

Transformers take this a step further by contextualizing enormous genomic, medical record, and literature datasets. Their attention mechanisms excel at surfacing novel treatment-patient correlations across this complexity. Clinicians leverage these tailored insights when guiding therapy decisions.

PREDICTIVE ANALYTICS FOR PROACTIVE CARE

Beyond reacting to present illnesses, personalized care requires anticipating disease progression. Generative models could analyze patients' longitudinal health and exposure data to forecast risk trajectories over their lifetimes.

Time-series and natural language processing techniques parse clinical notes, lab results, images and other multidimensional records. This unearths subtle patterns predicting future conditions, enabling proactive discussions around prevention and early intervention.

MULTI-OMICS INTEGRATION FOR COMPREHENSIVE PROFILING

Achieving true personalization demands considering interactions across biological layers from genome to metabolome. Deep learning integrates diverse "omics" data sources along patients' developmental timelines.

Such holistic profiling illuminates gene-environment interplays driving health statuses. It fuels predictive analytics optimizing screenings, lifestyle modifications and precision therapeutics on individual schedules.-AI Collaboration for Validation and Implementation

While generative systems power sophisticated modeling at scale, clinical expertise remains crucial. Doctors partner technology by reviewing simulations against other data sources.

Their specialized knowledge contextualizes population-scale findings for applicability to specific patients. Joint perspective advances care quality by fusing comprehension from lived experience and advanced analytics.

Together, these personalized analytics partnerships usher precision medicine into reality through cooperation ensuring technologies augment rather than replace human judgment or equitable access. Co-creation nurtures healthcare serving all communities according to their unique needs.

REAL-WORLD EVIDENCE (RWE)

Data Sources: RWE is derived from real-world data, which includes electronic health records (EHRs), claims and billing data, and data collected through wearable devices if any.

- Generative AI in RWE: Generative models can analyze these vast and diverse data sets to identify patterns and correlations that may not be evident in controlled clinical trials. This can be particularly useful for tailoring treatments to subpopulations of patients with unique characteristics.
- Applications: RWE can be used to monitor drug safety, assess treatment effectiveness, and even to support regulatory decisions. Generative AI can accelerate these processes by generating synthetic but statistically valid datasets for analysis.

AMBIENT CLINICAL INTELLIGENCE

- Microsoft Nuance: Nuance's ambient clinical intelligence technology uses advanced NLP and machine learning algorithms to capture and interpret clinical conversations, thereby automating the documentation process, freeing HCP time (accelerating the paperwork), and providing more space for the HCP to patient interpersonal relationship.
- Personalized Treatment: By integrating with EHRs, these technologies can provide real-time, personalized treatment suggestions based on both current medical literature and the patient's medical history.
- Efficiency and Accuracy: Such technologies not only streamline administrative tasks but also reduce errors, thereby improving the quality of personalized healthcare. They also enable the whole healthcare system to produce high quality clinical data that can be used in return to train high performance clinical Artificial Intelligence.

DECENTRALIZED CLINICAL TRIALS

Decentralized Clinical Trials (DCTs) represent a transformative approach to clinical research, leveraging digital technologies to facilitate remote patient participation. This model is particularly advantageous as it minimizes the logistical burdens on healthcare systems and enhances patient accessibility to trials.

Key Components and Advantages:

- Data Collection with ECOA: Electronic Clinical Outcome Assessments (ECOA) are often integrated into DCTs to collect patient-reported outcomes (PROs). These electronic methods standardize the data collection process, thereby reducing human errors and enhancing the overall efficiency of the trial.

- Real-time Insights: The immediacy of data capture through ECOA devices provides researchers with real-time insights into patient experiences and treatment outcomes. This timely data collection is invaluable for detecting trends or shifts in patient conditions more rapidly than traditional methods.
- Enhanced Patient Engagement: The use of ECOA devices, such as electronic diaries or mobile apps, significantly boosts patient engagement and compliance. Automated reminders and notifications further ensure a high rate of timely data submission from participants.
- Data Quality and Integrity: ECOA devices come with built-in data validation checks, which minimize the likelihood of missing or incomplete data. This feature enhances the overall quality of the data and reduces the need for subsequent data cleaning or corrections.
- Operational Efficiency and Cost-effectiveness: The electronic nature of ECOA eliminates manual data entry and reduces paperwork, thereby streamlining the data management process. This operational efficiency translates into cost savings for both researchers and participants.
- Remote Monitoring Capabilities: The ability to remotely monitor patient experiences and outcomes is especially beneficial in DCTs, where in-person visits to study sites may be infrequent or non-existent. This feature makes trials more accessible and convenient for a broader range of participants.
- Data Security and Regulatory Compliance: ECOA devices are designed with robust data encryption and secure storage capabilities. They adhere to relevant privacy regulations, ensuring the confidentiality and protection of patients' personal health information.
- Inclusion and Diversity: One of the most compelling advantages of DCTs is their potential to increase participant diversity. By eliminating geographical constraints, DCTs can include individuals who might otherwise be excluded from traditional clinical trials.
- Role of Generative AI: Generative AI technologies could further enhance the efficiency and effectiveness of DCTs by helping to automate various aspects of the trial process, from patient recruitment to data analysis.

DCTs, when augmented with eCOA and Generative AI, offer a more efficient, inclusive, and patient-centric approach to clinical research. They hold the promise of accelerating the advancement of evidence-based medicine and enriching our understanding of treatment impacts on patient lives.

ROLE OF GENERATIVE AI IN LIFE SCIENCES (SUMMARY OF USE CASES)

Data Augmentation: Generative models like GANs can create synthetic but statistically valid patient data, aiding in the preliminary stages of trial design.

Predictive Modeling: Generative AI can be used to simulate different trial scenarios, thereby helping researchers to optimize trial protocols and logistics.

Adaptive Trials: Generative AI can dynamically adjust trial protocols based on real-time data, making the trials more flexible and efficient. It is also solving a moral issue of clinical trials by enabling the patients in the placebo arm of the clinical trial to benefit from the innovative treatment as soon as possible (not waiting for the end of the trial).

All these features might be enabled by state-of-the-art digital / statistical platforms such as SAS Viya and SAS Software in general. SAS software provides robust statistical tools that can analyze both real and synthetic data generated by generative AI models, ensuring the integrity of the trial results. SAS Viya offers advanced data management capabilities, allowing for the secure and efficient handling of large datasets commonly associated with DCTs. SAS Viya offers real-time analytics features, which are crucial for adaptive trials that require dynamic adjustments.

These are some of the ways in which generative AI could possibly complement the rigorous capabilities of robust analytical technology frameworks such as SAS Viya;

- GANs: Primarily used for generating novel molecular structures. They can also be used in tandem with other machine learning models to predict the biological activity of the generated compounds.
- Transformers: Particularly useful in sequence-based drug discovery methods. Their attention mechanisms make them adept at capturing the complex relationships between different biological molecules.
- VAEs: Often used for generating molecular structures with specific properties. They are also useful for tasks that require an understanding of the entire data distribution, such as outlier detection in compound libraries.

While generative AI holds immense promise, challenges like data scarcity, model interpretability, and validation of AI-generated compounds remain. Ongoing research is focused on overcoming these challenges to make AI-driven drug discovery more robust and reliable. The future (like it was made for Alphafold) might be the extension of “attention mechanisms” (inherited from the transformers’ architecture) to other Deep Neural Network architecture (including CNN like that are more into “computer vision”) to provide multimodal AI with general attention (cross modality) capabilities.

CLINICAL DATA PROCESSING AUTOMATION

Generative AI serves as a critical component in the automation of clinical data management, offering a multi-faceted approach to enhance the efficiency and accuracy of healthcare processes.

- Drawing inspiration from existing AI-assisted tools like GitHub Copilot, which aids developers in expediting code development, SAS is in the process of developing its own Copilot like functionality. This tool aims to assist biostatisticians in generating functional and relevant statistical code, thereby significantly accelerating the data analysis process on the SAS Viya platform.
- In the realm of Synthetic Data Generation: Generative AI algorithms have the capability to synthesize clinical data that closely mimics real-world patient profiles. This synthesized data is invaluable, especially in scenarios where there is a lack of sufficient real-world data. It serves as a robust resource for training machine learning models, thereby enhancing their predictive accuracy.
- Using Synthetic Generation to augment existing data: Generative AI can augment existing clinical datasets by generating additional data points that are consistent with the original data distribution. This augmentation process not only increases the size of the dataset but also enhances its diversity. As a result, machine learning models trained on such augmented datasets exhibit improved performance and are more generalizable to new, unseen data.
- Predictive Modeling: Generative AI algorithms can sift through complex clinical data to generate predictive models. These models can analyze intricate patterns and relationships within the data, enabling healthcare providers to make more accurate diagnoses, forecast disease progression, and even tailor treatment plans to individual patient needs.

SAS IS WORKING ON A "COPILOT" LIKE FUNCTIONALITY

SAS is developing a "copilot" like functionality as an AI-powered assistant that can automatically generate statistical code and might lead to more standardization in the way clinical data is managed. A copilot like function for programming could offer the following functionalities;

- Automated Statistical Code Generation: generating statistical code for data cleaning, transformation, and analysis. This eliminates the need for manual coding and reduces the risk of human errors. Researchers and data analysts can focus more on interpreting the results rather than spending time on coding.
- Efficient Data Cleaning: automating data cleaning processes by identifying and handling missing values, outliers, and other data quality issues. This ensures that the clinical data used for analysis is accurate and reliable.
- Standardized Analysis: promoting standardization in the way statistical analyses are conducted, e.g., based on CDISC standards. By generating code based on best practices and industry standards, it ensures consistency in analysis methods and results. This is particularly important in multi-center clinical trials or collaborative research projects.
- Automated Reporting: generating automated reports that summarize the key findings and insights from clinical data analysis. These reports can be customized to meet specific requirements and can save significant time and effort in the reporting process.
- Efficient Collaboration: facilitating collaboration among researchers and data analysts by providing a common platform for data management, analysis, and reporting. It allows multiple users to work on the same project, share code, and collaborate seamlessly.

The outcomes of such disrupting features are:

- Time and Cost Savings: Partial automation, supported by human oversight, of the clinical data pipeline saves time and reduces costs by eliminating manual coding and repetitive tasks. Researchers and data analysts can focus on higher-value activities such as interpreting results and deriving insights.
- Improved Accuracy: using a generative AI approach, the risk of human errors in data cleaning, analysis, and reporting could be reduced. Automating these processes ensures that the results are accurate and reliable.
- Standardization and Consistency: promoting standardization in data management and analysis, ensuring consistent practices across different research projects or healthcare organizations. This improves the quality and comparability of clinical data.
- Enhanced Productivity: handling repetitive and time-consuming tasks, researchers and data analysts can work more efficiently and handle larger volumes of clinical data. This leads to increased productivity and faster turnaround times.
- Advanced Analytics: leveraging machine learning algorithms and advanced analytics techniques to derive meaningful insights from clinical data. Researchers can explore complex analyses and uncover valuable patterns and trends.

CHALLENGES AND SOLUTIONS IN CLINICAL DATA AUTOMATION

- **Data Quality and Standardization**: Clinical data often originates from disparate sources and formats, complicating standardization efforts. This process can be streamlined to enhance data normalization across multi-source datasets.
- **Data Security and Privacy**: The sanctity of patient data is paramount. Robust security protocols, HIPAA compliance, data encryption, and stringent access controls are essential components of a secure data environment.
- **Interoperability**: The lack of seamless data exchange between healthcare systems poses a challenge. Adopting standardized data formats like HL7 and implementing Health Information Exchange (HIE) systems can mitigate this issue.
- **Model Interpretability**: The complexity of AI models in healthcare necessitates transparency. Solutions include explainable AI, transparent prediction mechanisms, and feature importance analysis. Ethical AI principles must be integral to the design of Clinical Data Automation platforms.

ETHICAL CONSIDERATIONS ABOUT GENERATIVE AI IN LIFE SCIENCES

- **Regulatory Compliance**: Most AI-driven solutions in life sciences must undergo rigorous validation processes and comply with the relevant regulatory frameworks. This includes strict adherence to good clinical practices, obtaining the necessary approvals from regulatory bodies, and conducting robust validation studies to confirm the safety, efficacy, and reliability of AI applications.
- **Ethical Guidelines**: In the realm of life sciences, the application of AI technologies must strictly adhere to ethical principles as aligned with principles of bioethics. This encompasses benevolence, respecting patient autonomy, ensuring that informed consent is obtained, and safeguarding the privacy and confidentiality of sensitive health data. Developers and researchers should place the well-being and rights of individuals at the forefront when involved in AI-driven research or healthcare applications.
- **Data Security**: The nature of life sciences involves dealing with highly sensitive and personal health data. As such, responsible AI practices necessitate robust data privacy and security measures. Organizations must not only comply with relevant data protection laws and regulations but also implement advanced encryption techniques and secure storage methods. Furthermore, protocols should be established and updated regularly to protect patient data from unauthorized access or security breaches.
- **Bias Mitigation**: The algorithms that power AI in life sciences should be meticulously designed and continuously trained to minimize biases. Unexpected biases in data collection or algorithmic decision-making can result in unequal healthcare access, misdiagnoses, and treatment disparities. To counteract this, developers should curate datasets that are diverse and representative, and they should regularly evaluate the algorithmic performance to identify and rectify any biases.
- **Transparency and Explainability**: The AI models employed in life sciences must be both transparent and explainable. Understanding the rationale behind an AI's decision or recommendation is crucial, especially in life-or-death healthcare scenarios. Techniques such as interpretable machine learning models or rule-based systems should be utilized to provide deeper insights into the AI's decision-making process.
- **Human Oversight and Collaboration**: AI technologies should be developed with the intent to augment human expertise, rather than to replace it. Human oversight is essential for the responsible use of AI in life sciences. Healthcare professionals must be actively involved in the development, deployment, and interpretation of AI systems. Their expertise is invaluable for making informed decisions that consider the unique healthcare needs of individual patients.
- **Continual Monitoring and Improvement**: The field of AI is ever evolving, and as such, trustworthy AI practices in life sciences require continuous monitoring and improvement. Regular evaluations of AI systems are imperative to identify and address potential biases, errors, or unintended consequences. Frequent audits and updates to the AI models can ensure their ongoing accuracy, reliability, and alignment with evolving ethical, legal, and regulatory standards (concept of Machine Learning Operations: MLOps).

CASE OF EXPLAINABILITY OF LLM

Large Language Models (LLMs), characterized by their billions of parameters, have become a focal point in the artificial intelligence landscape. These models, such as GPT-x, are highly capable of generating human-like text and performing a myriad of language tasks. However, their complexity and extensive parameter space make them a black box, posing challenges to explainability.

Explainability in AI is a cornerstone for ensuring transparency, accountability, and ethical deployment. It becomes even more critical when these technologies are applied in sensitive areas like healthcare and life sciences. Here are some nuanced approaches to enhance the explainability of such complex models:

- Interpretability at the Architecture Level: Techniques like attention visualization and saliency maps can offer insights into the model's internal workings. These methods help pinpoint which parts of the input text have the most influence on the output, thereby shedding light on the model's reasoning process.
- Fine-Tuning and Domain-Specific Training: While LLMs are generally trained on broad datasets, fine-tuning them on domain-specific data can make their outputs more relevant and easier to interpret within that particular field. This focused approach simplifies the task of explaining the model's decisions.
- Rule-Based Post-Processing: Implementing post-processing rules can help align the model's outputs with ethical considerations, legal requirements, or specific project objectives. This makes the model's decisions more explainable and ensures they adhere to predetermined criteria.
- Addressing Biases and Ensuring Fairness: LLMs can inadvertently adopt biases present in their training data. Techniques like debiasing algorithms and regularization can help mitigate these biases, ensuring that the model's outputs are fair and do not perpetuate harmful stereotypes.
- Human-AI Collaboration: For complex or critical tasks, human experts can be brought into the decision-making loop. This hybrid approach combines human reasoning with machine efficiency, leading to outputs that are both comprehensive and easier to understand.

In the context of life sciences, a copilot for scientific programmers and data science engineers could serve as a prime example of responsible LLM application. It facilitates Human-AI collaboration, particularly in the validation of drug candidates. While the LLM assists in automating the clinical data pipeline and offers best-practice statistical code, the ultimate validation relies on the biostatistician's expertise and their effective use of the SAS Viya platform. The LLM essentially lowers the entry barrier, making the platform more accessible.

By incorporating these trustworthy AI practices, stakeholders in life sciences can unlock the transformative potential of AI technologies. This ensures not only advancements in medical research and patient care but also upholds the ethical standards, privacy protocols, and fairness that are imperative in healthcare application.

FUTURE PROSPECTS

The field of generative AI in life sciences is continuously evolving, driven by technological advancements, regulatory landscapes, and ethical considerations. In this section, we will explore some of the future prospects and factors that could shape the development and application of generative AI in the life sciences.

TECHNOLOGICAL ADVANCEMENTS:

Improved Deep Learning Models: Deep learning models like GANs and VAEs are advancing generative AI. Future gains may come from self-supervised and unsupervised learning, enhancing performance and efficiency. Meta-learning could enable real-time adaptation, amplifying their impact in life sciences.

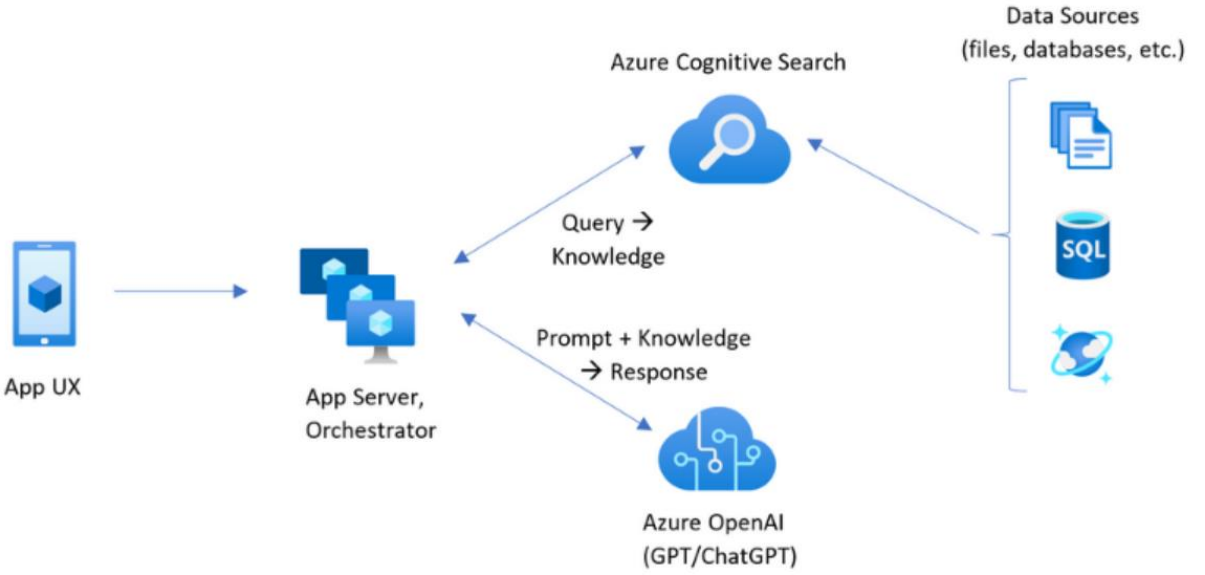
Multi-Modal Generative Models: Current generative AI models primarily focus on generating text or images separately. However, future advancements may lead to the development of multi-modal generative models capable of generating both text and images simultaneously. This could enable more comprehensive and accurate AI-generated outputs, enhancing applications in areas such as drug discovery, protein folding, and medical imaging analysis.

Integration of Domain Knowledge: In the evolving landscape of generative AI, the integration of domain-specific knowledge is becoming increasingly critical for enhancing the accuracy and relevance of generated outputs, particularly in the life sciences sector. While Fine-Tuning techniques have been the go-to approach, the advent of Retrieval Augmented Generation (RAG) is setting a new standard in the field. RAG ingeniously combines the strengths of retrieval models like BERT with those of generative models such as GPT-x or Transformers. This hybrid approach allows the system to pull from extensive knowledge bases, thereby significantly improving the quality and contextual relevance of its outputs.

The primary innovation behind RAG lies in its ability to overcome the limitations commonly associated with traditional generative models—limitations such as factual inaccuracies or the inability to effectively incorporate external knowledge. By adding a retrieval component to the generative process, RAG can access and utilize a wealth of

structured or unstructured data. This feature makes RAG exceptionally useful for tasks that demand accurate and contextually relevant responses, including but not limited to, question-answering systems, dialogue systems, and specialized content generation.

The RAG (Retrieval Augmented Generation: https://learn.microsoft.com/en-us/azure/search/retrieval-augmented-generation-overview) methodology represents a significant advancement in the field of generative AI, offering a more nuanced and informed approach to generating content that is both accurate and relevant, especially in data-intensive fields like life sciences.



Source: [RAG and generative AI - Azure Cognitive Search | Microsoft Learn]

A high-level summary of the pattern looks like this:

- *Start with a user question or request (prompt).*
- *Send it to Cognitive Search to find relevant information.*
- *Send the top ranked search results to the LLM.*
- *Use the natural language understanding and reasoning capabilities of the LLM to generate a response to the initial prompt.*

Cognitive search provides inputs to the LLM prompt but doesn't train the model. In RAG architecture, there's no extra training. The LLM is pretrained using public data, but it generates responses that are augmented by information from the retriever.

Quantum Computing: The advent of quantum computing could revolutionize the computational capabilities of generative models, allowing for more complex simulations and analyses especially when it comes to molecular design and simulation of the complex binding relationship between a ligand and its receptor with 3D models of either targets (cellular receptors) or molecular “keys” (proteins, antibodies, …).

Edge AI: The integration of AI algorithms directly into medical devices and sensors could enable real-time data analysis and decision-making at the bed of the patient to improve real-time patient’s clinical outcomes.

CONCLUSION

The integration of generative AI into the life sciences is a transformative development, offering unprecedented opportunities for innovation and efficiency. From drug discovery to personalized medicine, generative AI is reshaping traditional paradigms and accelerating the pace of research and treatment.

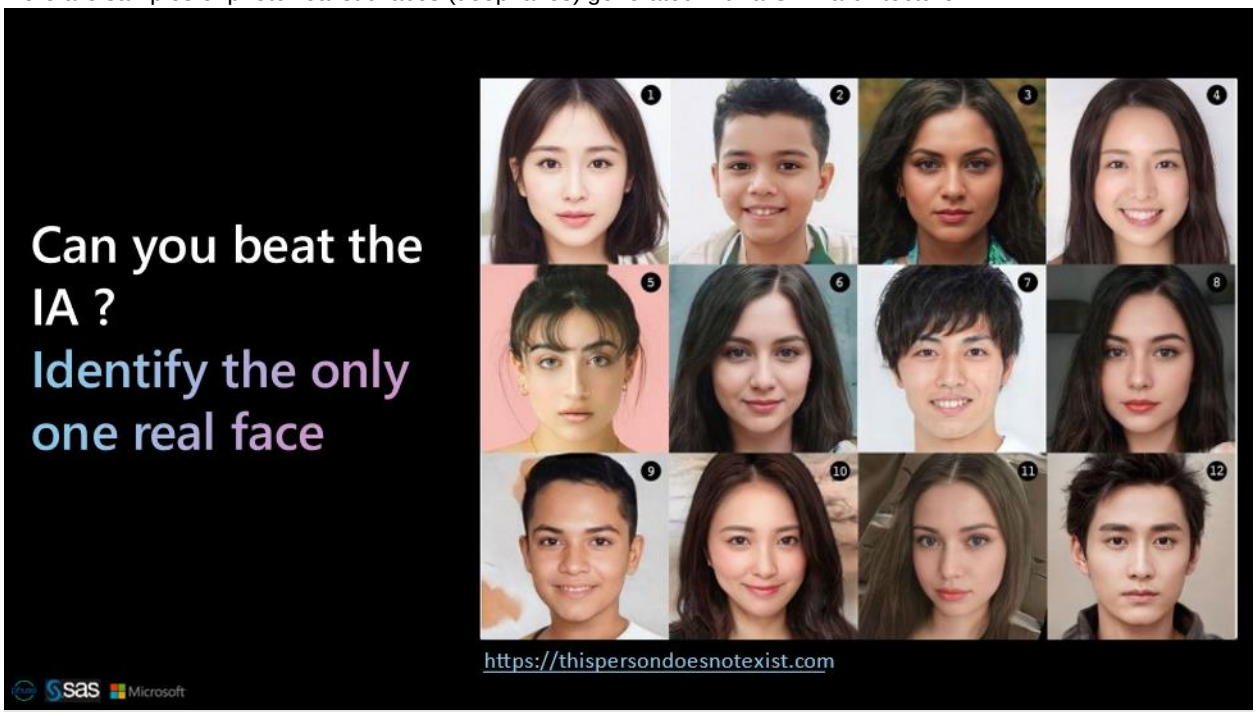
However, this technological leap is not without its challenges. Ethical considerations around data privacy and model interpretability, as well as regulatory constraints, necessitate a cautious and well-thought-out approach. The role of professional solutions like SAS Viya and a co-pilot like functionality with interactions to Microsoft's Azure OpenAI Service in providing standardized and reliable statistical analyses exemplifies the kind of interdisciplinary collaboration that will be essential for the responsible advancement of this field.

Looking ahead, the future is promising but uncertain. Technological advancements such as quantum computing and edge AI have the potential to further amplify the capabilities of generative AI. Yet, these advancements will also bring forth new ethical and regulatory challenges that will require international cooperation and governance.

As we stand on the cusp of this new era, it is imperative for stakeholders across academia, industry, and regulatory bodies to collaborate in defining the ethical and operational frameworks that will guide the future of generative AI in life sciences. Only through such collaborative efforts can we harness the full potential of this technology, while safeguarding the principles of ethical and equitable healthcare.

COULD YOU BEAT GENERATIVE AI?

Here are samples of photo-realistic faces (deep fakes) generated with a GAN architecture.



Remember these are “sophisticated stochastic parrots” which are as good as the SAS Viya platform when it comes to doing statistics.
GAN can do “Data Augmentation” because they can generate outputs that are statistically like their training dataset.
So how do you beat the Discriminator of this GAN who was not able to discriminate the outputs coming from the Generator, from “real portraits”?
You **must focus on details**, on something that is somehow different from the vast distribution of features you can see in the output.
Every portrait has “hair”, “eyes”, “mouth”, “teeth” … but only one portrait has earrings. We can guess that this specific “feature” was not statistically present in the learning dataset, hence the GAN did not make the inference that a portrait should have earrings.

ACKNOWLEDGEMENTS
The authors would like to thank SAS colleague Olivier Bouchard for his input and guidance.

CONTACT INFORMATION
Your comments and questions are valued and encouraged. Contact the authors at:

Jérôme Vétillard
Microsoft France
39, Quai du President Roosevelt
92130 Issy Les Moulineaux
France
Work Phone: +33 1 57 75 25 10
Email: jerome.vetillard@microsoft.com
Web: https://www.microsoft.com/en-us/industry/health/pharmaceuticals

Mark Lambrecht
SAS
Hertenbergstraat 6
3080 Tervuren
Belgium
Work Phone: +32 475 96 06 58
Email: mark.lambrecht@sas.com
Web: https://www.sas.com/en_us/industry/life-sciences.html