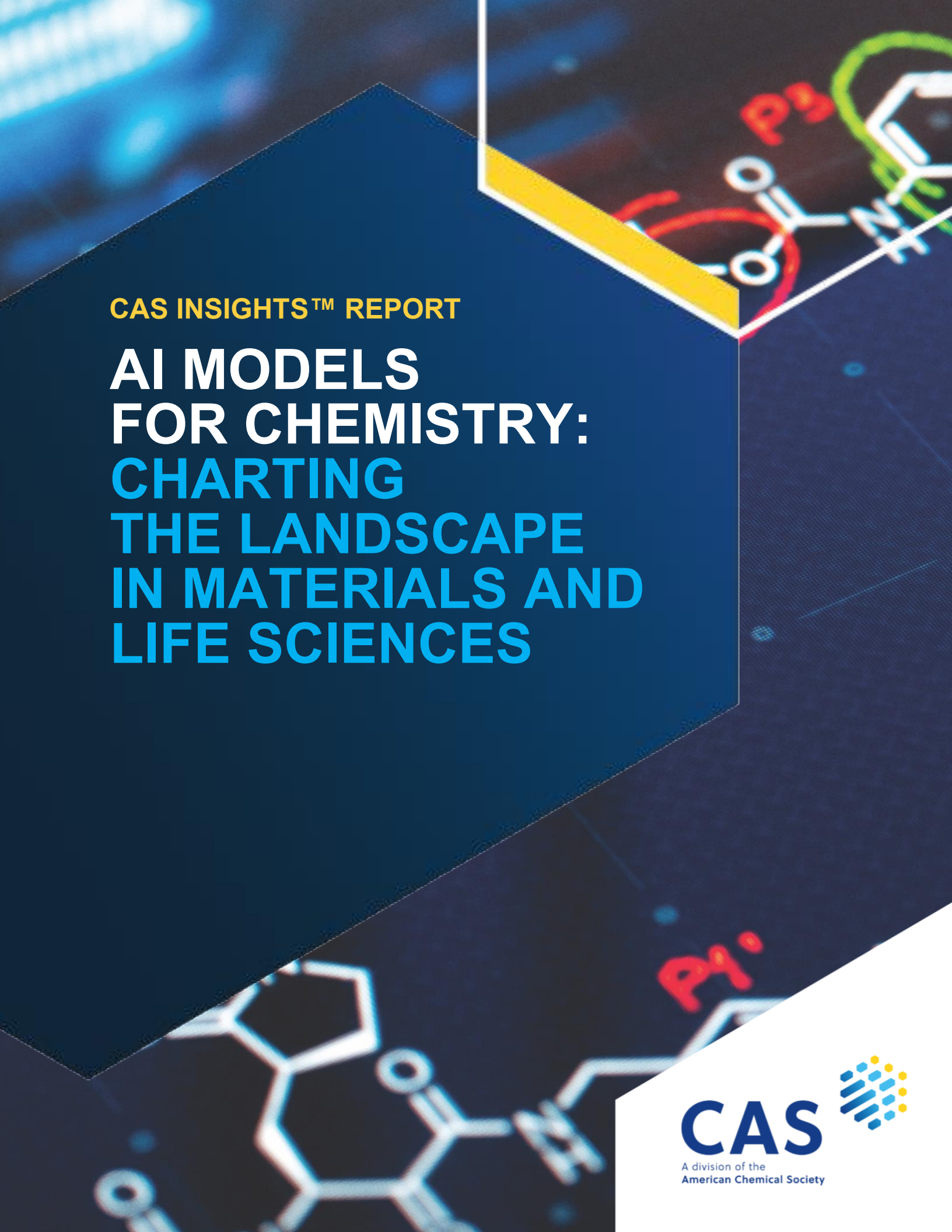


CAS INSIGHTS™ REPORT

AI MODELS FOR CHEMISTRY: CHARTING THE LANDSCAPE IN MATERIALS AND LIFE SCIENCES

CAS
A division of the
American Chemical Society

HOW AI MODELS ARE TRANSFORMING SCIENCE

In the landscape of modern scientific research, artificial intelligence (AI) has emerged as a transformative force, fundamentally altering how researchers conceptualize, conduct, and validate their work. The convergence of computational power, advanced algorithms, and vast datasets has driven AI from theoretical possibility to practical necessity across scientific disciplines.

This report of over 310,000 journal articles and patents can inform our views about the rise of AI for science and the current state and trajectory of AI integration in scientific research for researchers, institutions, and policymakers.

Table of Contents

Executive summary

In the landscape of modern scientific research, artificial intelligence (AI) has emerged as a transformative force, fundamentally altering how researchers conceptualize, conduct, and validate their work. The convergence of computational power, advanced algorithms, and vast datasets has driven AI from theoretical possibility to practical necessity across scientific disciplines.

This revolution addresses several longstanding challenges that have limited scientific progress: the overwhelming volume of data generated in fields like genomics, medicine, and materials science; the extensive time and costs of processes such as drug discovery; and cognitive limitations in hypothesis generation. AI offers new and promising pathways to overcome these challenges and find faster breakthroughs across many scientific disciplines.

For example, in biomedical sciences, AI has revolutionized protein structure prediction, accelerated the drug discovery process, and enabled personalized medicine. Similarly, in material sciences, AI is accelerating the discovery of novel materials. These advancements have paved the way for self-driving laboratories and inverse design frameworks, which are changing the scientific method itself. AI has also significantly advanced process optimization by enabling real-time experimental adjustment to improve yield and efficiency while reducing waste and cost.

We performed a quantitative analysis of the CAS Content Collection™, the largest human-curated repository of scientific information, to understand the implementation and impact of these technological advancements on scientific discovery. We systematically analyzed over 310,000 journal articles and patents from the CAS Content Collection spanning the years 2015-2025, employing advanced analytical techniques to identify AI-related publications and their associated methodologies, institutions, and research domains.

Our study explores AI method distribution across scientific fields with deep-dive analyses of two critical domains: biomedical sciences and materials science. These analyses detail dominant concepts, co-occurrence patterns of AI methods, and notable substance classes. Additionally, we investigate AI's role in optimizing industrial processes and emerging applications.

These findings provide essential insights into the current state and trajectory of AI integration in scientific research, offering valuable guidance for researchers, institutions, and policymakers in understanding technological adoption patterns, identifying collaboration opportunities, and making informed strategic decisions about future AI investments and research directions.

The landscape of AI models in science

To understand the context in which certain models are selected and their impact, we conducted a comprehensive analysis of scientific publications, including journals and patent families, that have incorporated AI-based methods. The visualization of the entire landscape highlights the increasing integration of these technologies across scientific disciplines as more models and functionalities are developed (see Figure 1).

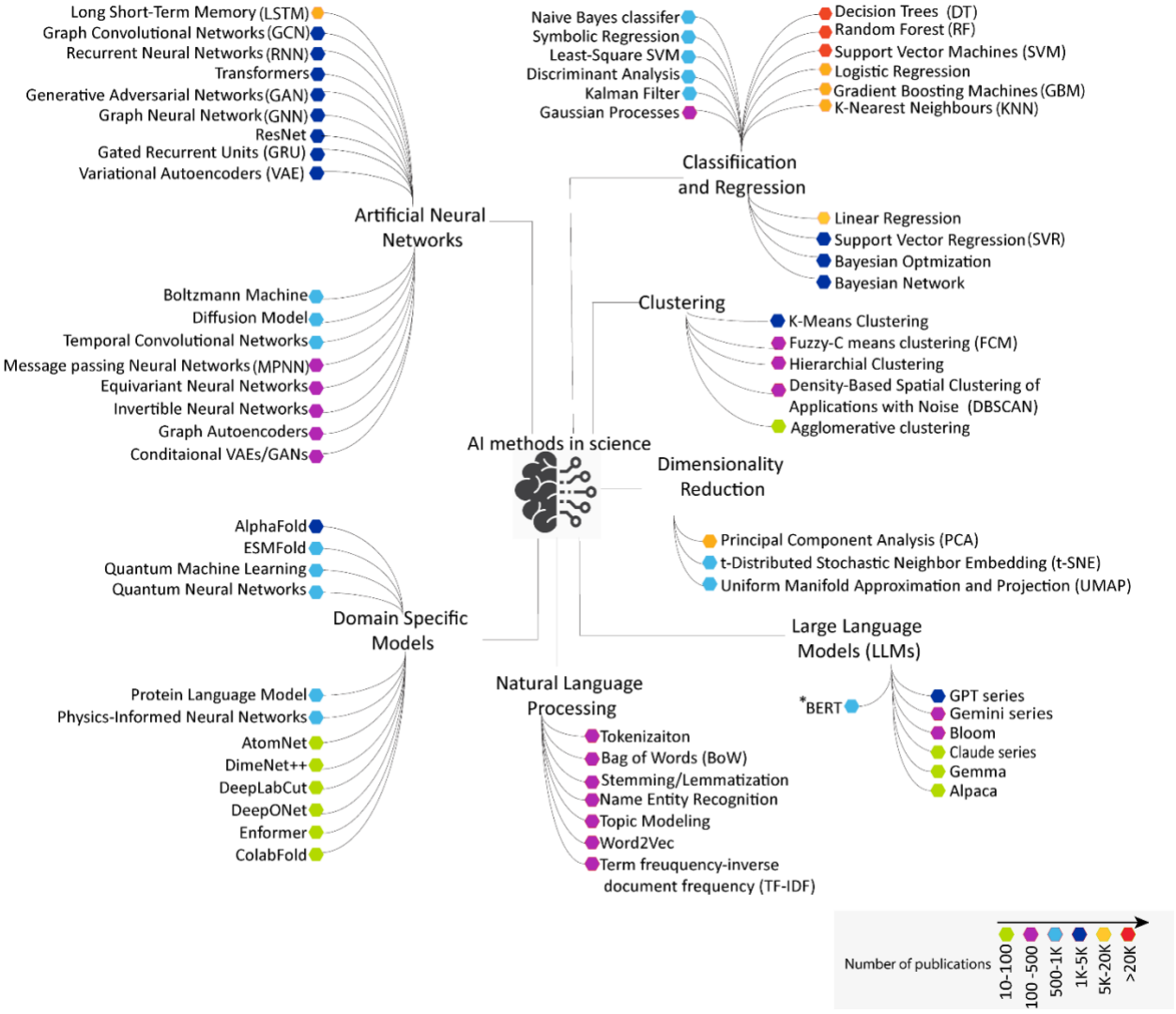


Figure 1: *Landscape of AI methods applied across scientific disciplines. Data includes journal and patent publications for the period 2015-2025, with 2025 data encompassing January to March. Source: CAS Content Collection.*

The main branches of this visualization include:

Classification, regression, and clustering models

Classification models are designed to predict discrete labels or categories and are especially effective in handling high-dimensional data and providing interpretable results. Common models include Decision Trees (DT), a model which classifies outcomes by recursively splitting data into branches. Random Forest (RF) is an ensemble of decision trees that reduces overfitting by averaging predictions. Another common model, Support Vector Machine (SVM), finds the optimal boundary between classes, while K-Nearest Neighbors (KNN) classifies the majority class of nearest neighbors.

These models are widely applied to labeled datasets (supervised learning) with categorical outcomes, such as spectroscopy, microscopy, or omics data with known classes. Example applications include classifying disease types from gene expression or imaging data, predicting material classes, and classifying compounds based on toxicity or reactivity.

Regression models predict continuous numerical values, making them ideal for forecasting and optimization. Some of the regression models observed in publications are Linear Regression, Logistic Regression, Support Vector Regression (SVR), and Symbolic Regression. These models are suited for numerical data from experiments, simulations, or time-series data from sensors or instruments. Applications include predicting energy levels, decay rates or particle trajectories, estimation of reaction yields, molecular properties, modeling temperature, precipitation, predicting conductivity, hardness, and band gaps.

Unlike classification and regression, clustering models uncover natural groupings in unlabeled data, revealing hidden structures and supporting exploratory analysis (unsupervised learning). These models are effective for high-dimensional, unlabeled datasets such as imaging, spectral, and molecular data. Suitable scientific data for this group includes but is not limited to grouping genes or samples based on expression profiles and discovering new material families from structural data.

Artificial neural networks (ANNs)

ANNs are a class of machine learning (ML) models designed to learn complex patterns through interconnected layers of artificial neurons. They are powerful for modeling intricate, non-linear relationships in data. As a result, they form the basis for deep learning and are the foundation of modern AI with specialized architectures such as Recurrent Neural Networks (RNNs) for sequential data, Long-Short Term Memory (LSTM), an enhanced RNN that better captures long-range dependencies, and Gated Recurrent Units (GRUs) which is a simplified version of LSTMs with fewer parameters.

These models are suitable for sequence and time series data such as gene expression, protein sequences, physiological signals, climate data, particle physics, and sensor network data. They often enhance image-text alignment by leveraging knowledge maps — structured representations of domain-specific concepts and their relationships — which helps to bridge the semantic gap between visual data (like medical images) and textual descriptions (such as clinical notes or diagnostic reports). Their adoption has surged in areas like drug discovery and development, precision medicine, medical imaging, material discovery and design, energy storage, manufacturing, and quality control.

Hybridized methods

The comparison between ANNs and conventional ML models, such as classification, regression, and clustering, is not simply about which is better but rather about understanding their complementary roles. ANNs dominate where the data is complex and unstructured, where there are large training datasets, and when representation learning is required. Conversely, conventional ML remains strong where the data is

small, research problems require strict interpretability, and the inputs possess well-understood statistical relationships and require uncertainty quantification.

Domain-specific models

Although domain-specific models have fewer publications, they include groundbreaking work such as AlphaFold, a deep learning approach that achieved near-experimental accuracy in protein structure prediction. ESMFold is another interesting model that directly leverages protein language modeling to generate high-quality structure predictions directly from single protein sequences.

Natural language processing (NLP)

NLP is used to analyze, understand, interpret, and generate human language. It involves tasks like tokenization (breaking down text into smaller units called tokens, typically word or sub-word) that can be processed by algorithms. A common method for representing text is the Bag of Words (BoW) model, which converts text into numerical vectors based on word frequency, ignoring grammar and word order. Another key technique is Named Entity Recognition (NER), which identifies and classifies entities such as names of people, organizations, and locations within the text.

There are extensive applications of NLP models in biomedical text mining and knowledge extraction, ranging from specialized language models pre-trained on biomedical literature (i.e., BioBERT, BioGPT) and electronic health record (EHR) analysis to the automated extraction of disease characteristics from clinical records and creating novel drug candidates using language models. In material science and chemistry, NLP has been used for synthesis protocol extraction, material property prediction, knowledge base construction, and inverse design.

Large language models (LLMs)

This transformative class of AI systems has revolutionized NLP and found extensive applications across scientific domains. LLMs are neural network-based systems that are trained on vast amounts of text data to learn statistical patterns of language. They are used in information extraction, summarization, knowledge graph construction, cross-domain integration, and generative applications.

Widely adopted LLMs include Generative Pre-Trained Transformer (GPT), which drives ChatGPT and GPT-3.5 and GPT-4. Bidirectional Encoder Representations from Transformers (BERT) remains a top language model, maintaining strong academic interest due to its bidirectional context understanding and superior performance in question answering and sentiment analysis.

BLOOM, a multilingual model, has emerged as a significant contributor to the research landscape. New models such as GEMINI developed by Google DeepMind and LLAMA from Meta AI are also becoming popular, as is Claude, developed by Anthropic. The latest generation of models Gemma, Falcon, Mistral, Qwen, and DeepSeek, launched between 2023 and 2025, show great promise for future developments.

Several specialized LLMs for chemistry, life sciences, and material science are also making an impact, such as chemLLM, PharmaGPT, and MatSciBERT.

Publication trends show AI's impact on science

As noted, we analyzed 310,000 documents in the CAS Content Collection related to AI in scientific research over the period 2015-2025. We noted steady growth over the entire period with prominent growth in the last five (see Figure 2). The rising number of patent families further suggests that AI is evolving from an emerging technology to an essential research tool across various industries.

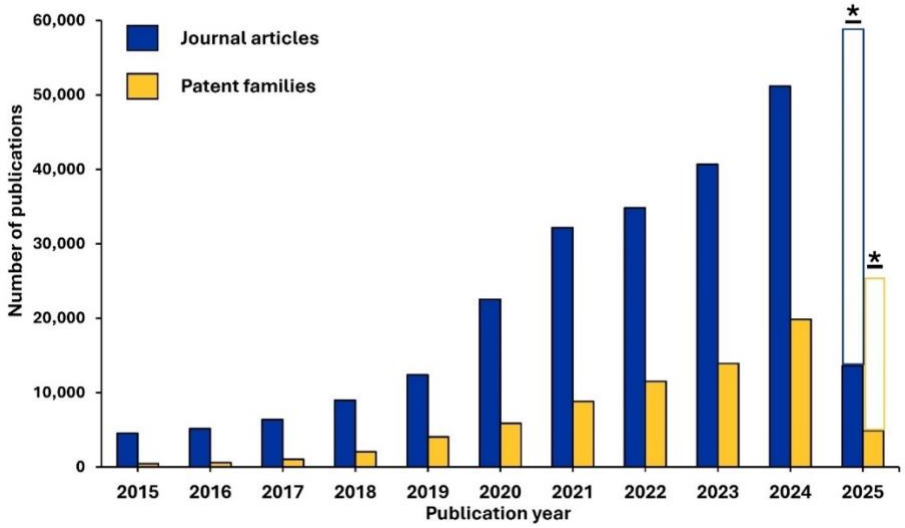


Figure 2: *Yearly trends for journal articles and patent families. Data includes publications for the period 2015-2025, with 2025 data encompassing January to March. Source: CAS Content Collection.*

- **Distribution by country/region and organization:** China leads in publication volume (journals and patents), while the U.S., India, South Korea, and Japan also show steady growth in publications (see Figure 3). In terms of patents, China and India together contribute to more than 75% of the global patents in this field — the top 15 institutions in terms of patent filings are from China and India. Chinese universities dominate in terms of publication volume, with the average citation count for most of them ranging between 10 and 20. In contrast, MIT and Stanford boast significantly higher average citation counts of 32 and 66, respectively. This suggests that although China excels in quantity, U.S. universities are leading in terms of the quality and impact of their journal articles in the field of AI.

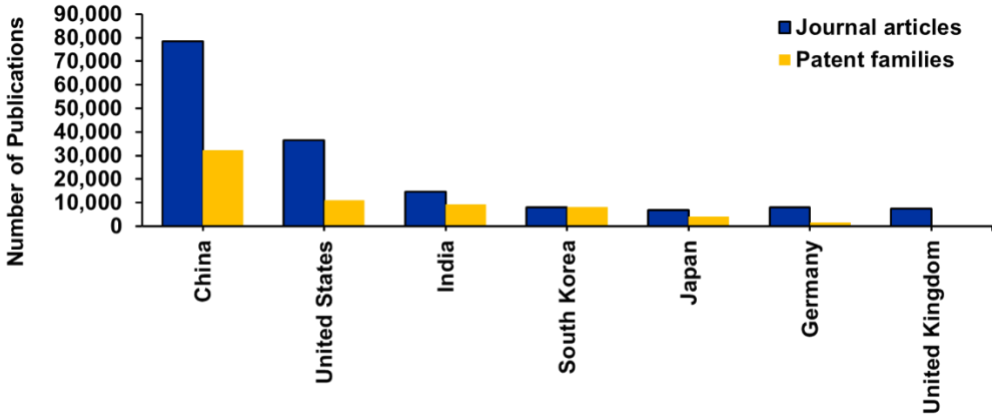


Figure 3: *Leading countries/regions by volume of AI-related publications.*

- **Distribution in particular publications:** We analyzed publication volume and citation impact of leading scientific journals publishing AI-based research. Publication volume is dominated by *Scientific Reports, Applied Sciences*, and *PLoS One*. However, when considering the average citation count per article, *Proceedings of the National Academy of Sciences (PNAS)* stand out, demonstrating exceptional influence with nearly 50 citations per publication, despite having relatively modest publication volume. Other journals such as the *Journal of Chemical Information and Modelling (JCIM)*, *Bioinformatics,* and *Nature Communications* also show remarkable average citation, significantly outperforming the journals with huge volume. This pattern suggests that while broad-scope journals like *Scientific Reports* publish the most AI-related research, specialized or prestigious journals like *PNAS, JCIM*, and *Nature Communications* publish more influential work that generates greater scientific impact.

- **Trends by scientific field:** Numerous fields have been greatly impacted by AI, but several stand out for exponential growth in publications (see Figure 4):

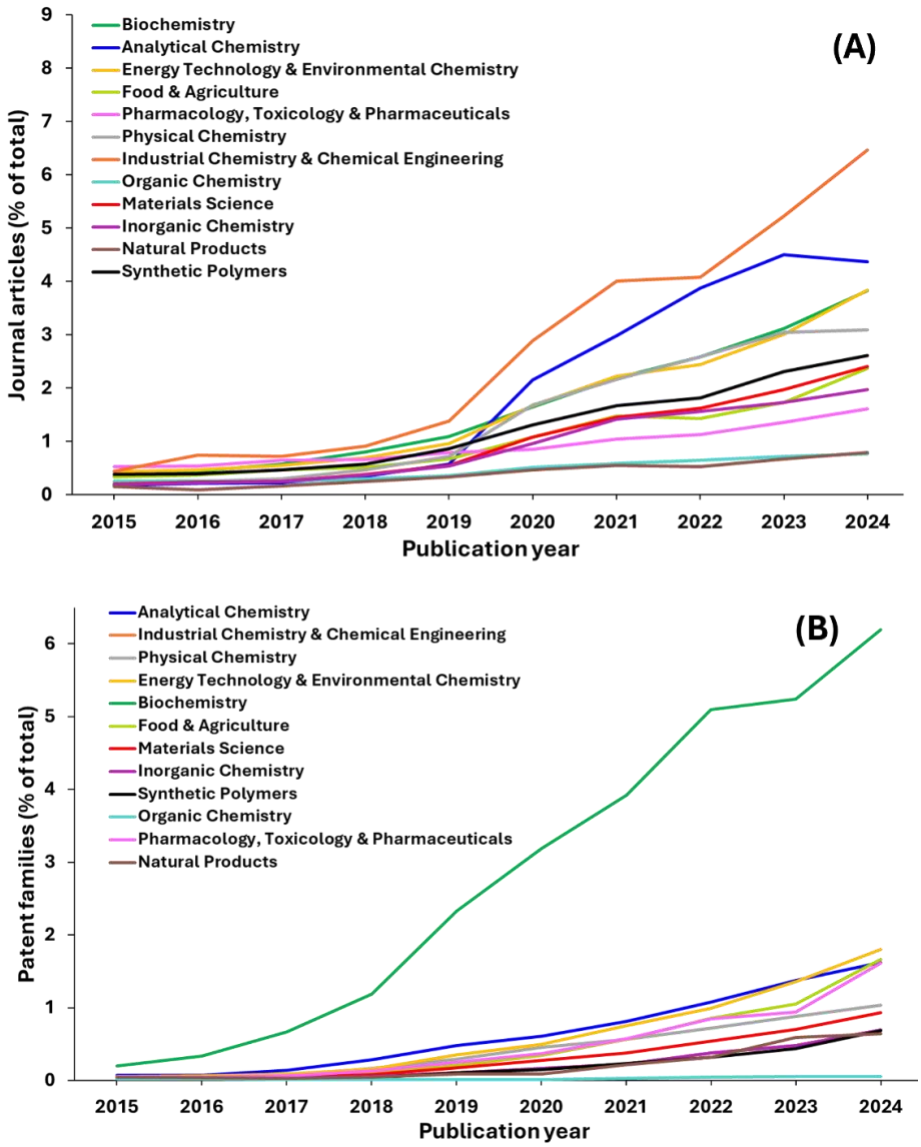


Figure 4: *Publication trends per discipline between 2015-2024 for (A) journal articles and (B) patent families. Source: CAS Content Collection.*

Among all disciplines, Industrial Chemistry & Chemical Engineering demonstrates the most dramatic growth in journal publications, with its trajectory reaching approximately 8% of the total documents by 2024. This exceptional growth likely reflects the field's broad applications across manufacturing, process optimization, and industrial innovation, as well as its critical role in developing sustainable industrial processes and green chemistry solutions.

Analytical Chemistry emerges as the second-fastest growing field in terms of journal publications, with the deep blue line showing robust growth throughout the period from 2019 onward. This strong growth pattern shows the field's fundamental importance across all areas of chemistry and its role in developing new measurement techniques, instrumentation, and analytical methods that support research across multiple disciplines. Energy Technology & Environmental Chemistry demonstrates solid growth, jointly ranking third among the fastest-growing fields along with Biochemistry. This reflects urgent global focus on climate change, environmental sustainability challenges, and emerging events.

The remaining disciplines, including Physical Chemistry, Pharmacology, Toxicology and Pharmaceuticals, Organic Chemistry, Inorganic Chemistry, Natural Products, and Synthetic Polymers, all demonstrate modest but consistent increases in publication volume. These fields represent mature areas of scientific research where fundamental principles are well-established, yet continued investigation yields incremental advances and refinements.

The patent family data presents a strikingly different landscape compared to journal publications, with highly varied and concentrated innovation patterns rather than the uniform growth seen in academic output. This contrast highlights the fundamental difference between academic research productivity and commercially viable innovation, where market forces, practical applications, and economic incentives play decisive roles in determining which scientific advances translate into patentable intellectual property.

Among patents, Biochemistry shows exponential-like growth, reaching approximately 8% by 2024 and completely overshadowing all other disciplines. This dramatic patent activity reflects the intense commercial interest and investment in life science research in general and further boosted in response to the COVID pandemic, where commercial research was given significant funding. The biochemistry patent landscape is dominated by biomedical methods and shaped by innovations that merge molecular science with advanced healthcare technologies. Key domains include disease detection, medical imaging, wearable monitoring, and clinical decision support, all leveraging biochemical markers. Biochemistry also drives progress in neurological interfaces, genomics, biomarker discovery, signal processing, surgical guidance, and biomedical manufacturing, highlighting its central role in modern biomedical innovation.

Based on our analysis, we found the majority of the AI-related publications are related to biomedical research and materials science. Hence, we conducted an in-depth analysis for these key research areas as these fields dominate the dataset, exhibit rapid AI adoption with notable annual growth rates, and yield higher citation impacts, indicating their scientific relevance.

Applications of AI models in biomedical research

Overview

Biomedical research has faced ongoing challenges in complexity and costs that AI can now address. For example, decoding complex protein structures and interactions, the high costs and failure rates associated with drug development, and the intricacies of understanding multifactorial disease mechanisms all lend themselves to AI-based approaches.

The rationale behind this analysis was to ensure a representative sampling of the most significant research areas by using document frequency as an objective metric to identify key concepts. Specifically, we identified concepts that have generated substantial scientific interest and research investment using several AI/ML methods within the biomedical field (see Figure 5).

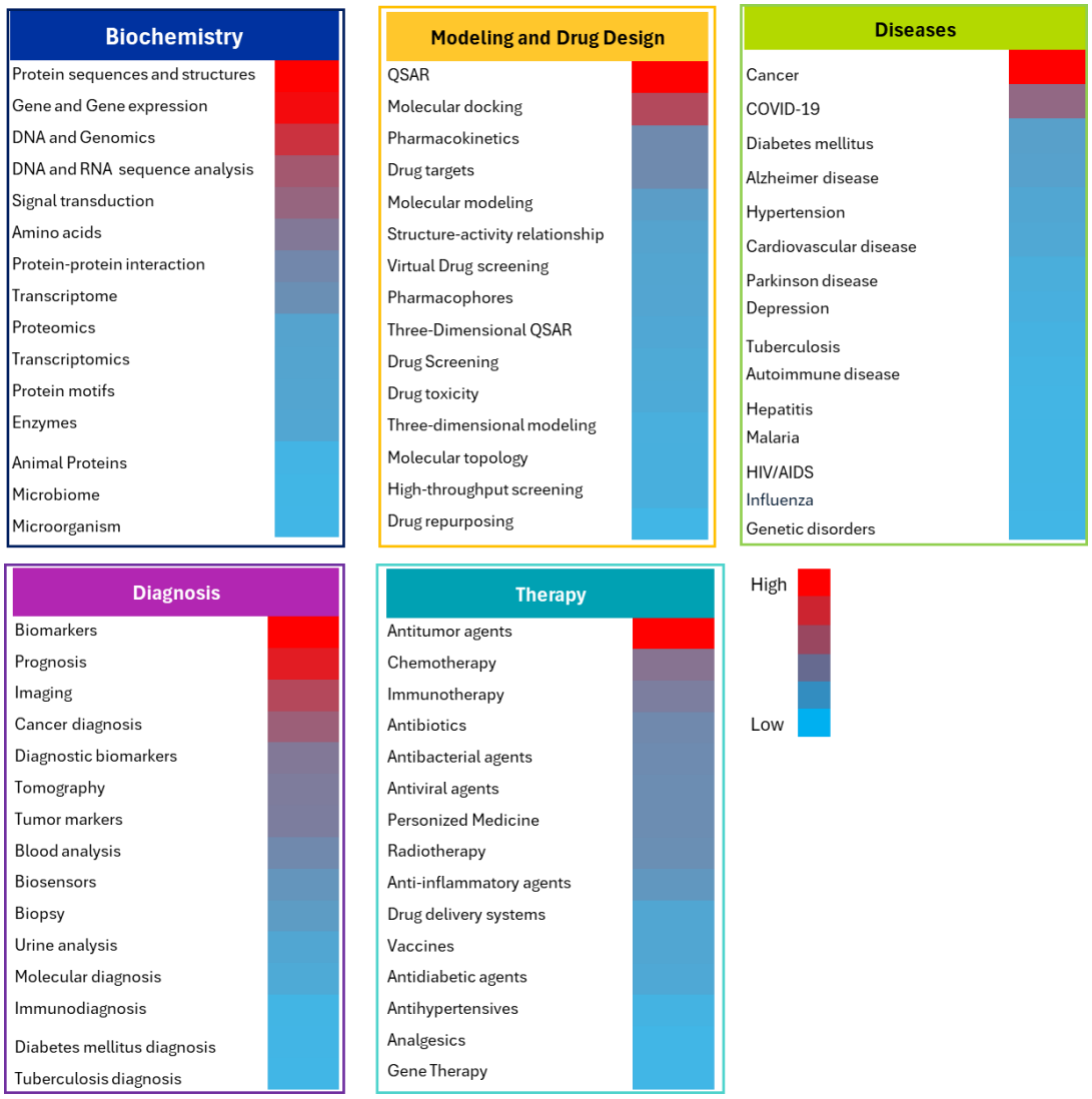


Figure 5: *Conceptual landscape of AI-driven Biomedical research publications from 2015-2025, organized into five distinct domains: Biochemistry, Modeling and Drug Design, Diagnosis, Diseases, and Therapy. Each domain encompasses 15 representative concepts, employing a color coding from red (highest) to blue (lowest), indicating document frequency. Source: CAS Content Collection.*

In the **Biochemistry domain**, protein sequences and structures, gene and gene expression and DNA/genomics have emerged as the most extensively studied concepts, utilizing diverse machine learning approaches. ML models thrive in these areas due to data abundance and standardization, especially in complex systems like protein hierarchies and gene regulatory networks. ML excels in tasks such as sequence-function mapping, protein family classification, domain prediction, structure-function analysis, tumor feature selection, gene interaction prediction in cancer, and clustering genes by expression in patterns in oncology, leveraging high-dimensional expression data with thousands of genes measured

simultaneously, complex non-linear relationships between genes and phenotypes, and hierarchical organization of genetic regulatory networks.

Advanced ML models like AlphaFold and BERT are used for protein structure prediction by learning from amino acid sequences and evolutionary patterns to predict 3D folding and functional domains. These models effectively handle complex spatial relationships and sequential inputs with long-range dependencies.

In the **Modeling and Drug Design domain**, QSAR modeling dominates this field. It utilizes ML methods to extract molecular descriptors (i.e., physicochemical properties, fingerprints) from chemical structures for predicting biological activities and toxicity endpoints through supervised learning approaches. These methods excel with feature rich molecular descriptors, structure-activity relationships, and quantitative endpoints suitable for regression tasks.

Molecular docking, the second most frequent concept, uses deep learning approaches (CNN and GNN) on 3D structural data and molecular graphs to predict protein-ligand binding poses and affinities, leveraging spatial conformations, energy landscapes, and molecular interactions. Pharmacokinetics follows as the next focus area, using ML to learn from molecular descriptors, physiological parameters, and time-series data for predicting ADMET properties (Absorption, Distribution, Metabolism, Excretion, Toxicity) and drug clearance rates.

Biomarkers dominate the **Diagnostic domain** where ML methods analyze high-dimensional genomics, proteomics, and metabolomics data to identify discriminative molecular signatures distinguishing disease states from healthy controls through feature selection and classification. AI methods are also used in prognosis research to analyze patient clinical data and treatment history for predicting survival times and disease progression through survival analysis techniques.

Medical imaging contains data from X-rays, CT, MRI, and structured spatial information that is ideal for deep learning methods. CNNs are widely used to analyze radiological images for tumor detection, fracture identification, disease classification, and pathological conditions through automated pattern recognition. Imaging can also benefit from the ability of deep learning to extract multi-scale features and perform segmentation tasks, particularly in cancer diagnosis where it aids in tumor detection, classification, malignancy assessment, and risk prediction.

In the **Disease domain**, cancer research received the highest attention, with various ML models applied based on the data types. Genomic and transcriptomic data are commonly analyzed using RF and SVM models for cancer subtype classification and treatment prediction. These tasks benefit from integrating multi-omics data, which captures diverse molecular layers, and from modeling the complexity of heterogeneous tumor microenvironments.

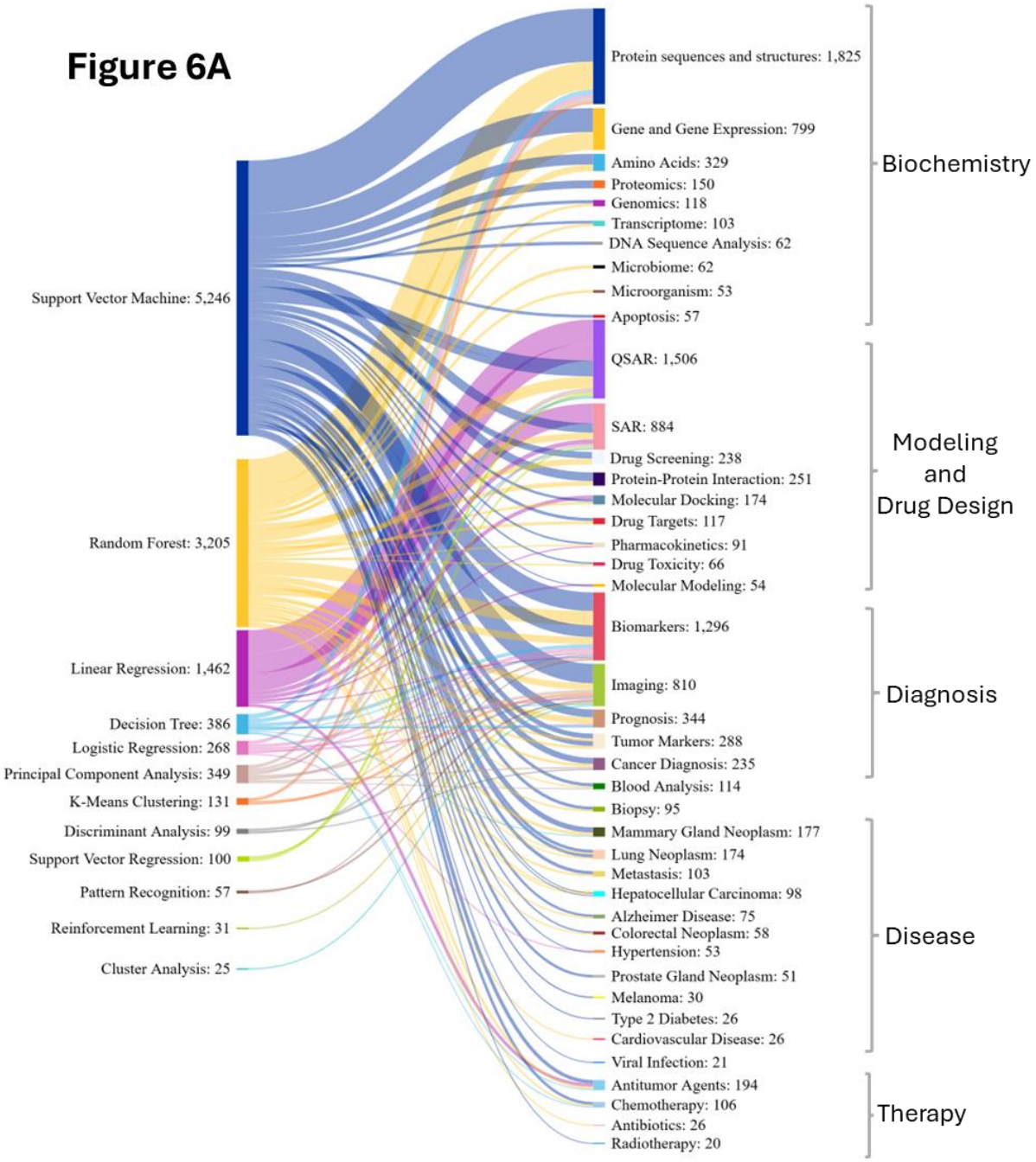
Various algorithms are used in diabetes research for predicting diabetic complications and optimizing treatment protocols taking advantage of longitudinal monitoring data, complex interactions between lifestyle factors and genetics, and time-series prediction problems for glucose levels. In Alzheimer's disease, RF and SVM models are widely used for early prediction by analyzing neuropsychological test scores, biomarker levels (amyloid-beta, tau proteins), and demographic data to classify patients into healthy or mild cognitive impairment categories.

In the **Therapy domain**, antitumor agents received the most research attention, employing ML methods for drug screening and activity prediction by analyzing molecular descriptors and chemical properties to identify compounds with potential anticancer activity. In chemotherapy, these methods predict treatment responses and toxicity by analyzing patient clinical data, genetic profiles, and tumor characteristics for personalizing drug selection and dosing protocols. Similarly, these models support immunotherapy by

analyzing patient immune profiles, tumor mutational burden, and biomarker expressions to predict response and optimize therapeutic outcomes, addressing the complexity of immune interactions and temporal dynamics.

Co-occurrences of AI models and biomedical concepts

Our analysis explores the co-occurrence patterns between biomedical concepts and various AI/ML methods, elucidating their strategic applications across the five domains (see Figure 6).



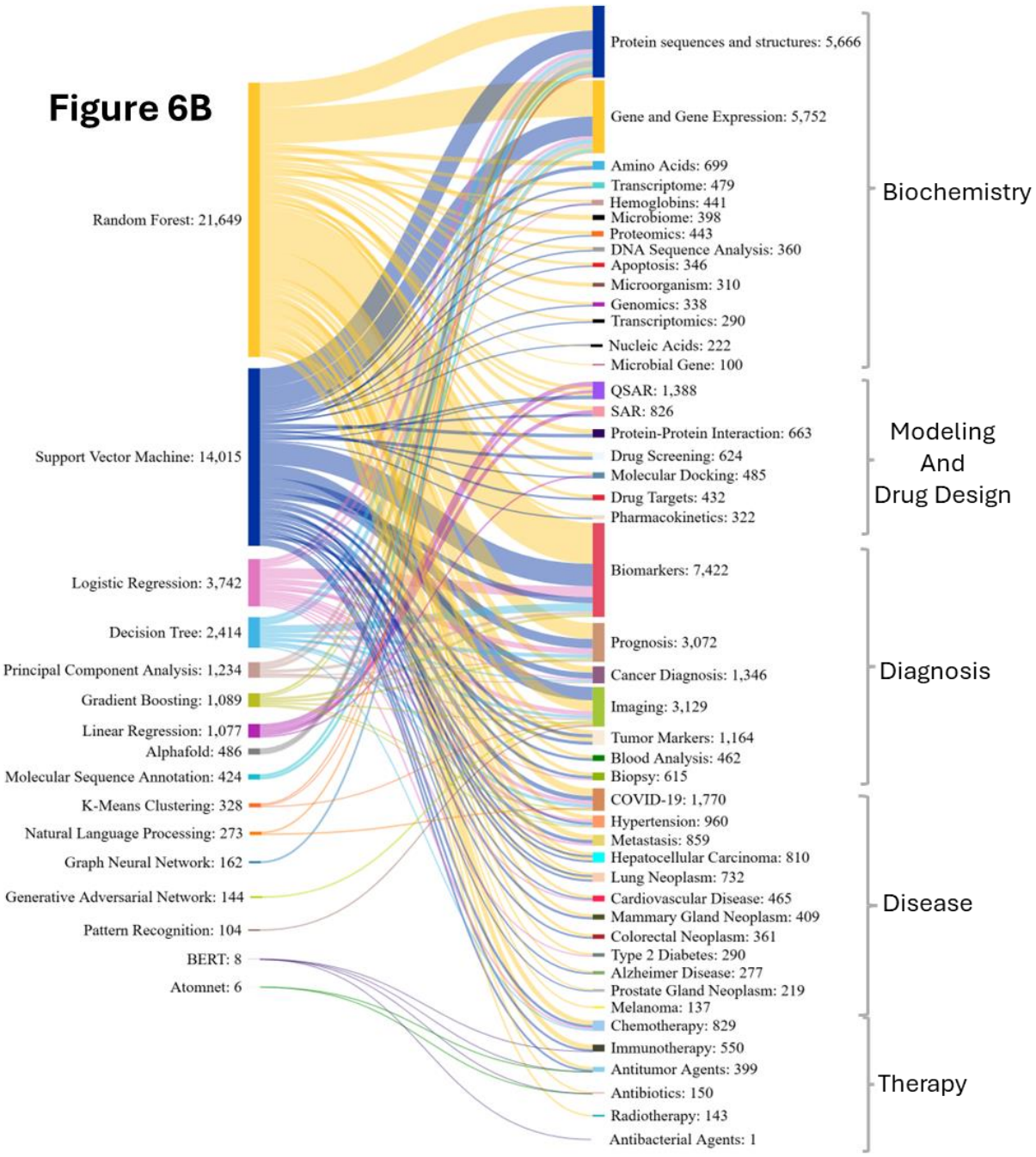


Figure 6 A and B: *Sankey diagram illustrating the co-occurrence of AI methods and biomedical concepts from (A) 2015-2019 and (B) 2020-2024. Source: CAS Content Collection.*

Random Forest (RF) has emerged as the dominant AI method in biomedical research, showing remarkable growth and overtaking Support Vector Machine (SVM) between the two periods. RF shows strong co-occurrence with concepts across all the domains of biomedical sciences, particularly in biochemistry, diagnosis, and disease-related research. RF is used for protein function prediction and classification by analyzing amino acid composition, sequence motifs, and physicochemical properties to classify proteins into functional families. In biomarker discovery, RF is used to analyze high-dimensional genomics, proteomics, and metabolomics data.

Support vector machine (SVM) while growing steadily between the periods, maintains significant presence with proteins, gene expression, DNA and genomics, biomarkers, prognosis, medical imaging, and cancer research. In protein studies, SVM is used for protein classification and functional annotation by analyzing amino acid sequences, structural features, and physicochemical properties. It also predicts subcellular localization and identifies enzyme classes using kernel-based high-dimensional feature space mapping.

In gene expression, SVM identifies genes involved in specific biochemical processes, classifies metabolic pathways, and predicts enzyme-encoding genes based on their expression patterns across different biochemical conditions and cellular states. For genomic variant classification, SVM processes DNA sequence features and annotations to distinguish pathogenic from benign mutations and identify disease-associated genomic regions through high-dimensional feature analysis of nucleotide patterns and genomic context.

In biomarker research, SVM supports identification, classification, and validation by integrating multiple biomarkers into predictive models for diagnosis, treatment response, disease progression risk assessment, and stratifying patients for personalized medicine approaches. SVM is used in cancer research to distinguish between cancer and normal tissues, classify cancer types, and identify molecular subtypes within specific cancers. It is also used for cancer prognosis and treatment prediction by integrating clinical parameters, biomarker profiles, and genomic features to predict patient survival outcomes, treatment responses, drug resistance patterns, and recurrence risk. Figure 6B reveals that SVM maintained strength in QSAR and SAR modeling while its connections to protein analysis weakened relative to RF in the later period.

Logistic Regression (LR) is a widely used, interpretable statistical method in biomedical research, often co-occurring with key concepts such as proteins, gene expression, biomarkers, prognosis, medical imaging, cancer, and COVID-19. In protein function prediction, LR analyzes amino acid composition, sequence features, and structural properties to classify proteins into functional or structural categories. The model's key advantage is providing interpretable coefficients that indicate the relative contribution of each feature to the classification decision, allowing researchers to identify which amino acid properties or structural features most strongly influence the predictions.

This interpretability is equally valuable in gene expression classification, where LR helps identify differentially expressed genes. For biomarker validation and diagnostics, LR models predict binary outcomes, such as disease/healthy or responder/non-responder and provide clinically meaningful odds ratios for diagnostic decision making. In prognostic modeling, LR uses a similar binary outcome prediction, by evaluating clinical parameters, biomarker profiles, and patient demographics to forecast outcomes like mortality/survival, disease recurrence/remission, favorable/unfavorable prognosis.

LR is used in COVID-19 diagnosis, severity prediction, and mortality by integrating comorbidities, vital signs, and biomarkers, supporting risk stratification and clinical decision-making. These diverse applications underscore LR's strong alignment with core biomedical research themes, making it a foundational tool for interpretable and clinically relevant modeling. Figure 6B illustrates the increased adaptability of Logistic Regression, likely due to the capability for multi-class classification and probability-based outputs, unlike linear regression, which is limited to continuous predictions.

As seen in Figure 6B, the 2020-2024 period witnessed the emergence of specialized deep learning approaches barely present earlier, such as AlphaFold. This period also saw a surge in the use of Graph Neural Networks (GNN) for molecular interaction modeling, Graph Adversarial Networks (GAN) for synthetic data generation, and Pattern Recognition for imaging applications and diagnostic challenges. Another major shift in recent years is the integration of NLP techniques into biomedical research, highlighted by the rise of models like BERT for mining clinical reports and analyzing biomedical texts. This

specialization reflects the field's maturation beyond applying generic AI methods toward developing tailored approaches for specific biomedical challenges.Research priorities shifted substantially between the two periods. Protein sequence and structure analysis and gene expression analysis grew significantly in 2020-2024(Figure 6B). Biomarker discovery and imaging applications also experienced dramatic growth, while disease-specific research expanded considerably, with COVID-19 emerging as a major focus and cancer research diversifying across multiple subtypes.

Substances in biomedical research

This comparison provides strategic insight into under-explored therapeutic opportunities, and it highlights substance classes that may benefit from increased research investment or innovative approaches to bridge the gap between scientific discovery and clinical application (see Figure 7). Please note that “substance classes” refers to categories of CAS-indexed substances with recognized potential therapeutic applications.

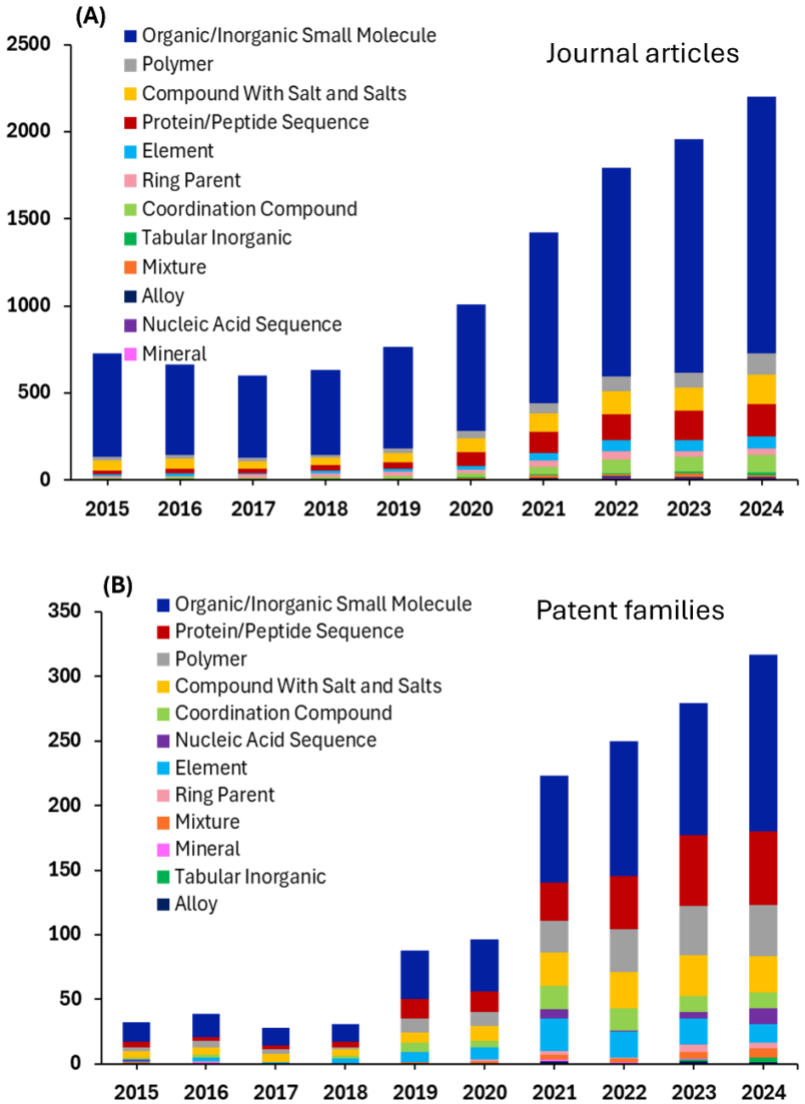


Figure 7: *Bar chart representation of key therapeutic substance classes, based on publication frequency for (A) journal and (B) patent publications. Source: CAS Content Collection.*

Small molecule drugs continue to dominate pharmaceutical research and development, with AI technologies accelerating their discovery. ML-based virtual screening, QSAR modeling for pharmacokinetic prediction, and deep learning systems optimizing synthetic pathways are now integral to this process. These innovations build upon the inherent advantages of small molecules, including established synthesis routes, well-characterized pharmacokinetics, oral bioavailability, and cost-effective manufacturing.

Protein/peptide therapeutics constitute the second-most-researched category. Tools like AlphaFold have revolutionized protein structure prediction, enabling the design and optimization of therapeutic proteins such as insulin for diabetes, tissue plasminogen activator for strokes, and monoclonal antibodies for cancer and autoimmune diseases. ML models further enhance this field by predicting protein-protein interactions and optimizing peptide stability.

Compounds with salt and salts benefit from AI-powered formulation optimization algorithms that predict solubility, bioavailability, and stability, reflecting the importance of drug formulation and bioavailability optimization. Polymers, which often serve as crucial therapeutic delivery systems, are being enhanced through AI-designed nanoparticle formulations, ML-optimized hydrogel properties, and predictive modeling for targeted drug delivery improving therapeutic efficacy and tissue specificity.

Coordination compounds are another promising class, where AI assists in ligand design and metal-organic framework (MOF) optimization. These compounds are being refined through computational chemistry and AI models that predict metal-ligand interactions and biological activity.

Applications of AI models in materials science

Overview

In materials science, understanding and predicting the complex relationships between material composition, structural features, and properties is essential for accelerating material discovery. The data involved are often multidimensional, hierarchical, heterogeneous, and generated through high-throughput, data-intensive processes. AI, particularly ML, is revolutionizing the field by enabling the analysis of these complex datasets. We analyzed the CAS Content Collection to identify key materials science concepts where AI methods are applied and assessed their significance in materials discovery and property prediction (see Figure 8).

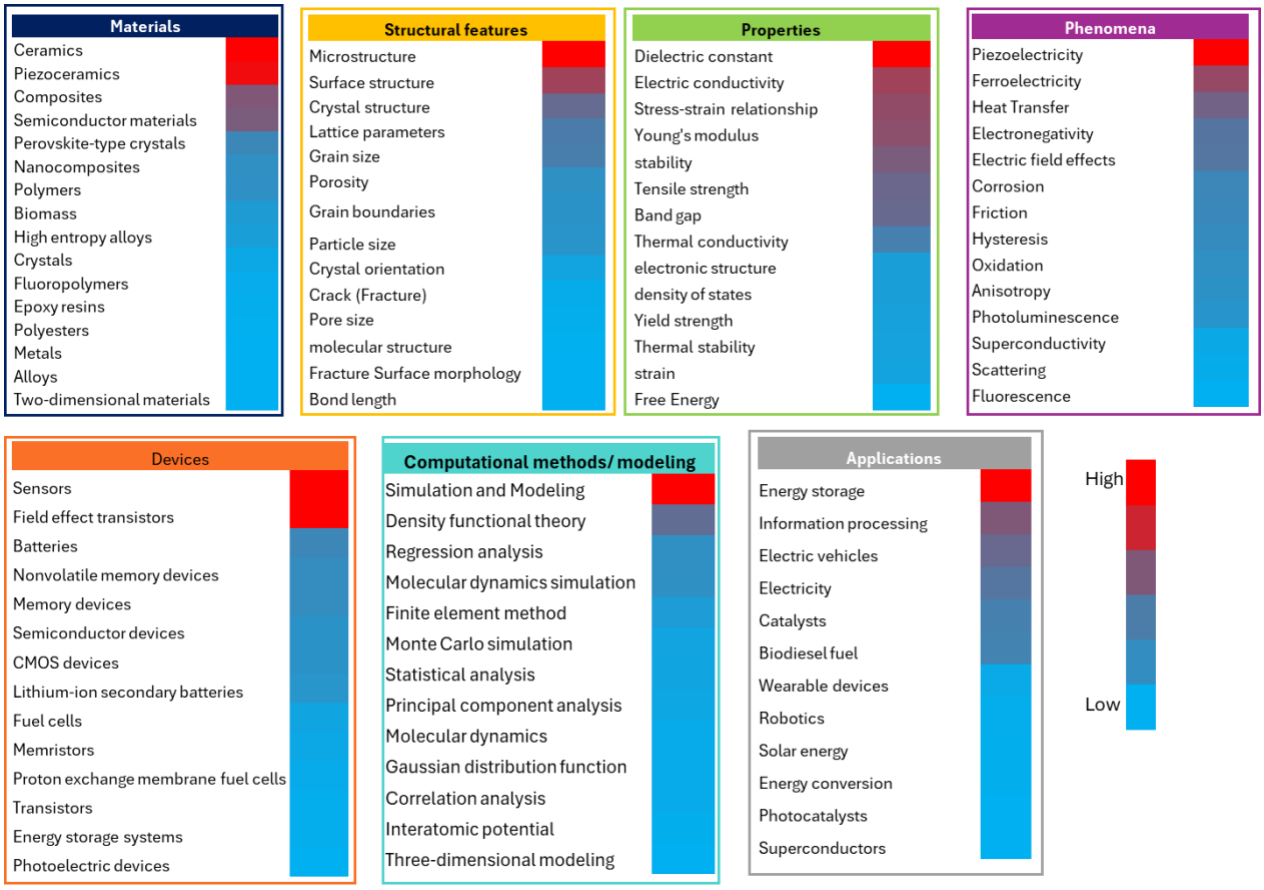


Figure 8: *Conceptual landscape of AI-driven materials science publications, categorized into seven categories. Each category includes representative concepts, with color coding red (highest) to blue (lowest) indicating publication frequency. Source: CAS Content Collection.*

Energy storage is the most AI-intensive research area in **Applications**, reflecting the global need for advanced battery technologies and sustainable energy solutions. ML techniques have become potent tools for battery material innovation, with researchers utilizing three main strategies: direct property predictions, machine learning potentials, and inverse design. The high concentration of AI research in this field stems from the complex multi-scale nature of battery systems, where traditional experimental approaches often fall short in optimizing the intricate relationships between material composition, structure, and electrochemical performance.

In the **Materials** and the **Devices** categories, the significant AI activity in ceramics, piezoceramics, sensors, and field effect transistors reflects the sophisticated nature of these materials and their devices. These advanced materials often exhibit complex structure-property relationships that are ideal candidates for ML approaches.

In **Structural Features**, we found that convolutional neural networks (CNNs) along with computer vision are used for microstructural analysis by enabling automated classification, segmentation, and feature extraction from microscopy images. These tools are helping uncover structure-property relationships that were previously difficult to quantify. In metallurgy, targeted AI approaches are advancing superalloy development, while systems like CAMEO (Closed-Loop Autonomous Materials Exploration and Optimization) enable combinatorial metallurgy by integrating AI with high-throughput experimentation.

Furthermore, machine learning tools for 3D tomography data are enhancing the analysis of porous materials, allowing for more accurate modeling of properties like permeability and mechanical strength.

Dielectric constant and electric conductivity are the leading concepts in **Properties**, while piezoelectricity leads in **Phenomena**. In **Simulation and Modeling**, computational modeling shows high AI adoption, which aligns with machine learning's natural compatibility with data analysis tasks.

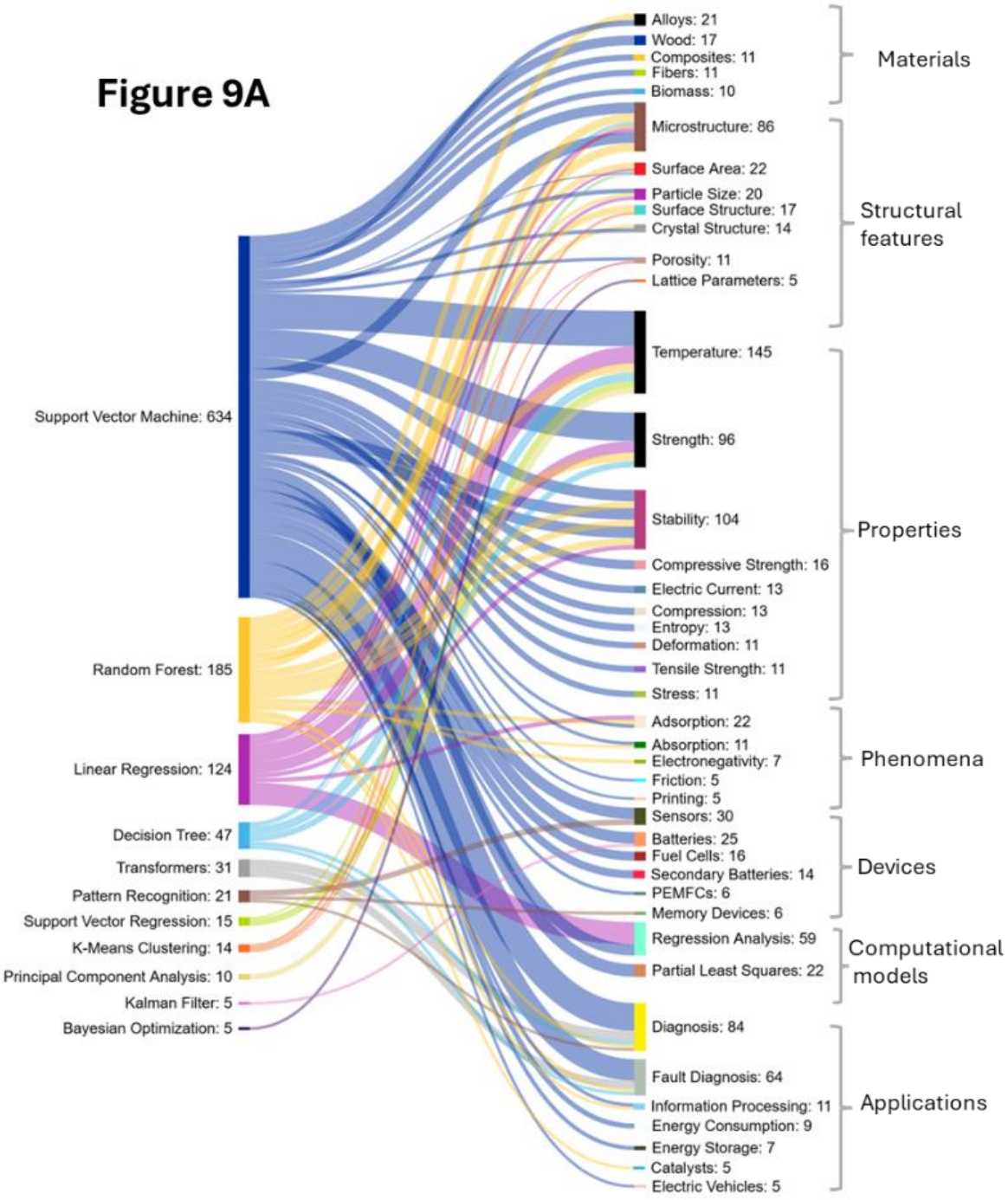
The moderate AI activity areas represented by purple and mixed-colored regions in Figure 7 are fields where AI adoption is accelerating but have room for improvement. For example, low activity in some materials areas, predominantly shown in blue, represents some of science's most established and fundamental domains, creating a paradoxical situation where the most mature fields show the least AI integration. Composite materials and polymers are prime examples of this category.

However, specialized AI models are being developed to address specific challenges. The model polyBERT, treats polymer structures as a chemical language, enabling more effective modeling of complex polymer systems. For crystalline materials, models like MEGNet, CGCNN, and SchNet incorporate crystal symmetry and quantum interactions, while ALIGNN captures higher-order atomic interactions essential for accurate alloy property prediction.

Co-occurrences of AI models and materials science concepts

As with our biomedical analysis, we looked closely at AI methods and co-occurring concepts in materials science. We found that traditional ML models such as Random Forest (RF), Support Vector Machine (SVM), Gradient Boosting Machines (GBM), and Decision Trees (DT) are widely used for prediction and feature importance analysis, largely due to the high-dimensional data with non-linear relationships that is common in materials informatics (see Figure 9).

Figure 9A


Support Vector Machine: 634
Random Forest: 185
Linear Regression: 124
Decision Tree: 47
Transformers: 31
Pattern Recognition: 21
Support Vector Regression: 15
K-Means Clustering: 14
Principal Component Analysis: 10
Kalman Filter: 5
Bayesian Optimization: 5
Alloys: 21
Wood: 17
Composites: 11
Fibers: 11
Biomass: 10
Materials
Microstructure: 86
Surface Area: 22
Particle Size: 20
Surface Structure: 17
Crystal Structure: 14
Porosity: 11
Lattice Parameters: 5
Structural features
Temperature: 145
Strength: 96
Stability: 104
Compressive Strength: 16
Electric Current: 13
Compression: 13
Entropy: 13
Deformation: 11
Tensile Strength: 11
Stress: 11
Properties
Adsorption: 22
Absorption: 11
Electronegativity: 7
Friction: 5
Printing: 5
Phenomena
Sensors: 30
Batteries: 25
Fuel Cells: 16
Secondary Batteries: 14
PEMFCs: 6
Memory Devices: 6
Devices
Regression Analysis: 59
Partial Least Squares: 22
Computational models
Diagnosis: 84
Fault Diagnosis: 64
Information Processing: 11
Energy Consumption: 9
Energy Storage: 7
Catalysts: 5
Electric Vehicles: 5
Applications

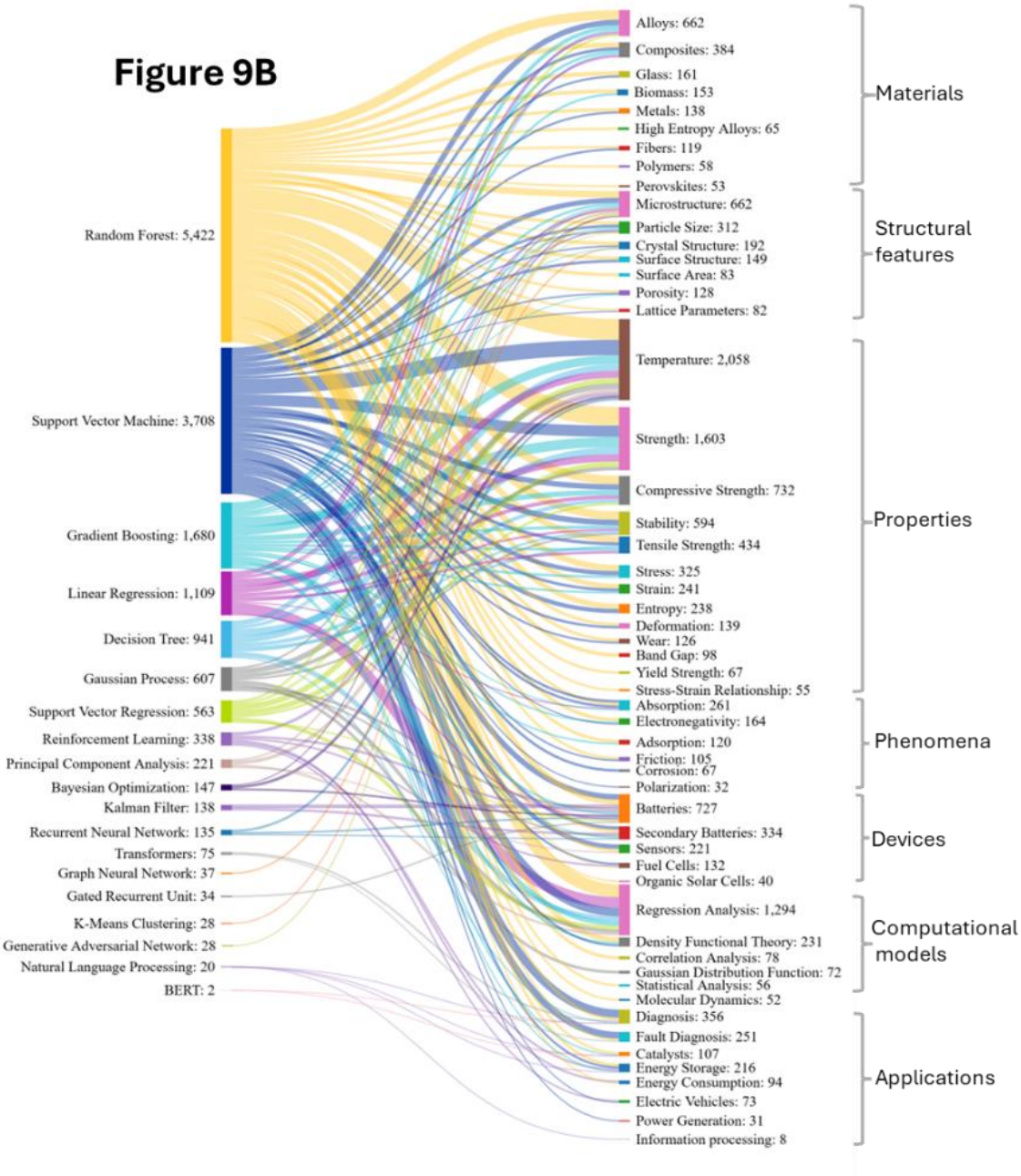


Figure 9: *Sankey diagram illustrating the co-occurrence of AI methods and materials science concepts from (A) 2015-2019 and (B) 2020-2024. Source: CAS Content Collection.*

RF models are widely used and distributed across all concept areas. Along with **GBM models**, they show strong co-occurrences with complex, heterogeneous materials due to their ability to model nonlinear relationships and handle mixed data types. They're often used for heterogeneous composites, predicting mechanical properties, and failure analysis by capturing the complex interactions between matrix and reinforcement.

We see these co-occurring with alloys, where RF is used for mapping nonlinear relationships in alloy compositions, aiding in high-entropy alloy design, precipitation strengthening prediction, and heat treatment optimization. GB is increasingly applied for precise property tuning, such as optimizing yield strength or corrosion resistance.

SVM is more frequently used with materials that have well-defined structure-property relationships. This type of model is effective for classifying alloy types and thresholds using fixed-dimensional feature vectors. Temperature-dependent properties also often follow non-linear patterns and depend on composition, processing history, and environmental conditions, which tree-based methods and SVMs capture effectively.

For structural features like microstructure, ensemble models are employed to identify features from complex images; classify grain boundaries, phases, and defects; and predict crystal structures, often treated as classification problems requiring symmetry-aware models. Surface properties, which span atomic to macroscopic scales, involve complex patterns in surface energy and reactivity. These applications demand the preprocessing of image data, extraction of multi-scale features, and encoding of symmetry and periodicity, especially for catalysis, corrosion, and adhesion studies, often requiring integration with molecular modeling.

Application domains have expanded considerably, with batteries research showing dramatic growth. Figures 9A and 9B confirm RF's strong correlation with electrochemical processes, handling complex interactions for battery capacity prediction, electrode material selection, and time-dependent processes. The emergence of energy-related applications (Energy Storage, Power Generation) in the 2020-2024 period demonstrates the field's diversification. Long-Short Term Memory (LSTM) captures temporal dynamics which are crucial for battery degradation prediction and cycling performance forecasting. Another emerging correlation is the Reinforcement Learning (RL) which is capable of optimizing charging protocols and material usage and can be used in battery management systems. Kalman filters are also observed with batteries and secondary batteries which are used to track material state during processing and forecast battery state-of-health estimation and remaining useful life prediction.
Transformers are used in extracting synthesis procedures and material properties from scientific literature and capturing long-range dependencies in complex material descriptions. Convolutional Neural Networks (CNNs) and GANs automate feature extraction from 2D/3D structures and support property prediction and pattern recognition. Despite their power, they require more data and computation and offer lower interpretability, thus complementing rather than replacing traditional models. NNs are increasingly used in image-based analysis and generative design, while traditional methods maintain advantages in composition-property mapping with limited datasets. This is another reason why Neural Networks and traditional models are used in combination. Another development is materials specific CNN architectures incorporating symmetry and physical constraints, while GNNs are gaining traction for atomic and molecular structures.

Substances in materials science

Returning to categories of CAS-indexed substances, we analyzed the number of publications relating to these substances in materials science (see Figure 10).

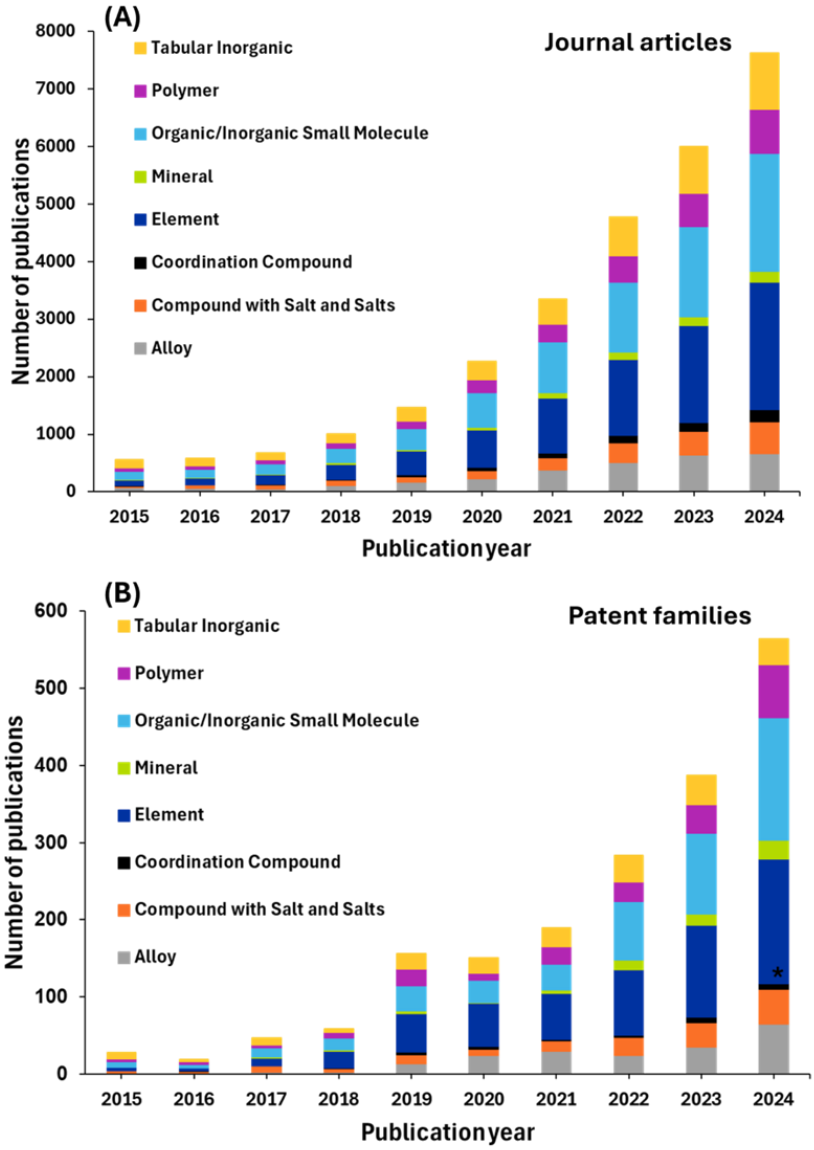


Figure 10: *Bar chart representation of key substance classes in materials science, based on publication frequency for (A) journal and (B) patent publications. Source: CAS Content Collection.*

The dominance of **organic/inorganic small molecules** in AI-driven research reflects the field's computational advantages and data abundance in handling small structures in terms of computational cost versus larger complex systems such as polymers. With approximately 6,500 journal articles, this category benefits from decades of accumulated chemical databases.

Small molecules are particularly amenable to ML approaches because their properties can be effectively encoded using molecular descriptors, fingerprints, and graph-based representations. The high volume of

research in this area also indicates the pharmaceutical and chemical industries' strong investment in AI for drug discovery, catalyst design, and molecular property prediction.

Elements, ranking second with around 5,500 articles, similarly benefit from extensive databases and the relative simplicity of their structures, making them ideal candidates for AI-driven property prediction and materials discovery.

Tabular inorganic materials, with approximately 4,500 journal articles, represent another area where AI has found substantial applications. These materials — which include ceramics, oxides, and other structured inorganic compounds — benefit from crystallographic databases and systematic property measurements that provide the large datasets necessary for effective machine learning. The structured nature of these materials allows for consistent feature engineering based on crystal structure, composition, and bonding characteristics. This category's prominence in AI research likely reflects the materials science community's focus on discovering new functional materials for energy storage, catalysis, and electronic and magnetic applications.

Polymer research shows moderate AI adoption with around 1,800 journal articles, which may seem low given the industrial importance of polymers. This likely reflects the unique challenges polymers present to AI applications, including their complex chain structures, molecular weight distributions, and processing-dependent properties. However, the growing research volume suggests increasing success in applying AI to polymer property prediction, synthesis, and processing optimization.

Another important finding from our analysis is how patent families represent only a fraction of journal articles across all materials science categories. The disparity between the number of journals and patents suggests that much of the AI research in materials science remains in the fundamental or exploratory stage, with significant time lags between academic discoveries and practical applications worthy of patent protection at the commercial stage.

The relatively small patent numbers may also reflect the challenge of translating AI-driven materials discoveries into commercially viable technologies, as many AI applications in this field focus on property prediction and fundamental understanding rather than immediately patentable inventions. The consistent ratio across different material types indicates that the academic-to-commercial translation challenge is universal across materials science, rather than being specific to specific material classes.

AI in process management

AI is transforming process management by enabling smarter, faster, and more adaptive decision-making. Analyzing AI's role in this domain reveals how automation, predictive analytics, and real-time optimization can drive operational excellence across industries (see Figure 11).

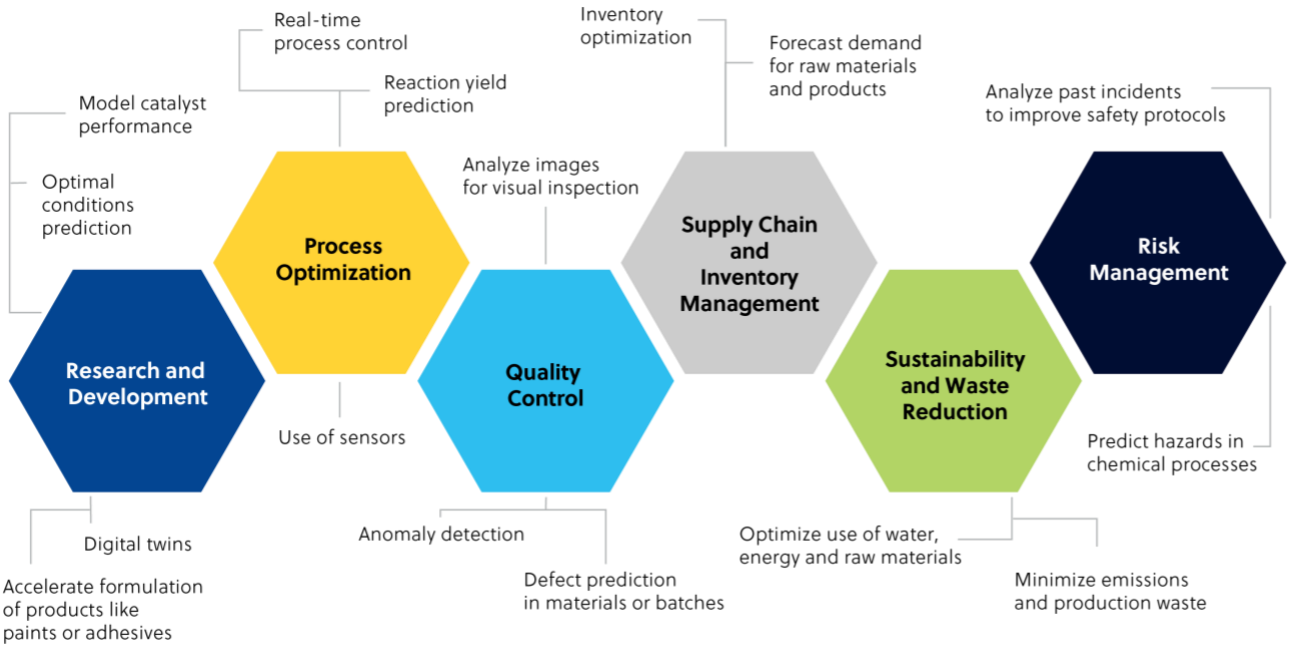


Figure 11: *Framework illustrating the role of artificial intelligence in modernizing research, development, and operational processes across scientific domains.*

AI-driven optimization is widely used in additive manufacturing, petrochemical industries, plastic processing, metallurgy, flow chemistry, catalytic processes, and drug discovery and synthesis. Commonly used models include response surface methodology (RSM), design of experiments along with ML techniques like ANNs, hybrid models, PINN, LLMs, and Reinforcement Learning approaches. Let’s explore a few other technologies that are extending predictive modeling into real-time decision-making and automation:

- **Real-time monitoring:** these systems analyze streaming data using ML to detect patterns, trends and early warning signs, facilitating dynamic forecasting and proactive decision making. Key innovations include autonomous decision making, where parameters are automatically adjusted based on real-time inputs; predictive maintenance, which anticipates failures before they occur; and self-improving models that adjust to changing conditions. Additionally, real-time data visualization and intelligent alert systems are transforming how operators interact with live process data. These advancements are being applied across industries like automated chromatography systems and drug reaction monitoring in pharmaceutical manufacturing. We’re also seeing their use in property optimization in polymer production, solid waste analysis in waste management, environmental monitoring, energy and resource management for large-scale hydrogen production, and cloud-based energy management for aging infrastructure.

- **Soft sensors:** Using real-time data from physical sensors, these leverage AI and statistical models to estimate variables that are difficult or costly to measure directly. They are widely applied across industries including bioprocesses monitoring and monitoring wastewater and air quality. In chemical process control, soft sensors are used to monitor complex sulfide ore floatation, isomerization, reactive distillation, and gasoline blending.

- **Digital twins:** These innovations are virtual replicas of physical systems that are rapidly evolving with AI integration to enable real-time optimization, predictive analytics, and autonomous operations. AI enhances digital twins by learning from historical and real-time data to simulate future states, detect anomalies, and predict failures, transforming them from static models into resilient, self-optimizing systems. In manufacturing, AI-driven digital twins predict equipment failures, optimize production processes, and detect quality issues using sensors and images. In healthcare, they simulate patient conditions, support personalized medicines, model diseases, clinical trials simulation, and optimize immunotherapy. In environmental and sustainability domains, AI-powered digital twins are applied in geological carbon storage, ore waste processing, and resource management. They also support materials science, enabling automated discovery, defect analysis, and inverse molecular design.

Challenges to the use of AI in scientific research

AI has clearly made many positive contributions to all scientific research fields, and it will continue to do so as it advances. However, these innovations are not without challenges that researchers and data scientists must address to ensure AI reaches its full potential in science. Some of the most pressing challenges include:

- **Data privacy and security:** Ensuring data privacy and security is one of the largest challenges in implementing AI since it needs a large volume of sensitive data for training and operation. Issues occur when sensitive information is gathered and used without proper permission, or when sensitive information is leaked or stolen from AI systems. Scientific datasets may contain unpublished experimental results, novel compound structures, and proprietary synthesis methods that represent significant competitive advantages. When researchers utilize cloud-based AI platforms or participate in collaborative research environments, they face substantial risks of intellectual property exposure. To combat the risk, some businesses are implementing safeguards to protect client information and maintain trust, with new start-ups finding market niches to safeguard against external AI threats.

- **Data quality and bias:** Datasets frequently suffer from systematic biases that can propagate through AI models, leading to skewed predictions and potentially reinforcing existing research inequities. Historical biases in published reaction data, where successful reactions are overrepresented and failed experiments underreported, can significantly impair the ability of ML models to explore novel chemical space, particularly for inorganic materials. This "publication bias" creates a self-reinforcing cycle where AI systems preferentially recommend compounds similar to those already well-studied. Generative AI models present additional concerns, as they can manipulate existing data and hallucinate to satisfy customer requests at the expense of real evidence.

- **Transparency:** The "black box" nature of these models can obscure the reasoning behind their predictions, making it difficult for researchers to evaluate their reliability or identify potential failure modes. A review of explainable AI methods in drug discovery highlighted how the lack of interpretability in complex models can undermine trust among medicinal chemists and impede regulatory acceptance of AI-guided decisions in pharmaceutical development. This challenge is particularly acute for graph neural networks and transformer models that operate on high-dimensional chemical representations, where the relationship between input features and predictions becomes highly non-linear.

- **Balancing automation with human expertise:** While AI systems can process vast datasets and identify patterns beyond human cognitive capacity, they lack the intuition, creativity, and contextual understanding that experienced scientists bring to research problems. This may lead to a generation of scientists who lack fundamental research skills and conceptual understanding. Conversely, it can also lead to "algorithmic aversion" when AI-driven conclusions contradict intuition.

AI models and the future of scientific research

AI is only going to increase in importance for scientific inquiry, and its revolutionary capabilities are already being realized in fields such as biomedicine and materials science. As our analysis of the CAS Content Collection shows, there are new applications, models, and methods being applied to difficult research questions all the time. We can expect breakthroughs in drug discovery, disease diagnosis, materials development, and much more to be influenced or entirely driven by AI-powered technologies.

As AI shifts from tool to collaborative scientific partner, we're also likely to see new paradigms like inverse design and self-driving laboratories redefining scientific methodology toward autonomous discovery engines. However, low patent-to-publication ratios indicate much research remains in academia, particularly in traditional materials science fields.

Further AI adoption requires appropriate trust calibration through technical solutions like uncertainty quantification and model transparency. The wide-ranging nature of AI's impact on science and the scope of the challenges in implementing it mean that collaboration and visibility will be key to maximizing its benefits for scientific advancement.