WHITEPAPER

Materials Informatics

Accelerating the Future of Materials Discovery with AI and Machine Learning

Executive Summary

This white paper explores how materials informatics the fusion of materials science, Artificial intelligence (AI), and Machine learning (ML) is transforming materials discovery. Combining data-driven modeling, automation, and active learning approach to accelerate innovation, reduces R&D costs, and enables self-driving laboratories that operate with minimal human intervention. The paper also addresses challenges around data quality, model generalization, and interdisciplinary integration. Finally, it outlines Tech Mahindra's role in advancing this field through its Makers Lab ecosystem, which leverages AI-powered data platforms and simulation tools to drive sustainable, high-impact innovations in materials research.

Key Takeaways

1 AI-Driven Data Curation & Aggregation Platforms

- **Capability:** Building centralized platforms to compile, clean, and structure data from diverse sources (lab notebooks, publications, patents).
- **Strategic Value:** Addressing a major industry pain point- data fragmentation and poor metadata quality. By creating machine-readable repositories, Tech Mahindra can enable more accurate and scalable ML model training.
- **Opportunity:** How Tech Mahindra is a leading data backbone provider for materials science, like what AWS is for cloud infrastructure.

2 AI-Powered Modeling & Simulation

- **Capability:** Using deep learning (e.g., graph neural networks) and quantum computing to simulate atomic-level interactions.
- **Strategic Value:** Reducing prototyping costs and speeding up material validation. This is especially valuable for clients in the semiconductor, energy, and advanced manufacturing sectors.
- **Opportunity:** Offering cloud-based virtual testing environments as a service-allowing clients to test materials digitally before physical synthesis.

3 Generative AI for Material Design

- **Capability:** Developing generative models (e.g., VAEs, transformers) to design materials based on desired properties.
- **Strategic Value:** Enabling **inverse design workflows,** where clients specify performance goals and receive candidate materials.
- **Opportunity:** Launching a **"Materials Design Studio"** platform-an intuitive interface for R&D teams to explore AI-generated material options.

4 Industry-Academia Collaboration

- **Capability:** Partnering with universities, national labs, and industry players to access specialized datasets and validate AI predictions.
- **Strategic Value:** Enhancing credibility and accelerates innovation cycles. Additionally, it helps **standardize AI practices** in materials science.
- **Opportunity:** Leading **joint initiatives** to co-develop open standards, benchmark datasets, and validation protocols for AI in materials discovery.

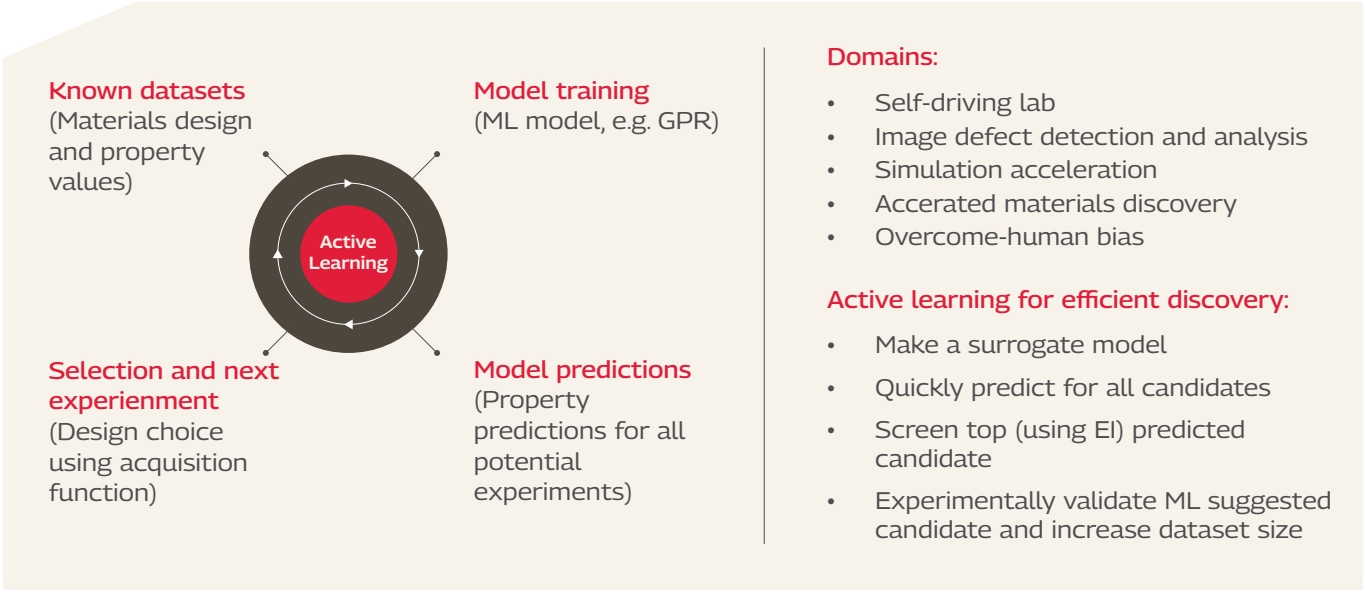
Introduction

Historically, advances in materials science relied on trial-and-error experimentation, researcher intuition, and discovery. While this empirical approach produced breakthroughs such as semiconductor-grade silicon, lightweight aluminium alloys, and high-strength steels, it is inherently slow, costly, and has a limited scope. The traditional rate of discovery cannot meet the pressing global challenges of developing next-generation energy technologies or sustainable materials for a greener economy.

Materials informatics offers a new paradigm by merging materials science with Artificial intelligence (AI) and Machine learning (ML). Researchers can accelerate progress from concept to implementation using extensive datasets, statistical learning, and computational modelling.

At its core, the approach uses surrogate ML models trained on existing data to predict material properties thousands of times faster-and at a fraction of the cost-of traditional methods. Raw materials data are converted into numerical formats, capturing essential chemical, structural, and processing characteristics. With these fingerprints, ML algorithms-ranging from decision trees to graph neural networks-can detect patterns, predict properties, and recommend new material formulations or synthesis routes.

Process diagram illustrating the workflow:



Domains:

- Self-driving lab
- Image defect detection and analysis
- Simulation acceleration
- Accerated materials discovery
- Overcome-human bias

Active learning for efficient discovery:

- Make a surrogate model
- Quickly predict for all candidates
- Screen top (using EI) predicted candidate
- Experimentally validate ML suggested candidate and increase dataset size

Active learning enhances this process by allowing ML models to recommend the next most informative experiments. Researchers can explore vast parameter spaces with minimal effort through closed-loop design-build-test-learn cycles. Combined with automated synthesis and characterization platforms, this enables self-driving laboratories—autonomous experimental systems that continuously generate, test, and refine materials without human intervention.

Beyond speed, materials informatics promotes a broader ecosystem of materials intelligence: generative models that "invert" discovery to design materials meeting target specifications; natural language processing (NLP) tools that mine decades of literature for hidden structure-property-processing relationships; and open repositories such as the Materials Project, OQMD, and NOMAD that house millions of entries. These developments foster a feedback-rich environment where computation, experimentation, and analytics interact seamlessly, ushering in unprecedented levels of AI-assisted scientific discovery.

Ultimately, the fusion of machine guidance and human intuition redefines the scientific method. This human-machine partnership enables researchers to meet emerging technological demands, optimize performance for specific applications, and systematically navigate the vast materials landscape. This will drive innovation across energy, healthcare, infrastructure, and sustainable manufacturing—ensuring that materials discovery keeps pace with the needs of the 21st century.

Why Material Informatics Matters

One of the main challenges facing science and engineering is the vastness of the materials space, which is why materials informatics is so important. Every combination of elements, crystal structures, and processing methods could yield a new material—but the sheer number of possibilities makes exhaustive exploration through traditional trial-and-error approaches practically impossible within the lifetime of the universe. Data-driven approaches in materials informatics help identify trends in existing experimental and computational data, highlighting the most promising candidates for focused research rather than exhaustively testing every option.

This approach's strength lies in its ability to convert material knowledge into machine-readable formats, known as "fingerprints," which capture key traits at various levels of detail-from atomic configurations to elemental chemical compositions. By linking these fingerprints to properties or performance results, machine learning algorithms can develop surrogate models that evaluate new candidates much faster than traditional simulations or experiments. This shift is not just about speed; it also enables entirely new research methods, such as inverse design-which creates candidate materials matching specified properties-and active learning, where algorithms guide experimental efforts toward high-value, unexplored areas of the design space.

Many successes have already showcased the impact: superhard ceramics, high-efficiency thermoelectrics, high-entropy alloys with exceptional strength, and metallic glasses with targeted properties have all been discovered through ML-guided screening. In each case, integrating data sources, domain expertise, and prediction algorithms revealed previously unknown structure-property relationships while reducing costly and time-consuming experimental cycles. Additionally, the accessible knowledge pool is expanding rapidly, thanks to large open repositories like Materials Project, AFLOW, and NOMAD, as well as natural language processing techniques that mine scientific literature.

The economic impact is significant. For a fraction of the cost of a single high-fidelity simulation or synthesis experiment, machine learning models can evaluate millions of hypothetical materials once trained. When combined with automated characterization, robotic synthesis, and advanced computational infrastructure, these tools form "materials intelligence ecosystems"—integrated, semi-autonomous research environments that continuously design, test, and optimize materials.
These systems have the potential to accelerate innovation, reduce R&D costs, and enable the rapid deployment of new materials in fields such as quantum computing and renewable energy. In summary, materials informatics represents a paradigm shift in the discovery and development of materials. By combining machine learning's predictive capabilities with curated data and automated experimentation, it transforms an otherwise intractable search into a guided exploration.

The Toolbox of Materials Informatics

Over the past decade, the materials informatics toolbox has expanded significantly, evolving from basic statistical techniques to a comprehensive ecosystem of algorithms, data pipelines, and automated experimentation. The field utilizes traditional regression and classification models to map materials to characteristics such as hardness, conductivity, or thermal stability. While more sophisticated algorithms, such as deep neural networks and Gaussian process regression, handle non-linear, high-dimensional relationships in richer datasets, more conventional techniques like linear regression, decision trees, kernel methods, and random forests remain helpful due to their interpretability and efficacy in low-data regimes.

One of its distinguishing features is the capacity of contemporary materials informatics to learn adaptively through Bayesian optimization and active learning. In this case, the machine learning model does more than just passively fit data; it suggests the next most instructive experiment or simulation to run, striking a balance between exploring uncharted material space and exploiting promising leads. Because fewer expensive physical or computational tests are required thanks to this closed-loop method, innovations such as the targeted discovery of high-performance piezoelectric and high-entropy alloys with limited initial data are made possible.

The incorporation of physics-informed machine learning is another developing aspect. These models incorporate well-known physical constraints, such as conservation laws, thermodynamic relations, or symmetry rules, directly into the learning process rather than treating materials as "black boxes." By adhering to these guidelines, physics-informed neural networks increase extrapolative power and predictive accuracy, enabling trustworthy predictions in uncharted areas of materials space. In fields where physical theory is developed but experimental data is limited, this blending of data-driven and physics-based reasoning is especially crucial.

New avenues for knowledge extraction have been made possible by large language models (LLMs) and natural language processing (NLP). Domain-specific LLMs can mine large databases of scientific publications and patents to create structured materials databases that capture properties, synthesis conditions, and relationships between processing, structure, and performance at scale. When generative design algorithms are combined with learned chemical intuition, the result is "materials-aware" assistants that can detect data gaps, suggest synthesis pathways, and generate novel material candidates.

Finally, specialized methods for processing complex microstructure data at high throughput are being developed. For example, deep neural networks-specifically, convolutional architectures-are used to quantify morphological variations from materials images, detect defects, and segment microstructural features. What once required painstaking manual analysis can now be automated, as these models can swiftly extract statistically meaningful features from massive microscopy and tomography datasets. Connecting these image-derived features with processing histories and measured properties enables researchers to uncover processing–structure–property relationships more efficiently, speeding up both scientific understanding and real-world applications.

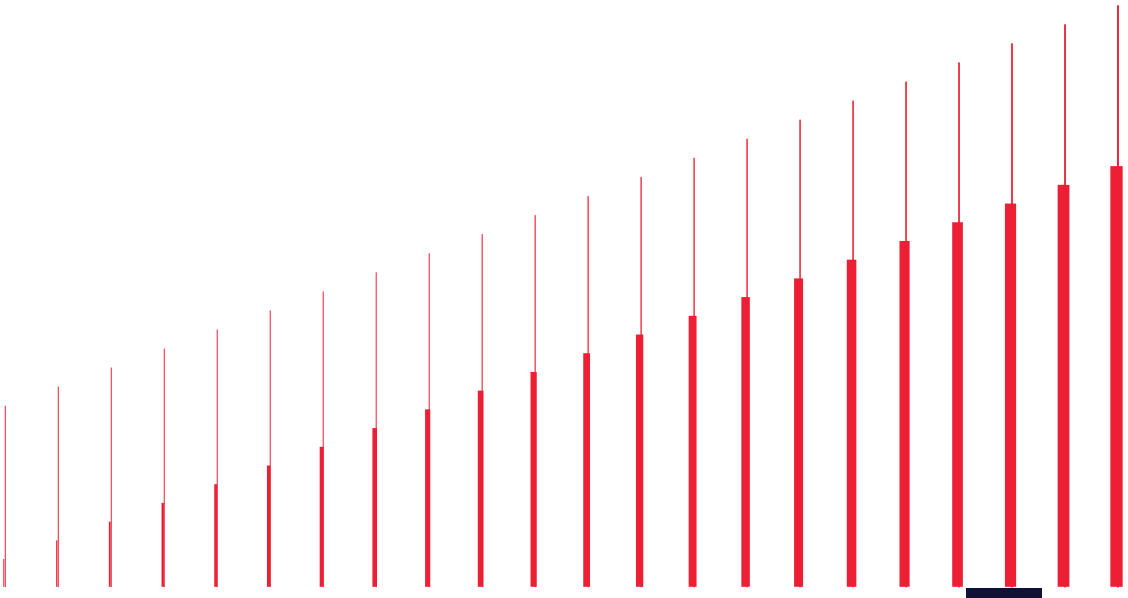
The Role of Materials Fingerprints

The concept of a materials fingerprint is crucial in materials informatics because it transforms raw scientific data into a format that machine learning models can understand and utilize efficiently. The key features of a material-such as its chemical composition, crystallographic structure, local bonding environment, and sometimes even its processing history-are represented numerically by its fingerprint. This conversion from physical reality to numerical data enables algorithms to identify correlations, recognize patterns, and make predictions without directly handling the physical systems. The effectiveness of a model and its ability to provide insights into the underlying science rely on how detailed and relevant these fingerprints are.

Creating a dependable fingerprint is both a computational and scientific challenge. Domain knowledge ensures that descriptors include the most relevant chemical and physical factors influencing a target property. At the same time, algorithmic compatibility requires that fingerprints respect the invariances of the material system, avoiding unnecessary or superfluous features that could mislead the model.

Fingerprints can range from simple scalar values, like average bond lengths or composition ratios, to complex high-dimensional vectors that encode statistical features from microscopy images, crystallographic symmetries, or detailed local atomic environments. Recent advancements enable the direct generation of fingerprints from raw experimental or simulated data. For instance, deep neural networks can automatically extract relevant features from electron microscopy or X-ray diffraction patterns.

Well-designed fingerprints enhance interpretability, reduce the amount of data needed, and improve predictive accuracy by incorporating scientific knowledge directly into the learning process. They act as a bridge between the abstract reasoning of machine learning models and the physical understanding of materials.



Emerging Materials Informatics Ecosystems

Recent progress has enabled interconnected ecosystems where computational tools, experiments, and data analytics operate together, enabling faster and more cost-effective research:

Active Learning for Discovery

Monte Carlo methods, genetic algorithms, and Bayesian optimization guide experiments toward promising candidates, concentrating resources efficiently.

Accelerated Simulation

ML-based surrogate models replace slow physics simulations, providing accurate predictions at lower cost.

Microstructure Analysis

Deep learning of microscopy images enables automatic phase classification and defect detection, improving throughput and reproducibility.

Structure–Property–Processing Correlations

ML uncovers hidden relationships across datasets, refining theoretical understanding and informing industrial design.

Autonomous Laboratories

At the frontier, robotics and AI perform synthesis, processing, and characterization with minimal human input, dramatically increasing scalability and reliability.

Self - Driving Labs
A New Research Paradigm

Self-driving labs, which combine the capabilities of robotics, artificial intelligence, and automated characterization into a single closed-loop system, mark a revolutionary advancement in materials research. Self-driving labs plan, carry out, analyze, and iterate experiments without human assistance, unlike high-throughput labs that merely carry out big batches of pre-programmed experiments. The efficiency of materials discovery and optimization is significantly increased by active learning algorithms that adaptively explore complex experimental spaces, enabling this continuous "experiment-analysis-decision" cycle. These platforms operate around the clock, enhance laboratory safety by minimizing direct handling of hazardous materials, ensure reproducibility by eliminating human error, and, above all, navigate high-dimensional parameter spaces that would be impossible for a human researcher to explore manually.

A striking example comes from the work of Prof. Andrew I. Cooper's group at the University of Liverpool, one of the pioneers in this field. In a groundbreaking study, researchers deployed a mobile robot chemist to autonomously search for improved photocatalysts that can convert water into hydrogen [1]. Using a batched Bayesian search algorithm, the system ran 688 experiments over eight days in a ten-variable space. The robot carried out the entire workflow autonomously—from loading and weighing solids to dispensing liquids, controlling reaction conditions, performing photolysis, and analyzing hydrogen output. Using the same approach, the group identified new solid-state materials [2]. They also integrated multiple robotic platforms to automate crystallization, sample handling, powder X-ray diffraction, and, later, exploratory synthetic chemistry using UPLC-MS and NMR [3]. These examples highlight the versatility of self-driving labs across diverse chemistries and experimental workflows.

The range of self-driving labs is quickly growing beyond just single case studies. Global initiatives have shown their usefulness in improving photovoltaic films [4], creating uniform conductive thin films through spray combustion synthesis [5], making complex polymer blends for stable organic photovoltaics [6], finding new perovskite single crystals [7], changing reaction conditions to get higher yields in different chemical syntheses [8,9], speeding up solid-state synthesis of new inorganic powders [10], and optimizing redox-active materials for flow batteries [11]. The common thread that runs through all these examples is that autonomous systems can learn from data in real-time and focus on the most promising experimental paths. This is something that even the most experienced human researchers can't do as quickly or as broadly.

Self-driving labs can help bridge the gap between scalable manufacturing and lab-scale experimentation [12]. To automatically optimize the processing conditions for electronic polymer films with low defects and high conductivity, the "Polybot" platform was created. Through the integration of stations for liquid handling, solution mixing, blade coating, annealing, in-line conductivity measurements, and imaging, the system determined the ideal fabrication parameters that produced transparent conductive thin films with conductivities exceeding 4500 S/cm. The identified parameters were notably applicable to large-scale production, underscoring the utility of such platforms in industrial settings.

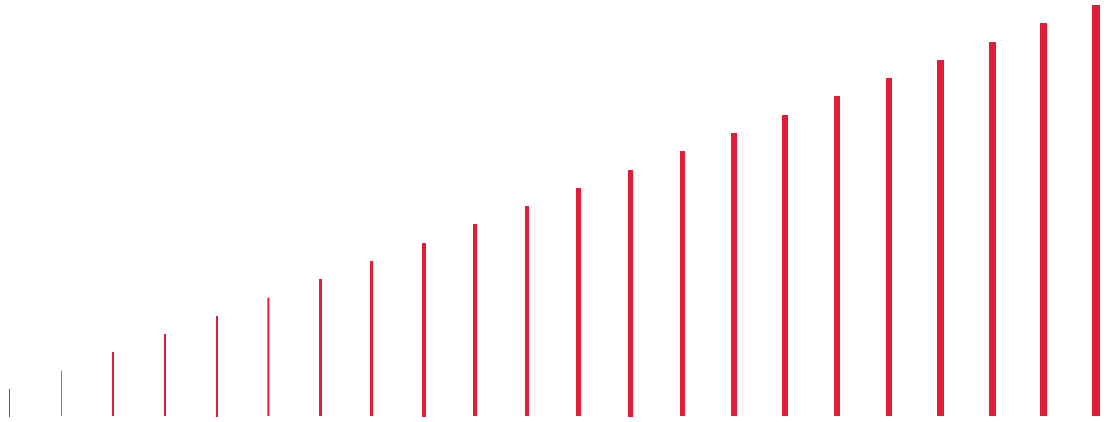
Active Learning for Efficient Discovery

Active learning has emerged as a transformative approach for accelerating materials discovery, particularly in cases where experimental budgets are limited and the design space is vast. Unlike traditional design-of-experiments strategies that select all experiments in advance, active learning iteratively decides on the next experiment based on the outcomes of previous ones, balancing the exploration of under-sampled regions with the exploitation of promising candidates. Interestingly, this approach is especially powerful when integrated with Bayesian optimization, where ML models predict material properties and quantify uncertainty, enabling informed decisions about where to sample next. In materials science, such adaptive experimentation has been instrumental in efficiently navigating high-dimensional parameter spaces to identify materials with optimal properties.

A compelling example comes from the discovery of high-electrostrain perovskite piezoelectric, where only 61 out of 605,000 compositions were initially tested. By starting with a small dataset and employing a group of ML models with different acquisition functions, researchers iteratively refined their search to identify a composition, $(Ba_{0.84}Ca_{0.16})(Ti_{0.90}Zr_{0.07}Sn_{0.03})O_3$, that exhibited a 50% higher electrostrain than the best candidate in the initial dataset. Similar active-learning strategies have been successfully applied to the design of high-strength high-entropy alloys, low-hysteresis shape-memory alloys, and morphotropic-phase-boundary piezoelectric, in each case dramatically reducing the number of required experiments while uncovering superior materials.

Recent work in peptide-based materials discovery provides another striking demonstration of the value of active learning. Researchers used an AI-driven, experiment-informed workflow to identify β-sheet–forming pentapeptides for self-assembling nanostructures. The researchers iteratively trained ML models and selected candidates from a large sequence library, starting from a small dataset. They experimentally validated them, focusing particularly on cases where ML predictions diverged from conventional β-sheet propensity tables. Over three active-learning loops, they synthesized and tested 268 pentapeptides, discovering 96 that formed β sheets, including many unconventional sequences that traditional rules would have overlooked. Interestingly, this strategy not only improved predictive accuracy but also expanded the known chemical space for peptide assembly.

The broader relevance of active learning in materials discovery lies in its ability to simultaneously accelerate exploration and build high-quality datasets for future research. By appending each new validated data point back into the training set, active learning creates a virtuous cycle in which models continuously improve, guiding experiments more effectively. This is especially helpful in materials informatics, where datasets are often sparse and experimental costs are high. In addition, focusing on "areas of disagreement" between different models or between models and domain heuristics can systematically uncover nonintuitive candidates, leading to breakthrough discoveries that would be missed by human intuition alone.



Challenges and the Road Ahead

Despite encouraging developments, several obstacles still exist

Data Quality and Availability

Materials datasets often lack standardization or metadata and omit negative results. Building FAIR (findable, accessible, interoperable, reusable) data pipelines and open repositories remains essential.

Generalization and Transferability

Models trained in specific chemical domains often fail elsewhere. Multitask learning, domain adaptation, and physics-informed models can improve cross-domain robustness.

Integration With Physical Knowledge

Data-only models may violate physical laws; incorporating thermodynamics and symmetry constraints ensures theoretical soundness.

Hardware and Experimental Constraints

Multi-step synthesis, hazardous reagents, and high-temperature chemistries still challenge autonomous labs. Standardized control protocols and adaptive planning are needed.

Interdisciplinary Collaboration

Close coordination among materials scientists, data scientists, roboticists, and chemists is essential for the practical deployment of these technologies.

Tech Mahindra's Role in the Materials Science Process

Here's how Tech Mahindra, through Makers Lab and its AI capabilities, can contribute to the field of material discovery

Platforms for Data Curation and Aggregation

AI depends on high-quality data. Tech Mahindra can build centralized platforms that compile, clean, and organize simulation and experimental data-from lab notebooks, publications, patents, and simulations-into machine-readable repositories. Its knowledge graphs and semantic search expertise will help AI models learn from decades of dispersed research.

AI-Powered Modelling and Simulation

Drawing on Maker's Lab work in AI and quantum computing (Tech Mahindra, n.d.), the company can simulate atomic-level interactions in novel materials using deep learning architectures, such as graph neural networks. Cloud-based virtual testing environments will enable industrial clients to evaluate AI-predicted materials before synthesis, thereby reducing prototyping costs.

Generative AI for Material Design

Generative models, including variational autoencoders and transformers, can design new materials that meet specified performance criteria. Maker's Lab could lead the development of an AI-based "materials design studio," enabling users to input desired properties and obtain candidate materials.

Collaboration with Industry and Academia

Through global partnerships with universities, national labs, and industry, Tech Mahindra can access specialized datasets and domain expertise. Joint initiatives may include co-developing AI standards for materials science or validating predicted materials in academic laboratories.

Tech Mahindra combines technological expertise, research networks, and innovative capacity to drive AI-enabled materials discovery. Investing in research platforms and sustainable practices can shape the next era of materials innovation.

The Way Forward

Automation, data science, and materials science come together powerfully in materials informatics. It can completely change how we find, develop, and utilize new materials to meet future energy and sustainability demands by facilitating data-driven decision-making, accelerating simulation and experimentation, and creating the possibility of self-driving labs. The idea of a future in which AI-guided systems continuously build, test, and improve materials in a closed loop, cutting the time to discovery from decades to months or even days, is intriguing. Better algorithms alone won't be enough to accomplish this; we also need more diverse datasets, more adaptable robotics, and a shared dedication to incorporating AI into science.

Authors

Dr. Rohit Batra

Phd | Assistant Professor | IIT - Madras

Dr. Rohit Batra is a leading expert in materials informatics and computational materials science, with a strong focus on AI-driven materials discovery. He has held research positions at Argonne National Laboratory and Georgia Tech and currently serves as Assistant Professor at IIT Madras. His work spans machine learning, polymer informatics, and autonomous laboratories, with several high-impact publications in top scientific journals.

Saikumar B

Group Delivery Head | Hi-Tech Vertical

Saikumar is a seasoned IT professional with 24+ years of experience in multinational corporations, specializing in delivery, transformation, and leadership. He excels in driving growth, managing portfolios, and achieving exceptional client outcomes. At Tech Mahindra, Saikumar leads global delivery operations for the semiconductor unit, overseeing high-tech client relationships worldwide. He directs high-performing teams, drives innovation, fosters strategic partnerships, and ensures P&L management and client satisfaction.

Dr. Gautam K Rangan

Project Manager | Hi-Tech - D&A

A Scientist, Innovator, entrepreneur, academic and now a delivery professional, Gautam has played diverse roles in the areas of Data Analysis, Marketing and AI. With about 18+ years of experience, he currently serves as the delivery manager for Tech Mahindra's Data & Analysis practice. He has published extensively with ~30 publications in reputed peer reputed and leading practitioner publications.

References

1. Burger, Benjamin, et al. "A mobile robotic chemist." Nature 583.7815, 237 (2020).
2. Lunt, Amy M., et al. "Modular, multi-robot integration of laboratories: an autonomous workflow for solid-state chemistry." Chemical Science 15.7, 2456 (2024).
3. Dai, Tianwei, et al. "Autonomous mobile robots for exploratory synthetic chemistry." Nature 635, 890 (2024).
4. MacLeod, Benjamin P., et al. "Self-driving laboratory for accelerated discovery of thin-film materials." Science Advances 6.20, eaaz8867 (2020).
5. MacLeod, B.P., Parlane, F.G.L., Rupnow, C.C. et al. A self-driving laboratory advances the Pareto front for material properties. Nature Communications 13, 995 (2022).
6. Langner, Stefan, et al. "Beyond ternary OPV: high-throughput experimentation and self-driving laboratories optimize multicomponent systems." Advanced Materials 32.14, 1907801 (2020).
7. Li, Zhi, et al. "Robot-accelerated perovskite investigation and discovery." Chemistry of Materials 32.13, 5650 (2020).
8. Bédard, Anne-Catherine, et al. "Reconfigurable system for automated optimization of diverse chemical reactions." Science 361.6408, 1220 (2018).
9. Coley, Connor W., et al. "A robotic platform for flow synthesis of organic compounds informed by AI planning." Science 365.6453, eaax1566 (2019).
10. Szymanski, Nathan J., et al. "An autonomous laboratory for the accelerated synthesis of novel materials." Nature 624.7990, 86-91 (2023).
11. Hickman, Riley J., et al. "Atlas: a brain for self-driving laboratories." Digital Discovery (2025).
12. Wang, Chengshi, et al. "Autonomous platform for solution processing of electronic polymers." Nature Communications 16.1, 1498 (2025).
13. Computational Drug Discovery (CDD): https://www.techmahindra.com/industries/healthcare-life-sciences/pharma/computational-drug-discovery/

About Tech Mahindra

Tech Mahindra (NSE: TECHM) offers technology consulting and digital solutions to global enterprises across industries, enabling transformative scale at unparalleled speed. With 149,000+ professionals across 90+ countries helping 1100+ clients, Tech Mahindra provides a full spectrum of services including consulting, information technology, enterprise applications, business process services, engineering services, network services, customer experience & design, AI & analytics, and cloud & infrastructure services. It is the first Indian company in the world to have been awarded the Sustainable Markets Initiative's Terra Carta Seal, which recognizes global companies that are actively leading the charge to create a climate and nature-positive future. Tech Mahindra is part of the Mahindra Group, founded in 1945, one of the largest and most admired multinational federation of companies.

*Figures as per Q3, FY26.

TECH mahindra

www.techmahindra.com
www.twitter.com/tech_mahindra
www.linkedin.com/company/tech-mahindra