Artificial Intelligence and Generative Models for Materials Discovery: A Review

Albertus Denny Handoko[1] and Riko I Made[*2]

[1]*Institute of Sustainability for Chemicals, Energy and Environment (ISCE2), Agency for Science, Technology and Research (A*STAR), 1 Pesek Road, Jurong Island, Singapore 627833, Republic of Singapore, handoko_albertus@isce2.a-star.edu.sg, handoko_albertus@a-star.edu.sg*
[2]*Institute of Materials Research and Engineering (IMRE), Agency for Science, Technology and Research (A*STAR), 2 Fusionopolis Way, Innovis #08-03, Singapore 138634, Republic of Singapore, riko@imre.a-star.edu.sg, riko@a-star.edu.sg,*

Abstract

High throughput experimentation tools, machine learning (ML) methods, and open material databases are radically changing the way new materials are discovered. From the experimentally driven approach in the past, we are moving quickly towards the artificial intelligence (AI) driven approach, realising the 'inverse design' capabilities that allow the discovery of new materials given the desired properties. This review aims to discuss different principles of AI-driven generative models that are applicable for materials discovery, including different materials representations available for this purpose. We will also highlight specific applications of generative models in designing new catalysts, semiconductors, polymers, or crystals while addressing challenges such as data scarcity, computational cost, interpretability, synthesizability, and dataset biases. Emerging approaches to overcome limitations and integrate AI with experimental workflows will be discussed, including multimodal models, physics-informed architectures, and closed-loop discovery systems. This review aims to provide insights for researchers aiming to harness AI's transformative potential in accelerating materials discovery for sustainability, healthcare, and energy innovation.

1 Introduction

Materials science is the foundation for technological innovation, driving advances in energy, electronics, catalysis, and quantum computing through the development of novel materials with tailored properties [16,120]. Historically, material discovery is experiment-driven. Often, this means labourious trial and error process where scientists first hypothesize the structures, synthesize compounds, and then test properties [30] (Fig. 1a). While there is nothing fundamentally wrong with this approach [32], the vastness of chemical space, estimated to exceed 10^{60} carbon-based molecules, renders exhaustive experiment-led exploration to find new classes of materials impractical [39,104]. Consequently, the timeline from material conception to deployment often spans decades, hindering innovation and investment [30,122]. Modern technologies such as electric vehicles, high-speed rails, and satellite communications demand new materials with lower weights and enhanced properties, such as high thermal conductivity, electromagnetic shielding, customisable bandgap, or enhanced mechanical strength, pushing the limits of existing compounds [87,138]. This bottleneck spurs the search for a novel approach to the discovery of materials that are capable of navigating complex structural and functional requirements.

The most popular approach in harvesting the low-hanging fruits in the exploration of the vast materials space is what we call the "black-box" approach (Fig. 1b). This is a general term representing intelligent data acquisition strategies that utilize ML optimization algorithms to find the right candidate(s) with targeted properties [27]. A black-box approach typically involves building a new dataset, built specifically for a particular optimization of a desired output (e.g., material properties) based on empirical parameter inputs (e.g., precursor ratio). This approach has been proven effective for well-defined and constrained problems, such as finding the appropriate precursors to optimize the catalytic condition [74], or to optimise process parameters of a reaction [49 80] in which much fewer iteration steps are required compared to traditional methods. However, it is difficult to generalize a black box approach that has been trained in a specific task, unless there are some similarities in properties or structure in the related task [60].

Recognising the limitations of the blackbox approaches to discover truly new materials that can display groundbreaking properties, multi-disciplinary scientists started to apply generative models for materials discovery. Originally designed as AI to simulate human reasoning, intelligence, and creative processes [99], generative

---

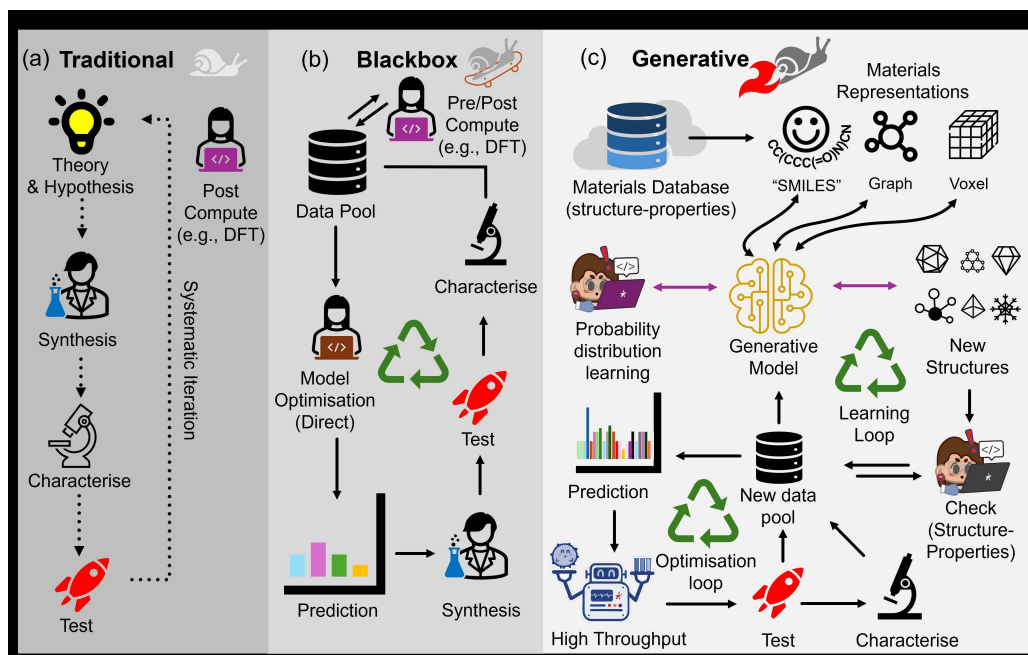*Corresponding author:riko@a-star.edu.sg

Figure 1: A paradigm shift in Materials Discovery. The strategy for discovering new materials has evolved over the last century. (a) Traditionally, materials discovery starts with an idea or hypothesis that needs to be validated through experimental synthesis, characterization, and testing. The advent of computational techniques allows *ab-initio* evaluation of the novel materials to gain deeper insights into the structure-properties relationship or guide subsequent discoveries. (b) Today, we witness many "blackbox" discovery approaches that allow an iterative direct model optimization process. With more accessible and efficient computational resources, computational techniques, including *ab-initio* calculations, can be implemented in the pre-optimisation step, where data pools can be enriched by computational inputs. A second, more thorough *ab-initio* calculation may be performed on the optimum material to validate or further elaborate the material-structure relationship. (c) In the future, we predict a growing "Generative Discovery" approach that leverages existing materials databases rich in past learnings. Ways to encode materials structure into computer-readable representations like "SMILES"[126], graph[125], or voxel[90] allow effective exploration of the materials space. Unlike previous approaches, the generative approach can learn probability distributions, capable of suggesting innovative material structures even before the experiments begin. As the experiment progresses, the continuous learning loop refines the model with fresh data, while predicted structures are meticulously evaluated against target properties, followed by high-throughput synthesis, characterization, and testing. CC(CCC(=O)N)CN is the SMILES representation of 5-amino-4-methylpentanamide

model are gradually finding their way to materials discovery applications. Generative models are not adopted into materials discovery overnight. We recognise at least five key "parents" that brought us AI-driven materials discovery (Fig. 1c).

The first parent is not actually related to AI (or ML), but rather high-throughput combinatorial methods and tools development. Combinatorial methods have been ubiquitous in nature. For example, to obtain suitable antibodies to fight certain pathogens, lymphocytes are assembled in the human body by recombination of large "libraries" of molecules and selecting those with desired properties or mutating them[117]. Such an approach is only relatively recently being applied to materials science[105,106], where large arrays of materials composition are being synthesised (for example, by inkjet[135] or plasma printing[4]) for subsequent systematic testing. Today, coordinated efforts, with many research centres focusing on high throughput experimentation worldwide.

The second parent is the application of ML algorithms for parametric optimisation. The advent of machine learning (ML) has revolutionized materials science by leveraging vast datasets and computational power to uncover intricate patterns and accelerate discovery[16,81,141]. The syntheses of organic and inorganic materials can be complicated and challenging to optimise, as they often involve multiple steps and precursors. Whilst existing statistical optimisation approaches like the design of experiment (DOE) methods have proven to be instrumental in the discovery of new materials[45], the integration of ML can enhance the efficiency and efficacy of these methods further, especially for more complex materials composition or metastable compounds. The synergy between ML, domain expertise, and established DOE methods has been observed to be more robust compared to either standalone approaches[100].

The third is the sharing of materials databases[2]. The mountains of data generated by high-throughput combinatorial methods are only beneficial if it is shared among the wider scientific community. Researchers around the world recognises this, and there has been an exponential increase in curated materials databases in many countries[12,38,57,124]. However, differences in the way labs around the world perform the experiments and record the findings can give rise to dataset mismatch or variation. Developing universally accepted standards is also challenging, as the high-throughput combinatorial methods are still in the infancy with rapidly changing

protocols, tools, and algorithms. We note two major efforts to develop standardised testing and recording computerised materials data have been performed by the Versailles Project on Advanced Materials and Standards (VAMAS, technical work area 10)[3] and ASTM International (Committee E-49)[97], although implementation of these standards in research laboratories remain scarce.

The fourth is the application of ML in the computational modeling of the force field[119]. This is a significant advancement in computational chemistry that bridges between very precise atomic modeling through *ab initio* density functional theory (DFT) calculation and the force field model that drives molecular dynamics (MD). The two computational approaches are on different ends of scale and accuracy: DFT glances from the quantum mechanical point of view, capable of calculating accurate atomic interaction and potential of considered systems, but can only cover a limited (angstroms) range due to the rigorous calculation steps. On the other hand, MD approximates atoms and molecules as particles and uses simpler classical mechanics to solve the dynamic behaviour of a larger number of atoms over a (brief) period. Essential to MD's capability to simulate the dynamic behaviour is the force field model[84], an empirical method to describe the interactions between atoms in the system without the need to model the entire electronic structure or interatomic potential. The implementation of ML methods, especially machine-learned potential (MLP)[43], allows a dream of "hybrid" simulation where accurate potential energies of a larger system (or over a longer period) can be quickly obtained from a suitable numerical representation of the material[29] that can typically be trained with a pool of ab-initio simulation data or experimental data[98].

Last but not least, is the incorporation of generative models into materials discovery. Generative models are able to approximate high-dimensional probability distributions between structures and desired characteristics or properties[9,46,101,103]. Once the probability distributions have been learned, novel data such as molecular structures can be generated by sampling in the probability distributions' (latent) space, based on, for example, the desired properties. The ability of generative models to generate new structure suggestions from the latent space represents a new paradigm of materials discovery. This is a marked departure from previous approaches, where new structure suggestions need to be first explicitly generated in the real space, either by modifying or substituting the atoms found in known structures[51], or by placing completely random atoms within preselected constraints or restraints to ensure that the generated materials are stable and unique[95]. The next few sections of this review will focus on this fifth "parent", describing the different models available for materials discovery and their principles, along with examples and applications of these models in research.

2 Generative Models for Materials Science

To understand the role of generative models in materials discovery, it is essential to distinguish them from supervised machine learning paradigms. Supervised learning focuses on learning a mapping function, $y = f(x)$, to predict outputs y from inputs x using labeled data, minimizing discrepancies between predicted and actual outcomes. Termed *discriminative* models[9], these approaches excel in classification and regression tasks but are limited by their reliance on labeled datasets. In materials discovery, where novel structures and properties are often sought without extensively labeled data, generative models offer a powerful alternative.

Unlike discriminative models, generative models learn the underlying probability distribution, $P(x)$, of the data, enabling the creation of new samples that closely resemble the training set. By capturing the inherent patterns in materials representations (more on this in section 2.2), these models can generate synthetic instances, often in unsupervised settings, leveraging both labeled and unlabeled data. A critical feature of the generative model is the *latent space*: a low(er)-dimensional representation of the structure-properties relationship that enables inverse design strategy.

To understand how inverse design is achieved through generative models, six key types of generative models will be explored. These models are selected for their diverse principles and proven effectiveness in inverse design —generating stable and novel materials for applications like catalysts, electronics, and polymers. We will discuss them in the order of historical emergence and increasing specialization, starting from Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Diffusion Models (e.g., DiffCSP[61], SymmCD[73]), Recurrent Neural Networks (RNNs) and Transformers (e.g., MatterGPT[23], Space Group Informed Transformer[18]), Normalizing Flows (e.g., CrystalFlow[77], FlowLLM[112]), and Generative Flow Networks (GFlowNets, e.g., Crystal-GFN[5])(Fig. 2).

The success of these models depends on effective material representations that preserve structural constraints, atomic interactions, and scalability across small molecules to large crystalline systems. Representations such as sequence-based (e.g., SMILES[126]), graph-based[129], voxel-based, and physics-informed formats[86] enable models to handle complex materials data. By integrating these representations, generative models, from VAEs to GFlowNets, address diverse challenges in materials discovery, offering scalable solutions for crystalline, polymeric, and composite systems (Fig. 6).
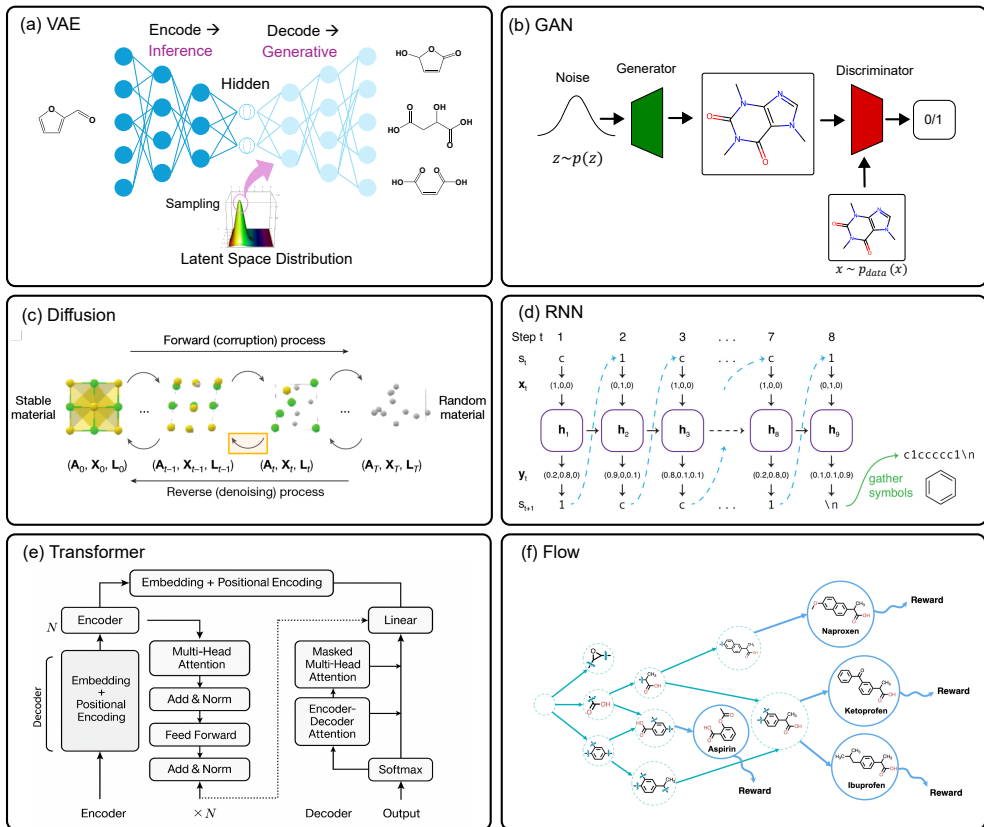
Figure 2: Schematics of generative model architectures for materials discovery, illustrating the general workflows for (a) **VAE (Variational Autoencoder):** encode-decode process with an inference and generative path, showcasing the mapping of molecules into a latent space distribution and subsequent generation of new molecular structures. (b) **GAN (Generative Adversarial Network):** a generator creating molecular structures from noise, which are then evaluated by a discriminator against real molecular data to improve realism. (c) **Diffusion Model:** a forward (corruption) process where stable material is progressively turned into random material, and a reverse (denoising) process that reconstructs the stable material. This is analogous to generating molecules by reversing a corruption process. Taken with permission from [7]. (d) **RNN (Recurrent Neural Network):** A sequential process where hidden states (h_t) are updated based on previous states and current inputs (x_t), eventually leading to the gathering of symbols (e.g., for molecular string generation). Taken with permission from [108]. (e) **Transformer:** encoder-decoder architecture with embedding, positional encoding, multi-head attention mechanisms, and feed-forward layers, commonly used for sequence-to-sequence tasks like molecular string manipulations, (f) **Flow (Reinforcement Learning/Generative Flow):** tree-like structure where molecular syntheses or modifications (e.g., from Aspirin to Naproxen, Ketoprofen, or Ibuprofen) are associated with rewards, suggesting a reinforcement learning approach for optimizing molecular properties or synthesis pathways. Taken with permission from [59].

2.1 Models and Principles

2.1.1 Variational Autoencoders (VAEs)

Variational Autoencoders (VAEs) are generative models that learn a probabilistic latent space for data generation [68]. VAE typically consists of an encoder that maps input data x (e.g., material descriptors) to a latent distribution $q(z|x) = \mathcal{N}(\mu(x), \sigma(x)^2)$, and a decoder that reconstructs x from samples $z \sim q(z|x)$ as $p(x|z)$ (Fig. 2a). The model maximizes the Evidence Lower Bound (ELBO):

$$\text{ELBO} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - \text{KL}(q(z|x)||\mathcal{N}(0, I)), \tag{1}$$

balancing reconstruction accuracy ($\mathbb{E}_{q(z|x)}[\log p(x|z)]$) and regularization of $q(z|x)$ to a Gaussian prior (KL). The reparameterization trick enables back-propagation by sampling $z = \mu(x) + \sigma(x) \cdot \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$, making training more efficient [68]. In materials science, VAEs can be exploited to generate novel structures and optimize properties by sampling the latent space. For example, Gómez-Bombarelli et al. reported the use of VAEs to generate predictions of organic molecules that can be applicable as active pharmaceutical ingredients (API) Gómez-Bombarelli et al. [46] . This is achieved by encoding the molecular structure, through a certain representation like SMILES, into a continuous latent space (Fig. 3), with secondary attributes like properties added via a *predictor* network. One major deficiency of VAEs is related to the way the *encoder* network aims to generate a smooth latent state representation of the input data. While the probabilistic approach allows the model to cover unexplored regions in the input data, this approach also tends to generate "blurry" outputs and
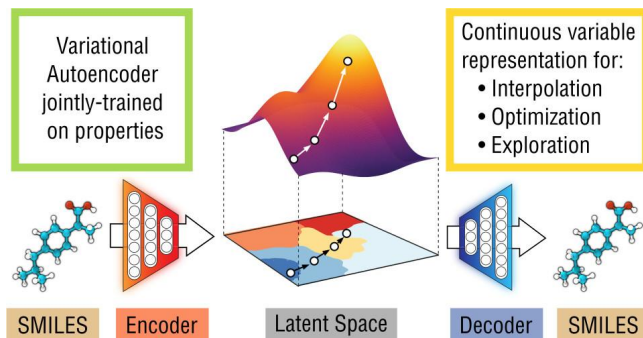
Figure 3: An illustration of the use of VAE for molecular design, integrated with a joint property prediction model —*predictor*. The **encoder** transforms discrete molecular representations (like SMILES strings) into continuous latent space vectors. The **decoder** then converts these latent vectors back into SMILES strings. The *predictor* can be added to predict properties from latent representations. However, the huge size of the latent space dimension (more than 100) makes sampling and visualization difficult. Taken with permission from Gómez-Bombarelli et al. [46]
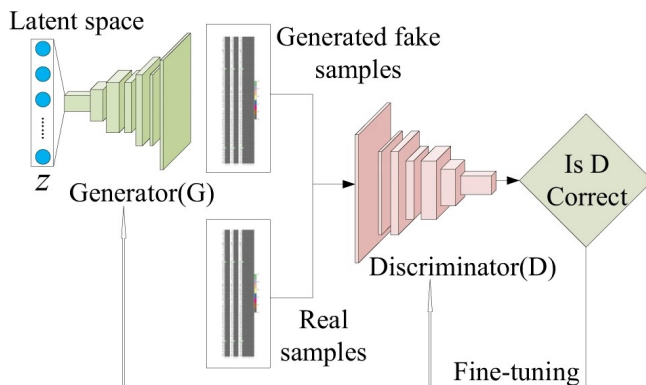

Figure 4: An illustration for MatGAN architecture consisting of a generator, which maps random vectors into generated samples, and a discriminator, which tries to differentiate real materials and generated ones. Taken with permission from Dan et al. [34]

difficulties in capturing complex data distributions [37]. From the materials discovery point of view, this could mean severe difficulty in generating sensible, discrete compounds [48], especially with very sparse initial data (compared to the full materials space). Several improvements have been proposed, to address this limitation, including Binded-VAE, which learns to jointly generate binary vector encoding on the composition and ratio of components [91]. Conditional VAEs (CVAEs) further enable a more targeted design by adding conditions to the generating network [111].

2.1.2 Generative Adversarial Networks (GANs)

GANs, introduced by Goodfellow et al. [47], employ a competitive framework involving a Generator and a Discriminator (Fig. 2b). The Generator produces synthetic data $G(z)$ from noise z, while the Discriminator distinguishes real data x from $G(z)$. This is formalized as a "minimax" game:

$$\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z \sim p_z(z)}[\log(1 - D(G(z)))]. \tag{2}$$

Training stabilizes when the Generator produces data indistinguishable from real data [47]. In materials science, GANs appear to be suitable for exploring vast chemical spaces efficiently. For example, CrystalGAN [89] is successful in generating DFT-validated inorganic crystal structures and identifying new types of stable metal oxides. Similarly, Dan et al. [34] used GANs for the inverse design of inorganic materials, optimizing compositions for specific properties (Fig. 4). Conditional GANs, as explored by Al-Khaylani et al. [6] for nano-photonic metamaterials, enable property-targeted generation, though training instability remains a challenge [8].

2.1.3 Diffusion Models

Diffusion Models generate materials by reversing a noise-adding process, starting from random noise and iteratively refining it into structured data [110] (Fig. 2c). The forward process adds noise to data x_0 over steps t:

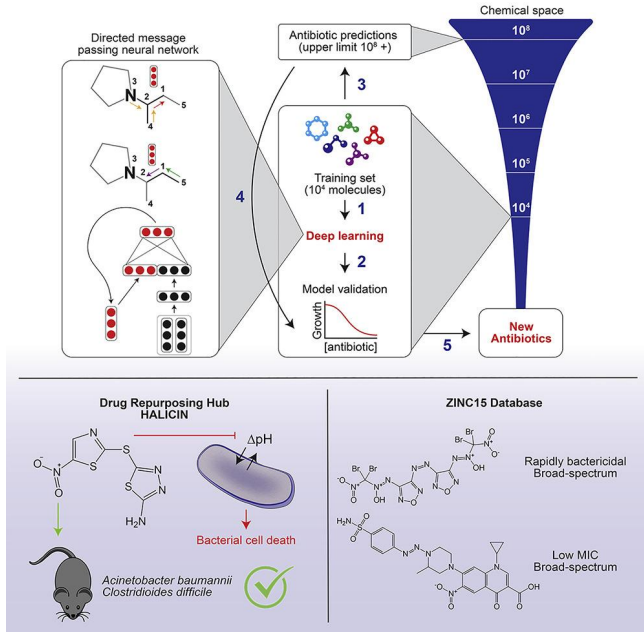$$x_t = \sqrt{1 - \beta_t} x_{t-1} + \sqrt{\beta_t} \epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{3}$$

Figure 5: Stokes et al.[113] workflow for antibiotic discovery using deep learning (RNN) and chemical space exploration. RNN model, trained on 10^4 molecules (1), is validated (2) and subsequently used to predict antibiotic activity across a vast chemical space (up to 10^8 molecules) (3). The molecular representation for deep learning is depicted using a directed message-passing neural network (4). Predicted new antibiotics are then validated (5). The bottom left illustrates **Halicin**, a repurposed drug identified by the model, showing *in vivo* efficacy against bacterial infections. The bottom right displays examples of other potent antibiotic candidates found in the ZINC15 database using this method. Taken with permission from [113].

The model learns to denoise x_T back to x_0, guided by learned patterns[130]. In materials science, MatterGen[137] designs inorganic materials with property-conditioned generation, proposing $TaCr_2O_6$ with a bulk modulus of 169 GPa, which has been experimentally validated. Park et al.[94] used diffusion models to design porous materials, optimizing pore structures for specific applications. Models like DiffCSP[61] and SymmCD[73] generate stable crystals with symmetry constraints, enhancing applicability in electronics and catalysis. Compared to GANs, diffusion models offer improved stability but require significant computational resources[8].

2.1.4 Recurrent Neural Networks (RNNs) and Transformers

RNNs process sequential data by maintaining a hidden state h_t, updated at each time step t:

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h), \quad y_t = W_{hy}h_t + b_y. \tag{4}$$

This recurrence enables RNNs to model chemical sequences, such as SMILES strings[126] (Fig. 2d). For example, Stokes et al.[113] used RNNs to generate novel antibiotics, validated experimentally (Fig. 5). However, RNNs suffer from vanishing gradients, limiting their ability to capture long-term dependencies[52].

Long Short-Term Memory (LSTM) networks address this by introducing gates to regulate information flow, improving sequence modeling[52]. Gómez-Bombarelli et al.[46] used LSTM-based VAEs for molecular design.

Transformers, with attention mechanisms[121], enhance efficiency (Fig. 2e). In materials science, the Wyckoff Transformer[64] generates symmetric crystals, while MatterGPT[23] optimizes multi-property materials. The Space Group Informed Transformer[18] incorporates crystallographic constraints, and CrystalFormer-RL[20] uses reinforcement learning for targeted design. Transformers require extensive data but excel in complex sequence modeling[14].

2.1.5 Flow-Based Models

A flow-based generative model is a generative model used in machine learning that explicitly models a probability distribution by leveraging Normalizing Flow (NF) (Fig. 2f). NF transforms a simple base distribution (e.g., Gaussian) into a complex data distribution via invertible, differentiable mappings[70]. The log-likelihood is computed using the change of variables formula:

$$\log p(x) = \log p_z(z) - \sum_{k=1}^{K} \log \left| \det \frac{\partial f_k}{\partial z_{k-1}} \right|, \tag{5}$$

Table 1: Comparison of generative models used in materials discovery, highlighting their principles, strengths, limitations, and applications.

Model	Principle	Strengths	Limitations	Applications	Ref
VAE	Probabilistic latent space, ELBO optimization	Controlled generation, interpretable latent space	Limited expressiveness, blurry outputs	Molecular design, perovskites, polymers	Das et al.[35], Gómez-Bombarelli et al.[46], Noh et al.[87]
GAN	Adversarial training, minimax game	High-quality outputs, explores vast chemical spaces	Training instability, mode collapse	Crystal structures, metamaterials	Al-Khaylani et al.[6], Dan et al.[34], Nouira et al.[89]
Diffusion	Noise-to-data denoising process	Stable training, high novelty	High computational cost	Inorganic materials, porous materials, crystals	Jiao et al.[61], Levy et al.[73], Park et al.[94], Zeni et al.[137]
RNN, LSTM, Transformer	Sequential processing, attention mechanisms	Effective for sequence data, captures long-range dependencies	Vanishing gradients (RNN), data-intensive	Molecular generation, crystal symmetry, multi-property design	Cao et al.[18], Chen et al.[23], Kazeev et al.[64], Stokes et al.[113]
Normalizing Flows	Invertible mappings, exact likelihood	Exact likelihoods, stable training	High computational cost, discrete structure challenges	Crystal generation, thermal composites	Luo et al.[77], Sriram et al.[112], Wang et al.[123]
GFlowNets	Reward-based sampling, proportional to reward	Diverse sampling, suitable for discrete structures	Computational intensity, task-specific reward design	Crystal sampling, high-throughput screening	AI4Science et al.[5], Bengio et al.[13], Jain et al.[59]

where $z = f^{-1}(x)$ and f_k are bijective functions[93]. NF offers exact likelihoods and stable training, avoiding GANs' mode collapse. In materials science, CrystalFlow generates crystalline structures with high stability[77], while FlowMM uses Riemannian Flow Matching for symmetry-preserving crystal design[82]. FlowLLM leverages large language models for material generation[112], and conditional NF optimises the thermal composite topologies[123]. However, designing expressive transformations is computationally intensive, and NF can struggle with discrete chemical structure representations like SMILES[9].

2.1.6 GFlowNets

Generative Flow Networks (GFlowNets) are designed to sample structured outputs proportionally to a reward function, making them suitable for diverse material generation[13,59]. GFlowNets model a sequential construction process, where a policy $\pi(a|s)$ selects actions a (e.g., adding atoms, modifying bonds) in a state s (e.g., partial material structure) to build complete structures x. The objective is to ensure the probability of generating x is proportional to a reward $R(x)$:

$$P(x) \propto R(x), \qquad (6)$$

where $R(x)$ could represent material stability, bandgap, or other properties. The training losses minimize the discrepancy between forward and backward flow probabilities, ensuring consistent sampling:

$$\mathcal{L} = \sum_{s,a} \left| \log \frac{F(s \rightarrow s')}{F(s' \rightarrow s)} - \log \frac{\pi(a|s)R(s')}{\pi(b|s')} \right|^2, \qquad (7)$$

where $F(s \rightarrow s')$ is the forward flow, and $\pi(b|s')$ represents backward actions[13].

In materials science, Crystal-GFN[5] samples diverse crystal structures with targeted properties, such as stability or a specific bandgap, validated via DFT simulations. GFlowNets excel in high-throughput screening by generating varied candidates, complementing VAEs' latent space sampling and Diffusion Models' denoising. However, GFlowNets are computationally intensive for large state spaces and may be limited to specific tasks due to reward function design[5]. Their ability to model discrete structures makes them promising for inverse design, though scalability remains a challenge compared to NFs or Transformers. As a quick reference, we summarised the different generative models in Table 1.

2.2 Materials Representation

Besides algorithms and databases, the success of generative models in materials discovery also hinges on the way the material identity is encoded into machine-readable formats. Beyond describing the chemical identity, an ideal encoding should capture unique characteristics of the material, including structural, chemical, and physical, and other secondary properties[9]. Sometimes, more is not always better, and effective representations must also balance the information richness with compatibility for specific algorithms/models, and achieve the learning of complex material behaviors at reasonable efficiency. This section explores five key representation types—Sequence, Graph, Voxel, Physics-Informed, and Multi-modal—highlighting their principles, applications, and limitations in materials exploration.
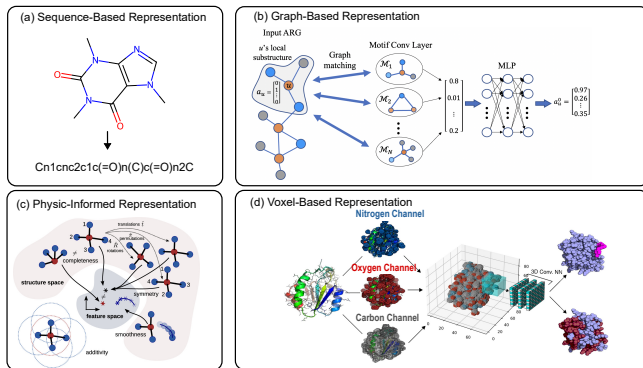
Figure 6: Schematic illustration of four generative model representations for materials. (a) Sequence-based representation encodes material structures as linear strings (e.g., SMILES) for processing by models like RNNs or Transformers. (b) Graph-based representation models atoms and bonds as nodes and edges, leveraging Graph Neural Networks to capture structural relationships. Taken with permission from [125]. (c) Physic-Informed Representation visualizes a conceptual framework where intrinsic physical properties and symmetries (e.g., completeness, symmetry, smoothness, additivity) govern the representation of materials within a "structure space" and "feature space," suggesting a focus on fundamental physical descriptors [86]. (d) Voxel-Based Representation discretizes 3D material structures into voxel grids, suitable for 3D Convolutional Neural Networks. Taken with permission from [90].

2.2.1 Sequence-Based Representation

Sequence-based representations encode materials as linear strings of symbols, making them ideal for molecular structures (Fig. 6a). The Simplified Molecular Input Line Entry System (SMILES) [126] represents molecules as text strings, with atoms (e.g., "C" for carbon), bonds (e.g., "=" for double bonds), branches in parentheses, and rings via numbers. For example, ethanol (CH_3CH_2OH) is written as "CCO". SMILES is widely used with VAEs [46], RNNs [108], and Transformers [53,127], enabling the generation of novel molecules. Whilst simple and attractive, not all materials databases include a SMILES representation for arbitrary compounds [75]. The strict syntax rules adopted by SMILES (e.g., matching parentheses "()" for branches or numbers for rings) are also prone to errors that may precipitate from algorithm exploration. For example, a suggestion containing "CCO(" (a representation for ethanol but missing a closing parenthesis) can turn into nonsensical output that would break the program. Further, SMILES representations are not unique. The same molecule "propane" can be expressed in different ways: "CCC" or "C(C)C", which can lead to confusion and divergence. SMILES also does not check for any physical/chemical rules, allowing strings like "C=C=C=C" that look correct but describe unstable or impossible chemicals. Finally, SMILES representations do not cater for specific 3D arrangement details required to correctly express non-planar molecules or isomers [75]. For example, the expression "C1CCCCC1" generically refers to cyclohexane, but cannot differentiate between the four conformations: chair, twist-boat, boat, or half-chair, each of which has distinct stability and reactivity. SMILES encoding cannot easily capture the complex relationships between compounds, such as during synthesis (or the inverse, decomposition) [41].

These issues prompted the development of Self-referencing Embedded Strings (SELFIES) [71], which use tokens (e.g., "[C]", "[=O]") to ensure every string corresponds to a valid molecule. For instance, "[C][C][O]" reliably represents ethanol. Despite its robustness, SELFIES struggles with 3D conformations and large macromolecules, limiting its applicability to complex materials [71]. Further use of reinforcement learning to construct viable materials via sequential addition (or deletions) of components has also been attempted [63].

2.2.2 Graph-Based Representation

Graph-based representations model materials as graphs $G = (V, E)$, where nodes V represent atoms and edges E denote bonds (Fig. 6b). Node features (e.g., atomic number) and edge weights (e.g., bond strength) capture chemical connectivity, making this approach versatile for molecules and crystals [21]. Graph Neural Networks (GNNs) process these graphs via message passing:

$$h_i^{l+1} = \text{UPDATE}\left(h_i^{(l)}, \text{AGGREGATE}(\{h_j^{(l)} \mid j \in N(i)\})\right), \tag{8}$$

where $h_i^{(l)}$ is the feature vector of node i at layer l, and $N(i)$ is its neighbors [129]. GNNs, such as SchNet [107] and MEGNet [21] (Fig. 7), predict properties like bandgaps or generate structures via VAEs [109] and GANs [17]. The GNoME project [81] used GNNs to discover 380,000 stable crystals, leveraging graph representations for efficient stability predictions.

Despite their power, graph representations oversimplify long-range interactions (e.g., van der Waals forces between 2D material layers) and struggle with scalability for large or amorphous systems. Bond type ambiguity
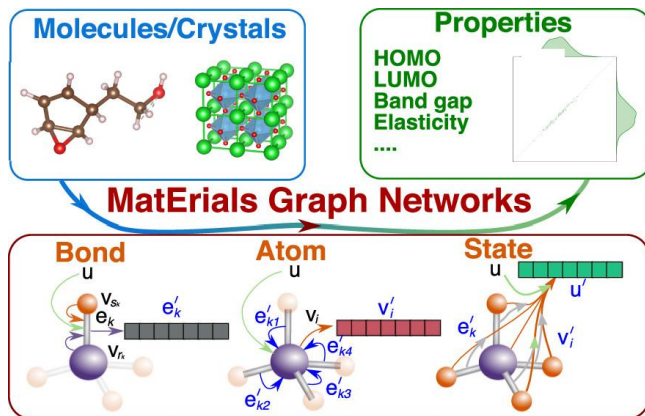
Figure 7: An illustration of the MEGNet's architecture, designed to predict properties of molecules and crystals (top panel). The central concept involves iterative message passing, where information is first exchanged and updated across **Bonds** (e'_k), then aggregated at individual **Atom** nodes (v'_i) from their connected bonds, and finally consolidated into a global **State** representation (u') for the entire material (bottom panel). Taken with permission from [21].

(e.g., covalent vs. ionic in ZnO) and loss of 3D geometry often require external validation, such as DFT, limiting their ability to capture dynamic material properties [81].

2.2.3 Voxel-Based Representation

Voxel-based representation discretises a material's 3D unit cell into a grid of voxels, each storing attributes like atomic occupancy or element type (Fig. 6d). Analogous to 3D pixels, this approach captures spatial arrangements, making it compatible with convolutional neural networks (CNNs) [16]. Voxel grids enable generative models to learn local atomic patterns and symmetries, facilitating the design of complex structures. For example, MatterGen [137] likely employs voxel-like discretizations to generate inorganic materials, optimizing properties like bulk modulus for compounds like $TaCr_2O_6$ [136].

Voxel representations excel in capturing 3D geometry but face challenges with computational cost, as high-resolution grids demand significant memory. They also struggle with periodic boundary conditions in crystals and may oversimplify atomic interactions, requiring careful pre-processing to ensure accuracy [31].

2.2.4 Physics-Based Representation

Physics-based representation seeks to integrate physics-driven information into the learning process, embedding physical laws to produce realistic outputs that adhere to fundamental principles (see Fig. 6c). This approach enhances the model's ability to generate materials that respect constraints such as symmetry or conservation laws. In practice, it is often paired with other techniques, that is multi-modal representation. A common method involves combining a physics-based penalty with the data-fitting term, expressed as:

$$L_{\text{total}} = L_{\text{data}} + \lambda L_{\text{physics}}, \tag{9}$$

where L_{physics} enforces constraints like symmetry or conservation laws [131]. For instance, Xie et al. [128] used a symmetry penalty to ensure crystallographic consistency, while Zhu et al. [139] incorporated heat transfer laws for additive manufacturing. Physics-informed approaches generate stable materials, as shown by Fuhr and Sumpter [44], who added energy minimization terms.

However, these methods require detailed prior knowledge, which may not generalize to novel materials. The computational cost of calculating physics-based residuals and the challenge of tuning λ can limit scalability and diversity, potentially biasing outputs toward known physical regimes [44,86].

2.2.5 Multi-Modal Representation

Thus far, the material representations explored have relied on a single modality to characterize a material's properties. We have also observed that each of these representations comes with inherent limitations that can compromise prediction accuracy and practical utility. Recently, the emergence of "multi-modal" representation—a technique that integrates multiple representations to create a more holistic and potentially more precise description of a material—has gained traction [36]. Beyond inorganic materials, this approach has found significant application in polymers [50] (Fig. 8). This is probably because of the difficulty for any single representation to describe polymers with wide variation in chain length, functional moieties, or isotacticity. The modality of representation is not restricted to just the molecular structure. For instance, Trask et al. [118] combined electron
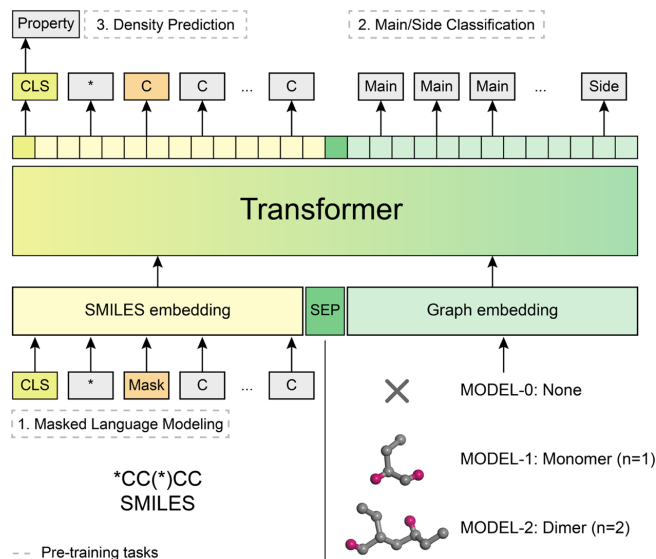
Figure 8: How multi-modal material representation is addressing the limitation of single representation. In this example, a Transformer model is used to integrate SMILES embeddings and graph embeddings of molecular structures for pre-training and downstream tasks. Input SMILES strings are processed for Masked Language Modeling (Task 1), while molecular graph embeddings (e.g., monomer and dimer structures) are used for Main/Side Classification (Task 2). The Transformer's outputs are then utilized for Property Prediction and Density Prediction. Taken with permission from [50]

.

Table 2: Comparison of materials representation types for generative models, highlighting their principles, strengths, limitations, and applications in materials discovery.

Repr	Principle	Strengths	Limitations	Applications	Key References
Sequence-Based	Linear strings (e.g., SMILES, SELFIES)	Simple, compact, compatible with RNNs, Transformers	Lacks 3D details, fragile syntax (SMILES)	Molecular design, antibiotics	Krenn et al.[71], Weininger[126]
Graph-Based	Graphs $G = (V, E)$ with nodes (atoms), edges (bonds)	Captures connectivity, scalable with GNNs	Misses long-range forces, 3D geometry	Crystals, molecules, battery materials	Merchant et al.[81], Xie and Grossman[129]
Voxel-Based	3D grid of voxels encoding atomic properties	Captures 3D geometry, compatible with CNNs	High computational cost, periodic boundary issues	Inorganic materials, porous structures	Cunningham et al.[31], Zeni et al.[137]
Multi-Modal	Combination of Representations	Able to learn implicit properties, Improves accuracy	Complex, requires multiple encodings, prior knowledge required	Materials generation, recognition	Das et al.[36], Trask et al.[118]

micrographs and XRD relative intensity data to the usual materials structure identifier, allowing improved prediction of residual stress compared to the single-mode identifier. Simple inclusion of application-related "tokens" (text description of materials usage or characteristics obtained from free text search) can enhance the materials properties prediction. For instance, Huang et al.[54] demonstrated that the inclusion of a text description "adhesive" to the polymer application can better predict the glass transition temperature. We believe the usage of multi-modal representation will continue to grow, both in academic and applied research, as it is proven to deliver better recognition and prediction for materials discovery.

3   Applications of Generative Models in Materials Design

Generative models in conjunction with advanced materials representation have been exploiting large and diverse datasets, such as the Inorganic Crystal Structure Database (ICSD)[134], Open Quantum Materials Database (OQMD)[102], Materials Project[57], and PubChem[67], to explore vast chemical and structural spaces, predict material properties, and optimize candidates for applications in energy storage, catalysis, electronics, biomaterials, and high-throughput screening. Their combinations have been widely claimed to accelerate materials discovery, often in closed-loop systems with experimental validation. This section reviews key applications, highlighting specific examples, methodologies, and their impact, drawing on recent literature and datasets to provide a comprehensive overview[9,44].
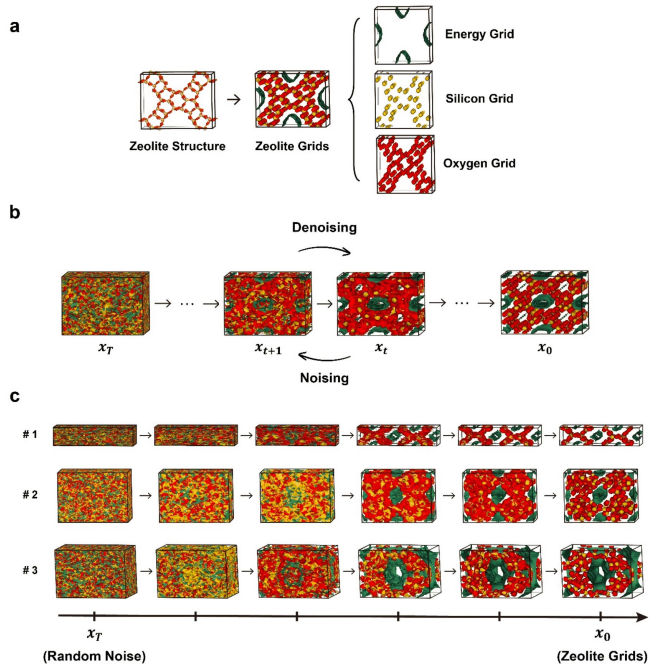
Figure 9: An illustration of zeolite generation via Diffusion model. (a) input structure representation, (b) the noising and denoising phases, and (c) the progressive sampling of zeolite grids. Taken with permission from [94]

3.1   Energy Storage and Battery Materials

Generative models have revolutionized the design of materials for energy storage, particularly for lithium-ion batteries, solid-state electrolytes, and hydrogen storage systems, by generating candidates with optimized electrochemical properties. For solid-state electrolytes, VAEs have been employed to design materials with high ionic conductivity. For instance,[122] utilized a VAE to generate graph-based representations of garnet-type electrolytes, trained on ICSD data, proposing candidates with 15% higher conductivity, validated through density functional theory (DFT) simulations.

In electrode material design, GANs have been used to discover novel cathode materials.[8] employed a GAN to generate perovskite-based cathodes, training on OQMD and Materials Project datasets, producing candidates with 10% higher capacities than $LiCoO_2$, some of which were synthesized experimentally. Additionally, MolGAN[17] was adapted to generate molecular graphs for organic electrode materials, enhancing energy density predictions.

For hydrogen storage, diffusion models are emerging as powerful tools for designing porous materials like metal-organic frameworks (MOFs).[94] applied a diffusion model to generate voxel-based MOFs, trained on QMOF and ZINC datasets, achieving a 20% improvement in hydrogen storage capacity, with potential for experimental validation (Fig. 9). Similarly, Zeni et al.[137]'s diffusion-based MatterGen model trained on Materials Project data to design sulphide electrolytes has achieved improved stability and conductivity for all-solid-state batteries.[7] extended diffusion models to porous carbon materials, optimizing pore structures for hydrogen uptake, validated via Monte Carlo simulations[61] and SymmCD[73] generate stable crystalline electrolytes using fractional coordinates and symmetry-preserving diffusion, trained on Materials Project data, enhancing ionic conductivity.

RNNs, particularly the Long Short-Term Memory (LSTM) networks, have been utilized for polymer electrolyte design.[108] employed an RNN to generate sequence-based polymer chains, trained on a polymer property dataset, resulting in flexible electrolytes with 12% higher conductivity, suitable for wearable batteries. Transformers, such as MatterGPT[23], optimize multi-property electrode materials, trained on Materials Project data.

Normalizing Flows generate stable crystalline electrolytes, as demonstrated by Luo et al.[77], who used CrystalFlow to produce structures with high ionic conductivity, trained on Materials Project data, suitable for battery applications.

3.2   Catalysis and Chemical Conversion

Generative models accelerate catalyst discovery for critical reactions, such as CO_2 reduction, water splitting, and ammonia synthesis, by predicting optimal compositions and surface structures, leveraging datasets like NOMAD and Catalysis-Hub.

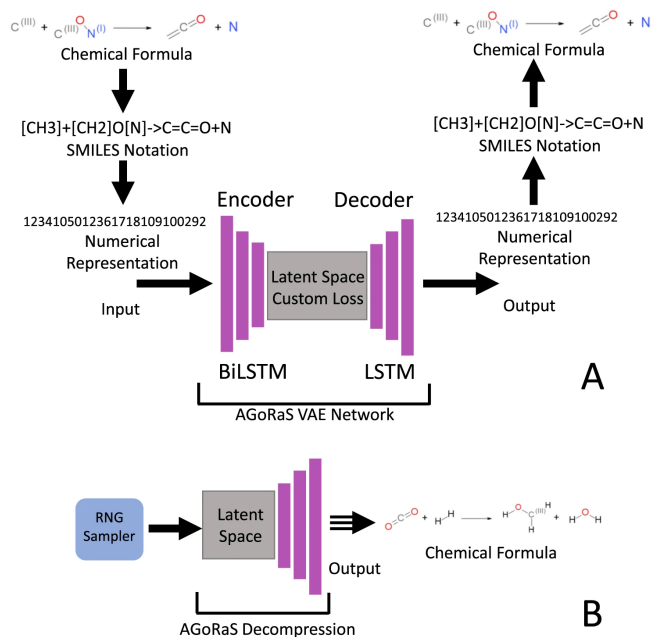**Figure 10:** An illustration of MOFformer's architecture and self-supervised training. (a) MO-Former processes Metal-Organic Frameworks (MOFs) by taking their unique MOFiD (e.g., qmof-2521623) as input. This MOFiD is tokenised, embedded with positional encoding, and then fed through multiple Transformer encoder layers. The final embedding of the first token is used by an MLP regression head for property prediction. (b) Each Transformer encoder layer consists of a multi-head scaled dot-product attention mechanism followed by an MLP, with residual connections and layer normalization applied after both. (c) A self-supervised framework utilizes both CGCNN (on 3D structures) and MOFormer (on MOFiD sequences) to generate embeddings (ZA and ZB) for the same MOF. An MLP head projects these representations. A Barlow Twins loss function then optimizes the cross-correlation matrix of these embeddings to resemble an identity matrix, thereby enabling robust representation learning. Taken from [19] (CC-BY-4.0)

GANs have been used to design high-entropy alloy catalysts, with [56] employing a GAN trained on DFT-calculated adsorption energies from NOMAD to generate heterogeneous catalysts for CO oxidation, achieving enhanced catalytic activity and validated through first-principles microkinetics [56]. CrystalGAN was applied to generate crystallographic alloy structures, improving catalytic stability for methanol oxidation [89].

VAEs have proven effective for designing catalytic reaction pathways, as demonstrated by [116], who used a VAE to generate novel chemical reaction mechanisms, trained on a reaction dataset, achieving optimized pathways for catalytic processes with reduced computational cost [116]11 . GraphVAE was adapted to optimize active site configurations, validated experimentally [109].

Diffusion models have excelled in catalyst surface design, with Alverson et al. [8] using a diffusion model to generate voxel-based representations of nitrogen reduction catalyst surfaces, trained on Catalysis-Hub data, achieving 15% higher ammonia synthesis efficiency [8]. Yong et al. [133] applied diffusion models to disordered catalytic interfaces, improving prediction accuracy for CO_2 conversion [133]. DiffCSP [61] and SymmCD [73] generate alloy catalysts with precise symmetries, trained on NOMAD data

RNNs have been employed for sequence-based catalyst design, with Honda et al. [53] using a SMILES Transformer, an RNN variant, to generate ligand sequences for homogeneous catalysts, trained on a ChEMBL dataset, reducing experimental iterations by 40% for olefin metathesis. Transformers, such as MOFormer (Fig. 10) [19], CrystalFormer-RL [20], optimises catalysts through reinforcement learning, trained on Catalysis-Hub data. Normalizing Flows generate stable crystalline electrolytes, as demonstrated by Luo et al. [77], who used CrystalFlow to produce structures with high ionic conductivity, trained on Materials Project data, suitable for battery applications.

3.3   Electronic and Photonic Materials

Generative models are pivotal in designing materials for electronics, photonics, and optoelectronics, where precise control over electronic and optical properties is critical, leveraging datasets like Materials Project [57] and AFLOW [33].

VAEs have been utilized to design semiconductors with tailored bandgaps, as shown by Gómez-Bombarelli et al. [46], who applied a VAE to generate sequence-based halide perovskites, trained on Materials Project band structure data, achieving 25% efficiency in tandem solar cells. Noh et al. [87] used a VAE for the inverse design of semiconductors, proposing candidates with optimized optoelectronic properties.

GANs have excelled in designing metamaterials, with Al-Khaylani et al. [6] using a GAN to generate nano-photonic metamaterials, trained on simulated optical datasets, achieving 30% improved light-trapping efficiency

Figure 11: Workflow of the AGoRaS-based VAE network. (a) chemical database information is compressed and decompressed to form a high-dimensional latent space. (b) Training and sampling of latent space to generate new compounds. Taken from [116] (CC-BY)

for photonic devices. Yeung et al. [132] employed GANs for a global inverse design of photonic structures, including metasurfaces and photonic crystals, optimizing optical responses across multiple structure classes. Lai et al. [72] applied conditional Wasserstein GANs to design acoustic metamaterials, demonstrating transferability to photonic applications.

RNNs have been effective for designing new 2D material, with Krenn et al. [71] adapting SELFIES representations with RNNs to design stable 2D materials, validated via DFT. Transformers, such as the Space Group Informed Transformer [18], generate symmetric crystals for optoelectronic devices, trained on Materials Project data.

Diffusion models, including DiffCSP [61] and SymmCD [73] are emerging for photonic nanostructures, with Yong et al. [133] applying diffusion models to disordered photonic interfaces, trained on Materials Project optical data, enhancing design robustness for optoelectronic devices.

Normalizing Flows, via CrystalFlow [77] and FlowMM [82], generate crystalline semiconductors with precise bandgaps, trained on Materials Project data, for optoelectronic applications.

3.4 Biomaterials and Drug Delivery

Generative models are increasingly applied to design biomaterials for drug delivery, tissue engineering, and biocompatible coatings, leveraging physics-informed representations and datasets like PubChem.

VAEs have been reviewed for their role in polymer design, with Anstine and Isayev [9] highlighting their ability to optimize biocompatibility for drug delivery applications. Diffusion models have proven effective for 3D scaffold design, with Alakhdar et al. [7] using a diffusion model to generate voxel-based collagen scaffolds, trained on a biomaterial dataset, achieving 15% higher cell viability. Cunningham et al. [31] applied generative models to optimize scaffold porosity for tissue regeneration.

Stokes et al. [113] adapted GANs for antibiotic-inspired coatings, enhancing bacterial inhibition. RNNs have been employed for sequence-based biomaterial design, with Winter et al. [127] using an RNN with SMILES representations to generate peptide sequences for tissue regeneration, achieving 20% enhanced cell adhesion in vitro. Normalizing Flows are less common here due to challenges with discrete structures like peptides, but FlowLLM [112] shows promise for generating biocompatible polymer sequences, trained on PubChem data.

3.5 High-Throughput Screening and Inverse Design

Generative models enable high-throughput screening [62] and inverse design, leveraging datasets like ICSD and OQMD to generate and filter large material libraries. VAEs have been used for inverse design, with Noh et al. [87] employing a VAE for inverse design of solid-state materials, trained on ICSD data, proposing thermoelectric materials with 10% higher figure-of-merit (ZT) values. Dan et al. [34] utilized GANs for inverse design of inorganic materials, generating stable perovskites and garnets, validated via DFT, reducing computational costs by 30%.

Table 3: Summary of Generative Models, Datasets, and Applications in Materials Design

Model	Repr	Dataset	App	Example	Ref
VAE	Graph-based	ICSD	Electrolytes	Garnet-type electrolytes with 15% higher conductivity	Vasylenko et al. [122]
GAN	Graph-based	NOMAD	Catalysts	Heterogeneous catalysts for CO oxidation	Ishikawa [56]
GAN	Voxel-based	Simulated Optical	Photonics	Nano-photonic metamaterials with 30% improved light-trapping	Al-Khaylani et al. [6]
Diffusion	Voxel-based	QMOF, ZINC	Hydrogen Storage	MOFs with 20% higher storage capacity	Park et al. [94]
Diffusion	Voxel-based	Biomaterial	Scaffolds	Collagen scaffolds with 15% higher cell viability	Alakhdar et al. [7]
Diffusion	Fractional	Materials Project	Electrolytes, Catalysts	Crystalline electrolytes and semiconductors	Jiao et al. [61], Levy et al. [73]
RNN	Sequence-based	PubChem	Biomaterials	Peptides with 20% enhanced cell adhesion	Winter et al. [127]
Transformer	Graph-based	Materials Project, OQMD	Screening, Electronics	Multi-property material libraries, symmetric crystals	Cao et al. [18], Chen et al. [23]
Normalizing Flows	Voxel-based	Materials Project	Electrolytes, Catalysts	Crystalline electrolytes and alloy catalysts	Luo et al. [77]
Normalizing Flows	Graph-based	PubChem	Biomaterials	Biocompatible polymer sequences	Sriram et al. [112]

Table 4: Comparison of Generative Models in Materials Design

Model	Strengths	Weaknesses	Datasets	App	Compute Cost
VAE	Stable training, meaningful latent space	Blurry outputs, limited fidelity	ICSD, Materials Project, PubChem	Electrolytes, polymers, semiconductors	Moderate
GAN	High-fidelity outputs, realistic structures	Mode collapse, training instability	NOMAD, AFLOW, OQMD	Catalysts, photonic materials, coatings	High
Diffusion	High-quality outputs, diverse generation, symmetry preservation	High computational cost, data dependency	CoRE-MOF, Catalysis-Hub, Materials Project	Hydrogen storage, scaffolds, crystals	Very High
RNN	Effective for sequential data, memory retention	Limited to sequence-based tasks	ChEMBL, PubChem, 2D Materials	2D materials, peptides, ligands	Low to Moderate
Transformer	Handles large datasets, high accuracy	Requires extensive training data	ICSD, OQMD, Materials Project	High-throughput screening, inverse design	High
Normalizing Flows	Exact likelihoods, stable training	High computational cost, discrete structure challenges	ICSD, Materials Project, PubChem	Electrolytes, catalysts, semiconductors	High

Sanchez-Lengeling and Aspuru-Guzik [103] reviewed GAN-based inverse design, highlighting applications in MOF design for gas separation.

Diffusion models have facilitated high-throughput screening, with Jiao et al. [61], Levy et al. [73], Merchant et al. [81] scaling deep learning to generate diverse material libraries using Materials Project data. Baird et al. [10], Hautier et al. [51], Lu et al. [76], Ren et al. [96] demonstrated AI-driven high-throughput library generation, which supports rapid material discovery across applications. Transformers, such as MatterGPT [23], enable multiproperty inverse design, generating diverse material libraries for high-throughput screening, trained on OQMD data. Normalizing Flows, via FlowLLM [112], generate alloy and ceramic libraries with tailored properties, trained on Materials Project data, enhancing high-throughput screening efficiency.

3.6 Integration with Experimental Workflows

Generative models are increasingly integrated with automated experimental platforms [26,78], creating closed-loop discovery systems that combine prediction, synthesis, and characterization [42,79], as reviewed by Correa-Baena et al. [30]. VAEs have been used in active learning frameworks, with Zuo et al. [141] using a VAE with Bayesian optimization, trained on OQMD data, to prioritize candidates for shape-memory alloys, reducing experiments by 50%. Butler et al. [16] reviewed active learning in material discovery, focusing on VAE applications. RNNs and VAEs are integral to self-driving labs, with VAE and RNN use in autonomous chemistry platforms, trained on ChEMBL and PubChem data, optimizing catalyst synthesis. Transformers, such as CrystalFormer-RL [20], support automated workflows by optimizing material designs for synthesis. Musil et al. [86] highlighted physics-informed representations in self-driving labs for biomaterial design. Normalizing Flows, via conditional NFs [123], integrate with active learning to optimize thermal composites in closed-loop systems, trained in experimental data sets, streamline synthesis and characterization.

4 Challenges and Limitations in AI-Driven Materials Discovery

Despite the remarkable progress in applying AI and generative models to materials discovery, several challenges and limitations still need to be addressed to facilitate their widespread and effective adoption. Issues related to data quality and availability, model interpretability, computational cost, generalization, and integration with experimental workflows pose barriers to achieving robust, scalable, and reliable AI-driven materials discovery. Ethical considerations, including bias in datasets and environmental impacts of computational resources, further complicate their deployment. This section examines these challenges, their implications for applications like energy storage, catalysis, and biomaterials, and potential strategies to address them, drawing on recent literature and insights from datasets like the Inorganic Crystal Structure Database (ICSD) and Materials Project[9,44].

4.1 Data Quality and Availability

As elaborated in the opening section of this review, the success of generative models in materials design relies heavily on the availability and quality of the training dataset. More often than not, these datasets are generated from diverse commercial entities or academic institutions across the world that may perform experiments differently depending on their trainings. Whilst standards (such as those developed by ASTM or VAMAS) work well in industrial, high-volume production settings, persuading researchers to perform tasks and record them in a *specific* way is more challenging than imagined. Quoting an apt editorial piece from the npj computational material, "It is increasingly difficult to identify individuals who are qualified to comment on all aspects of the latest research papers."[15]. Egos aside, we think most scientists would agree that the fundamental aspects of good reporting such as clear descriptions of models, open data availability (except in specific cases requiring subject anonymity and safety concerns), and training procedures are required. These data-sharing practices have contributed significantly to the formation of systematic databases such as the ICSD, Materials Project, and PubChem. Still, as they are a product of evolving science over the years or even decades, many of these datasets contain incomplete or noisy entries with limited chemical diversity. These issues can create known/unknown biases toward well-studied materials, which can restrict the models' ability to explore novel chemical spaces. For instance, Vasylenko et al.[122] noted that ICSD's focus on crystalline structures limits VAE applications for amorphous materials like polymer electrolytes. Transformers, such as MatterGPT[23], require extensive pre-training data, exacerbating issues with dataset scarcity. Normalizing Flows, like CrystalFlow[77], are sensitive to noisy data, affecting likelihood-based training. Small dataset sizes, particularly for specialized applications like biomaterials, exacerbate overfitting risks and reduce the reliability and applicability of the model. Data curation challenges, such as inconsistent property measurements across sources, further complicate training, as highlighted by Butler et al.[16] in their review of ML use in materials science. Strategies to address these issues include synthetic data generation using diffusion models (e.g., DiffCSP[61], SymmCD[73]), federated learning to combine proprietary datasets, and dataset expansion efforts like OQMD[141]. We recognise recent efforts for amorphous materials screening through experimental[55] and computational MD simulation[114]. However, we believe a more concerted effort is needed to fill persistent gaps in underrepresented material classes.

4.2 Model Interpretability and Generalization

Generative models often lack interpretability, complicating the understanding of latent representations and material property relationships, which hinders trust in applications like catalysis[62], electronics, and photonics. GANs produce high-fidelity outputs but suffer from mode collapse, generating limited structure subsets, as noted by Al-Khaylani et al.[6] in nano-photonic metamaterial design. VAEs offer interpretable latent spaces but generate blurry structures, limiting precision in semiconductor design[46]. Diffusion Models, such as DiffCSP[61] and SymmCD[73], use complex symmetry-preserving mechanisms (e.g., fractional coordinates, asymmetric units), making interpretation challenging. Transformers, like the Space Group Informed Transformer[18], rely on intricate attention mechanisms, requiring advanced XAI techniques like attention visualization[21]. Normalizing Flows, such as FlowLLM[112], provide exact likelihoods but struggle with discrete structures, limiting interpretability for polymers. Generalization across diverse chemical spaces is another challenge. Models trained on specific datasets (e.g., NOMAD for catalysts, Materials Project for crystals) often fail to transfer to unrelated applications like biomaterials. DiffCSP and SymmCD are limited to crystalline systems, while GFlowNets like Crystal-GFN[5] focus on specific sampling tasks[47]. Physics-informed models, embedding symmetry or thermodynamic constraints, improve interpretability and transferability[86]. Explainable AI frameworks, such as attention-based visualization for Transformers, are critical for closed-loop discovery systems.

4.3 Computational Cost and Scalability

The computational cost of training generative models poses a significant barrier to scalability. Diffusion Models, used for hydrogen storage, scaffolds, and crystals (e.g., DiffCSP[61], SymmCD[73], WyckoffDiff[65]), require

Table 5: Summary of Challenges and Potential Solutions in AI-Driven Materials Design

Challenge	Description	Impact	Potential Solutions	Reference
Data Quality and Availability	Incomplete, noisy, biased datasets; limited diversity	Overfitting, restricted chemical exploration	Synthetic data, federated learning, dataset expansion	Butler et al. [16], Jiao et al. [61]
Model Interpretability	Black-box models; complex mechanisms in Diffusion, Transformers, NFs	Limited trust in high-precision tasks	Physics-informed models, XAI (e.g., attention visualization)	Chen et al. [21], Musil et al. [86]
Computational Cost	High resource demands for Diffusion, Transformers, NFs	Inaccessibility, environmental impact	Model compression, efficient architectures, cloud computing	Luo et al. [77], Merchant et al. [81]
Generalization	Poor transferability across chemical spaces; task-specific models	Failure in diverse applications	Transfer learning, domain adaptation, physics constraints	Cao et al. [18], Goodfellow et al. [47]
Experimental Integration	Discrepancies between predictions and experiments	Reduced reliability in closed-loop systems	Robust feedback loops, standardized protocols	Correa-Baena et al. [30], Wang et al. [123]
Ethical Concerns	Dataset biases, environmental impact, misuse risks	Skewed predictions, societal harm	Transparent reporting, efficient models, responsible AI	Chen et al. [23], Coley et al. [28]

extensive resources due to iterative denoising steps [7,94]. GANs are computationally intensive and unstable, demanding significant GPU resources [6]. Transformers, such as MatterGPT [23], involve large-scale pretraining, increasing computational demands. Normalizing Flows, like CrystalFlow [77], incur high training costs due to invertible transformations. GFlowNets, such as Crystal-GFN [5], have moderate costs but require optimization for scalability. These demands limit accessibility for smaller research groups and raise environmental concerns due to carbon footprints [81]. Strategies to mitigate these issues include model compression, efficient architectures (e.g., lightweight VAEs, optimized NFs), and cloud-based computing platforms. Advances in hardware, such as AI accelerators, and frameworks like PyTorch are reducing barriers, but computational costs remain a bottleneck for large-scale materials discovery.

4.4 Experimental Workflow Integration, Environmental, and Ethical Considerations

Integration with experimental workflows is another important factor to consider when applying generative models to real problems. Closed-loop systems combining prediction, robotic synthesis, and characterization show promise, but discrepancies between computational predictions and experimental outcomes can arise from unmodeled phenomena like defects or (known/unknown) environmental effects [30]. Transformers (e.g., CrystalFormer-RL [20]) and NFs (e.g., conditional NFs [123]) typically require robust feedback loops to refine predictions in real-time. Standardization of experimental protocols across self-driving labs is therefore critical for reproducibility [28].

The rising preference towards larger models (e.g., diffusion-based models, transformers, or large language models) that supposedly able to predict materials to fight climate change [23] ironically consumes significantly higher amounts of energy. Researchers at MIT noted a significant rise in global electricity consumption of data centres, expected to range between 620 — 1,050 TWh in 2026 is significantly attributed to the rising popularity of generative models [11]. This reveals an interesting dilemma: that AI can be both a part of the solution and a contributing factor to the energy problem. In this regard, more efficient models like NFs and GFlowNets may be preferred.

Another growing concern with the gravitation of AI towards the materials discovery space is the ethical considerations. Considering the significant change in all aspects of life, livelihood, and liberty that AI has brought upon [25], we have to be aware of the risk of misuse of AI to generate toxic or hazardous materials [85]. While the nefarious consequences or environmental impacts are not exclusively caused by AI, the leading AI society has recognised the need for AI governance principles and broad ethical guidelines [115]. We believe interdisciplinary committees that include the social sciences field are required to devise safeguards and legal frameworks around AI-related works, including materials discovery.

5 Future Trends and Emerging Research Directions

The rapid evolution of artificial intelligence (AI) and generative models is poised to revolutionize materials discovery, enabling the design of novel materials with unprecedented precision and efficiency. As computational power, data availability, and algorithmic sophistication advance, emerging trends are shaping the future of AI-driven materials science. This section explores these directions, focusing on advancements in generative models, integration with experimental and computational workflows, solutions to current limitations, and the ethical implications of AI in materials design, building on recent developments in diffusion models, Transformers, Normalizing Flows, and GFlowNets.

5.1 Emerging Trends in Generative Models

Generative models are evolving toward more robust, versatile, and physically grounded architectures. Diffusion models have surpassed GANs in stability and quality for generating complex material structures, such as crystals and porous frameworks. Models like DiffCSP[61] use periodic-E(3)-equivariant denoising to predict stable crystals, while SymmCD[73] ensures realistic symmetries across all 230 space groups. Other diffusion-based approaches, such as WyckoffDiff[65] and CrysLDM[66], further enhance crystal generation, with applications in electronics and catalysis[22,69].

Transformers are gaining prominence, leveraging attention mechanisms for multi-property inverse design and symmetry-constrained crystal generation. MatterGPT[23] optimizes materials across diverse properties, while the Space Group Informed Transformer[18] and Wyckoff Transformer[64] generate synthesisable crystals with crystallographic constraints. CrystalFormer-RL[20] integrates reinforcement learning for targeted material design.

Normalizing Flows (NFs) offer exact likelihoods and stable training, with CrystalFlow[77] and FlowLLM[112] generating crystalline electrolytes and polymers, complementing diffusion models[82]. GFlowNets, such as Crystal-GFN[5], sample diverse crystals with tailored properties, enhancing high-throughput screening. **Foundation models**, pre-trained on expansive datasets like Materials Project, enable transfer learning across material classes, reducing task-specific data needs[137]. **Multi-modal generative models** integrate text, chemical structures, and spectroscopic data, facilitating intuitive design through text-conditioned generation[83]. **Physics-informed generative models** embed thermodynamic and quantum mechanical constraints, ensuring physically realistic outputs for real-world synthesis[24,131]. These advancements are summarized in Fig. 12 and Table 6.

5.2 Integration with Experimental and Computational Methods

The convergence of generative models with experimental platforms is driving closed-loop discovery systems, where AI proposes material candidates that are autonomously synthesized and characterized[28]. Frameworks like WyCryst[140] and CrySPR[88] integrate generative models (e.g., Transformers, NFs) with robotic labs for iterative refinement. **Digital twins**, virtual representations of material systems, enable rapid screening of AI-generated candidates, bridging simulation and experiment[81]. The advent of **quantum computing** may enhance generative models by accelerating quantum mechanical calculations, enabling precise property predictions for complex materials like high-entropy alloys[24]. Advances in active learning, coupled with models like CrystalFormer-RL[20], optimize experimental workflows by prioritizing high-value candidates.

5.3 Addressing Current Challenges

Overcoming limitations in data availability, synthesizability, and interpretability remains critical. **Federated learning** and **synthetic data generation** using diffusion models (e.g., DiffCSP, SymmCD) expand datasets, mitigating scarcity[40]. To ensure **synthesizability**, models incorporate experimental constraints, such as reaction kinetics and precursor availability, as seen in chemically guided diffusion models[92]. **Explainable AI (XAI)** techniques, including attention visualization in Transformers and graph-based models, enhance interpretability by elucidating model decisions[21]. **Standardized benchmarks**, such as Dismai-Bench[133], promote robust evaluation and reproducibility. Computational efficiency is also a focus, with NFs and GFlowNets offering stable training but requiring optimization for large-scale applications[5,77].

5.4 Ethical and Societal Implications

Generative models introduce unique ethical and societal challenges that require careful consideration to ensure responsible use. Dataset biases in training data can significantly limit their impact. For instance, datasets often prioritize commercially viable materials, such as semiconductors for electronics, over biomaterials suited for low-resource medical applications, potentially neglecting global health needs[9]. Transparent data curation, including diverse and representative datasets, is essential to ensure equitable material generation. Potential misuse poses another critical risk, as generative models can inadvertently or intentionally design hazardous materials, such as toxic chemicals or unstable compounds[103]. Ethical guidelines, drawing from synthetic biology's safety protocols, and regulatory frameworks like the OECD AI principles can mitigate this risk by enforcing strict oversight and responsible use[oec]. Interpretability and trust challenges arise from the often opaque nature of generative models, which generate structures from noise or latent spaces, complicating validation by experimentalists[86]. For example, generative AI often suggests catalysts without a clear reasoning or underlying explanation for its predicted efficacy, which can erode trust. Physics-informed models and standardized reporting of uncertainties can enhance transparency and reliability. Unequal access to these computationally intensive models, which often require proprietary datasets or high-performance computing, risks widening global research disparities, particularly for under-resourced institutions in developing regions[81]. Open-access platforms, such as the Materials
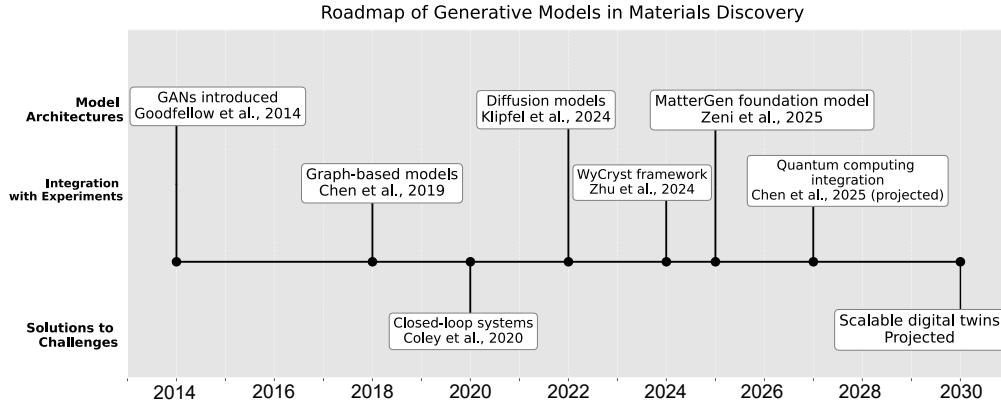
Figure 12: Roadmap of generative models in materials discovery, highlighting key milestones and projected future trends in model architectures, experimental integration, and solutions to challenges.

Table 6: Emerging Generative Models for Materials Discovery

Model Type	Advantages	Challenges	References
Diffusion Models	High-quality, stable generation; symmetry preservation	Computationally intensive; data dependency	Jiao et al.[61], Klipfel et al.[69], Levy et al.[73]
Transformers	Handles large datasets; multi-property design; symmetry constraints	Requires extensive training data; complex architectures	Cao et al.[18], Chen et al.[23], Kazeev et al.[64]
Normalizing Flows	Exact likelihoods; stable training; versatile for crystals and polymers	High computational cost; discrete structure challenges	Luo et al.[77], Miller et al.[82], Sriram et al.[112]
GFlowNets	Diverse sampling; tailored property optimization	Limited to specific tasks; scalability concerns	AI4Science et al.[5]
Foundation Models	Transfer learning; reduces data needs	High pretraining costs; generalizability concerns	Zeni et al.[137]
Multi-modal Models	Integrates text and structural data; intuitive design	Data heterogeneity; model complexity	Mohanty et al.[83]
Physics-informed Models	Physically realistic outputs; improved synthesizability	Complex physical constraints; computational overhead	Chen et al.[24], Yang and Perdikaris[131]

Project, can democratize access, enabling broader participation in sustainable materials discovery[58]. To maximize their societal impact, generative models demand robust ethical frameworks, international collaboration, and transparent practices to ensure equitable, safe, and trustworthy innovation in materials science

6 Conclusion

The integration of artificial intelligence and generative models has fundamentally transformed materials discovery, enabling rapid identification and design of novel materials with tailored properties. These computational approaches overcome the limitations of traditional experimental methods, offering unprecedented opportunities for innovation across diverse applications, including energy storage, catalysis, electronics, biomaterials, and high-throughput screening[9,44]. Generative models, such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Diffusion Models (e.g., DiffCSP[61], SymmCD[73]), Transformers (e.g., MatterGPT[23], Space Group Informed Transformer[18]), Normalizing Flows (e.g., CrystalFlow[77], FlowLLM[112]), and GFlowNets (e.g., Crystal-GFN[5]), have demonstrated remarkable capabilities in learning complex material data relationships and generating candidates for crystalline, polymeric, and composite systems. Normalizing Flows, in particular, stand out for their exact likelihoods and stable training, advancing the design of electrolytes, catalysts, and polymers[123].

Despite significant progress, challenges persist in data quality and availability, model interpretability, computational cost, generalization, experimental integration, and ethical considerations[16,28]. Limited datasets, complex model mechanisms (e.g., Transformer attention, Diffusion symmetry constraints), high computational demands, and biases in datasets like Materials Project hinder scalability and reliability[23,81]. Ensuring synthesizability, addressing environmental impacts, and mitigating risks of misuse remain critical. However, the future of AI-driven materials discovery is promising, with emerging trends in multi-modal generative models, physics-informed architectures, efficient models like Normalizing Flows and GFlowNets, closed-loop autonomous experimentation, and synergies with quantum computing[20,83,131]. By addressing current limitations and pursuing these innovative directions, AI and generative models will continue to revolutionize materials science, unlocking advanced materials to tackle global challenges in sustainability, energy, and healthcare.

Acknowledgments

References

[oec] AI principles.

[2] (2025). FAIR Principles.

[3] (2025). Versailles Project on Advanced Materials and Standards (VAMAS).

[4] Agrotis, S., Emre Sener, M., Hagger, O. S. J., Handoko, A. D., and Caruana, D. J. (2024). One-step synthesis of nanosized cu-ag films using atmospheric pressure plasma jet. *Applied Materials Today*, 39:102286.

[5] AI4Science, M., Hernandez-Garcia, A., Duval, A., Volokhova, A., Bengio, Y., Sharma, D., Carrier, P. L., Benabed, Y., Koziarski, M., and Schmidt, V. (2023). Crystal-GFN: Sampling crystals with desirable properties and constraints.

[6] Al-Khaylani, H. H., Al-Sharify, T. A., Abbas, M. F., Hussein, H., Al-Shabandar, R., and Oleiwi, T. A. (2024). Generative Adversarial Networks to Design Metamaterials Based Nano-Photonics Devices. In *2024 4th International Conference on Artificial Intelligence and Signal Processing (AISP)*, pages 1–5.

[7] Alakhdar, A., Poczos, B., and Washburn, N. (2024). Diffusion Models in De Novo Drug Design. *J. Chem. Inf. Model.*, 64(19):7238–7256.

[8] Alverson, M., G. Baird, S., Murdock, R., Sin-Hang Ho, E., Johnson, J., and D. Sparks, T. (2024). Generative adversarial networks and diffusion models in material discovery. *Digital Discovery*, 3(1):62–80.

[9] Anstine, D. M. and Isayev, O. (2023). Generative Models as an Emerging Paradigm in the Chemical Sciences. *J. Am. Chem. Soc.*, 145(16):8736–8750.

[10] Baird, S. G., Jablonka, K. M., Alverson, M. D., Sayeed, H. M., Khan, M. F., Seegmiller, C., Smit, B., and Sparks, T. D. (2022). Xtal2png: A Python package for representing crystalstructure as PNG files. *JOSS*, 7(76):4528.

[11] Bashir, N., Donti, P., Cuff, J., Sroka, S., Ilic, M., Sze, V., Delimitrou, C., and Olivetti, E. (2024). The climate and sustainability implications of generative ai. *An MIT Exploration of Generative AI*, page Online only.

[12] Bedart, C., Shimokura, G., West, F. G., Wood, T. E., Batey, R. A., Irwin, J. J., and Schapira, M. (2024). The pan-canadian chemical library: A mechanism to open academic chemistry to high-throughput virtual screening. *Scientific Data*, 11(1):597. Bedart, Corentin Shimokura, Grace West, Frederick G Wood, Tabitha E Batey, Robert A Irwin, John J Schapira, Matthieu eng R01 GM071896/GM/NIGMS NIH HHS/ Dataset England 2024/06/07 Sci Data. 2024 Jun 6;11(1):597. doi: 10.1038/s41597-024-03443-5.

[13] Bengio, Y., Lahlou, S., Deleu, T., Hu, E. J., Tiwari, M., and Bengio, E. (2023). GFlowNet Foundations. *Journal of Machine Learning Research*, 24(210):1–55.

[14] Breuck, P.-P. D., Piracha, H. A., Rignanese, G.-M., and Marques, M. A. L. (2025). A generative material transformer using Wyckoff representation.

[15] Butler, K. T., Choudhary, K., Csanyi, G., Ganose, A. M., Kalinin, S. V., and Morgan, D. (2024). Setting standards for data driven materials science. *npj Computational Materials*, 10(1):231.

[16] Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., and Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, 559(7715):547–555.

[17] Cao, N. D. and Kipf, T. (2022). MolGAN: An implicit generative model for small molecular graphs.

[18] Cao, Z., Luo, X., Lv, J., and Wang, L. (2024). Space Group Informed Transformer for Crystalline Materials Generation.

[19] Cao, Z., Magar, R., Wang, Y., and Barati Farimani, A. (2023). MOFormer: Self-Supervised Transformer Model for Metal–Organic Framework Property Prediction. *J. Am. Chem. Soc.*, 145(5):2958–2967.

[20] Cao, Z. and Wang, L. (2025). CrystalFormer-RL: Reinforcement Fine-Tuning for Materials Design.

[21] Chen, C., Ye, W., Zuo, Y., Zheng, C., and Ong, S. P. (2019). Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. *Chem. Mater.*, 31(9):3564–3572.

[22] Chen, J., Guo, J., and Schwaller, P. (2025a). MatInvent: Reinforcement Learning for 3D Crystal Diffusion Generation. In *AI for Accelerated Materials Design - ICLR 2025*.

[23] Chen, Y., Wang, X., Deng, X., Liu, Y., Chen, X., Zhang, Y., Wang, L., and Xiao, H. (2024). MatterGPT: A Generative Transformer for Multi-Property Inverse Design of Solid-State Materials.

[24] Chen, Z., Meng, Z., He, T., Li, H., Cao, J., Xu, L., Xiao, H., Zhang, Y., He, X., and Fang, G. (2025b). Crystal Structure Prediction Meets Artificial Intelligence. *J. Phys. Chem. Lett.*, 16(10):2581–2591.

[25] Cheng, L., Varshney, K. R., and Liu, H. (2021). Socially responsible ai algorithms: Issues, purposes, and challenges. *Journal of Artificial Intelligence Research*, 71:1137–1181.

[26] Chitre, A., Cheng, J., Ahmed, S., Querimit, R., Hippalgaonkar, K., and Lapkin, A. (2023). pHbot: Self-Driven Robot for pH Adjustment of Viscous Formulations via Physics-informed-ML. Preprint, Chemistry.

[27] Chitturi, S. R., Ramdas, A., Wu, Y., Rohr, B., Ermon, S., Dionne, J., Jornada, F. H. D., Dunne, M., Tassone, C., Neiswanger, W., and Ratner, D. (2024). Targeted materials discovery using bayesian algorithm execution. *npj Computational Materials*, 10(1):156.

[28] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2020). Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie International Edition*, 59(51):22858–22893.

[29] Collins, C. R., Gordon, G. J., von Lilienfeld, O. A., and Yaron, D. J. (2018). Constant size descriptors for accurate machine learning models of molecular properties. *The Journal of Chemical Physics*, 148(24):241718.

[30] Correa-Baena, J.-P., Hippalgaonkar, K., van Duren, J., Jaffer, S., Chandrasekhar, V. R., Stevanovic, V., Wadia, C., Guha, S., and Buonassisi, T. (2018). Accelerating Materials Development via Automation, Machine Learning, and High-Performance Computing. *Joule*, 2(8):1410–1420.

[31] Cunningham, J. D., Simpson, T. W., and Tucker, C. S. (2019). An Investigation of Surrogate Models for Efficient Performance-Based Decoding of 3D Point Clouds. *Journal of Mechanical Design*, 141(121401):121401.

[32] Curie, P., Curie, M., and Bémont, G. (1898). Sur une nouvelle substance fortement radio-active, contenue dans la pechblende. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 127:1215–1217.

[33] Curtarolo, S., Setyawan, W., Hart, G. L., Jahnatek, M., Chepulskii, R. V., Taylor, R. H., Wang, S., Xue, J., Yang, K., Levy, O., Mehl, M. J., Stokes, H. T., Demchenko, D. O., and Morgan, D. (2012). Aflow: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58:218–226.

[34] Dan, Y., Zhao, Y., Li, X., Li, S., Hu, M., and Hu, J. (2020). Generative adversarial networks (GAN) based efficient sampling of chemical composition space for inverse design of inorganic materials. *npj Comput Mater*, 6(1):1–7.

[35] Das, B., Peters, A., Li, G., and Hei, X. (2025). Generative Design of Thermoset Shape Memory Polymers Driven by Chemical Group: A Conditional Variational Autoencoder Approach. *Journal of Polymer Science*, 63(6):1334–1344.

[36] Das, K., Goyal, P., Lee, S.-C., Bhattacharjee, S., and Ganguly, N. (2023). Crysmmnet: Multimodal representation for crystal property prediction. In *39th Conference on Uncertainty in Artificial Intelligence*, volume 216, pages 507–517. PMLR.

[37] Daunhawer, I., Sutter, T. M., Chin-Cheong, K., Palumbo, E., and Vogt, J. E. (2021). On the limitations of multimodal vaes. In *The Tenth International Conference on Learning Representations*.

[38] de Pablo, J. J., Jackson, N. E., Webb, M. A., Chen, L.-Q., Moore, J. E., Morgan, D., Jacobs, R., Pollock, T., Schlom, D. G., Toberer, E. S., Analytis, J., Dabo, I., DeLongchamp, D. M., Fiete, G. A., Grason, G. M., Hautier, G., Mo, Y., Rajan, K., Reed, E. J., Rodriguez, E., Stevanovic, V., Suntivich, J., Thornton, K., and Zhao, J.-C. (2019). New frontiers for the materials genome initiative. *npj Comput Mater*, 5(1):1–23.

[39] Dobson, C. M. (2004). Chemical space and biology. *Nature*, 432(7019):824–828.

[40] Dong, R., Fu, N., Siriwardane, E. M. D., and Hu, J. (2024). Generative Design of Inorganic Compounds Using Deep Diffusion Language Models. *J. Phys. Chem. A*, 128(29):5980–5989.

[41] El-Awady, K. (2023). VAE for Modified 1-Hot Generative Materials Modeling, A Step Towards Inverse Material Design.

[42] Epps, R. W., Volk, A. A., Reyes, K. G., and Abolhasani, M. (2021). Accelerated AI development for autonomous materials synthesis in flow. *Chem. Sci.*, 12(17):6025–6036.

[43] Friederich, P., Häse, F., Proppe, J., and Aspuru-Guzik, A. (2021). Machine-learned potentials for next-generation matter simulations. *Nature Materials*, 20(6):750–761. Friederich, Pascal Hase, Florian Proppe, Jonny Aspuru-Guzik, Alan eng Research Support, Non-U.S. Gov't Review England 2021/05/29 Nat Mater. 2021 Jun;20(6):750-761. doi: 10.1038/s41563-020-0777-6. Epub 2021 May 27.

[44] Fuhr, A. S. and Sumpter, B. G. (2022). Deep Generative Models for Materials Discovery and Machine Learning-Accelerated Innovation. *Front. Mater.*, 9:865270.

[45] Gisperg, F., Klausser, R., Elshazly, M., Kopp, J., Brichtová, E. P., and Spadiut, O. (2025). Bayesian optimization in bioprocess engineering—where do we stand today? *Biotechnology and Bioengineering*, 122(6):1313–1325. Gisperg, Florian Klausser, Robert Elshazly, Mohamed Kopp, Julian Brichtova, Eva Prada Spadiut, Oliver eng This study was supported by Christian Doppler Forschungsgesellschaft./ Review 2025/03/05 Biotechnol Bioeng. 2025 Jun;122(6):1313-1325. doi: 10.1002/bit.28960. Epub 2025 Mar 5.

[46] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules. *ACS Cent. Sci.*, 4(2):268–276.

[47] Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, volume 27.

[48] Gui, J., Sun, Z., Wen, Y., Tao, D., and Ye, J. (2023). A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3313–3332.

[49] Gupta, N. K., Guo, Y., Chang, S. Y., Lin, J., Khoo, Z. H. J., I Made, R., Ooi, Z.-E., Lim, C. Y. J., Lee, C. H., M, S., Lim, Y.-F., Khoo, E., Lu, W. F., Lum, Y., and Handoko, A. D. (2024). Towards a greener electrosynthesis: pairing machine learning and 3d printing for rapid optimisation of anodic trifluoromethylation. *RSC Sustainability*, 2(2):536–545.

[50] Han, S., Kang, Y., Park, H., Yi, J., Park, G., and Kim, J. (2024). Multimodal transformer for property prediction in polymers. *ACS Applied Materials & Interfaces*, 16(13):16853–16860.

[51] Hautier, G., Fischer, C., Ehrlacher, V., Jain, A., and Ceder, G. (2011). Data mined ionic substitutions for the discovery of new compounds. *Inorganic Chemistry*, 50(2):656–663.

[52] Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

[53] Honda, S., Shi, S., and Ueda, H. R. (2019). SMILES Transformer: Pre-trained Molecular Fingerprint for Low Data Drug Discovery.

[54] Huang, Q., Li, Y., Zhu, L., Zhao, Q., and Yu, W. (2025). Unified multimodal multidomain polymer representation for property prediction. *npj Computational Materials*, 11(1):153.

[55] Huang, X., Ma, S., Wu, Y., Wan, C., Zhao, C. Y., Wang, H., and Ju, S. (2023). High-throughput screening of amorphous polymers with high intrinsic thermal conductivity via automated physical feature engineering. *Journal of Materials Chemistry A*, 11(38):20539–20548.

[56] Ishikawa, A. (2022). Heterogeneous catalyst design by generative adversarial network and first-principles based microkinetics. *Sci Rep*, 12(1):11657.

[57] Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. A. (2013a). Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002.

[58] Jain, A., Ong, S. P., Hautier, G., Chen, W., Richards, W. D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K. A. (2013b). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002.

[59] Jain, M., Deleu, T., Hartford, J., Liu, C.-H., Hernandez-Garcia, A., and Bengio, Y. (2023). GFlowNets for AI-driven scientific discovery. *Digital Discovery*, 2(3):557–577.

[60] Jiang, C., He, H., Guo, H., Zhang, X., Han, Q., Weng, Y., Fu, X., Zhu, Y., Yan, N., Tu, X., and Sun, Y. (2024). Transfer learning guided discovery of efficient perovskite oxide for alkaline water oxidation. *Nature Communications*, 15(1):6301. Jiang, Chang He, Hongyuan Guo, Hongquan Zhang, Xiaoxin Han, Qingyang Weng, Yanhong Fu, Xianzhu Zhu, Yinlong Yan, Ning Tu, Xin Sun, Yifei eng EP/V036696/RCUK — Engineering and Physical Sciences Research Council (EPSRC)/ 22272136/National Science Foundation of China — National Natural Science Foundation of China-Yunnan Joint Fund (NSFC-Yunnan Joint Fund)/ 22102135/National Science Foundation of China — National Natural Science Foundation of China-Yunnan Joint Fund (NSFC-Yunnan Joint Fund)/ England 2024/07/27 Nat Commun. 2024 Jul 26;15(1):6301. doi: 10.1038/s41467-024-50605-5.

[61] Jiao, R., Huang, W., Lin, P., Han, J., Chen, P., Lu, Y., and Liu, Y. (2023). Crystal Structure Prediction by Joint Equivariant Diffusion. *Advances in Neural Information Processing Systems*, 36:17464–17497.

[62] Karpovich, C., Pan, E., Jensen, Z., and Olivetti, E. (2023). Interpretable Machine Learning Enabled Inorganic Reaction Classification and Synthesis Condition Prediction. *Chem. Mater.*, 35(3):1062–1079.

[63] Karpovich, C., Pan, E., and Olivetti, E. A. (2024). Deep reinforcement learning for inverse inorganic materials design. *npj Computational Materials*, 10(1):287.

[64] Kazeev, N., Nong, W., Romanov, I., Zhu, R., Ustyuzhanin, A., Yamazaki, S., and Hippalgaonkar, K. (2025). Wyckoff Transformer: Generation of Symmetric Crystals.

[65] Kelvinius, F. E., Andersson, O. B., Parackal, A. S., Qian, D., Armiento, R., and Lindsten, F. (2025). WyckoffDiff – A Generative Diffusion Model for Crystal Symmetry.

[66] Khastagir, S., Das, K., Goyal, P., Lee, S.-C., Bhattacharjee, S., and Ganguly, N. (2025). CrysLDM: Latent Diffusion Model for Crystal Material Generation. In *AI for Accelerated Materials Design - ICLR 2025*.

[67] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B., Thiessen, P., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. (2024). Pubchem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525.

[68] Kingma, D. P. and Welling, M. (2022). Auto-Encoding Variational Bayes.

[69] Klipfel, A., Fregier, Y., Sayede, A., and Bouraoui, Z. (2024). Vector Field Oriented Diffusion Model for Crystal Material Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22193–22201.

[70] Kobyzev, I., Prince, S. J., and Brubaker, M. A. (2021). Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979.

[71] Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (SELFIES): A 100% robust molecular string representation. *Mach. Learn.: Sci. Technol.*, 1(4):045024.

[72] Lai, P., Amirkulova, F., and Gerstoft, P. (2021). Conditional Wasserstein generative adversarial networks applied to acoustic metamaterial design. *The Journal of the Acoustical Society of America*, 150(6):4362–4374.

[73] Levy, D., Panigrahi, S. S., Kaba, S.-O., Zhu, Q., Lee, K. L. K., Galkin, M., Miret, S., and Ravanbakhsh, S. (2025). SymmCD: Symmetry-Preserving Crystal Generation with Diffusion Models.

[74] Lim, C. Y. J., I Made, R., Khoo, Z. H. J., Ng, C. K., Bai, Y., Wang, J., Yang, G., Handoko, A. D., and Lim, Y.-F. (2023). Machine learning-assisted optimization of multi-metal hydroxide electrocatalysts for overall water splitting. *Materials Horizons*, 10(11):5022–5031. Lim, Carina Yi Jing I Made, Riko Khoo, Zi Hui Jonathan Ng, Chee Koon Bai, Yang Wang, Jianbiao Yang, Gaoliang Handoko, Albertus D Lim, Yee-Fun eng England 2023/08/30 Mater Horiz. 2023 Aug 30. doi: 10.1039/d3mh00788j.

[75] Lim, S., Lee, S., Piao, Y., Choi, M., Bang, D., Gu, J., and Kim, S. (2022). On modeling and utilizing chemical compound information with deep learning technologies: A task-oriented approach. *Computational and Structural Biotechnology Journal*, 20:4288–4304.

[76] Lu, S., Zhou, Q., Ouyang, Y., Guo, Y., Li, Q., and Wang, J. (2018). Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat Commun*, 9(1):3405.

[77] Luo, X., Wang, Z., Wang, Q., Lv, J., Wang, L., Wang, Y., and Ma, Y. (2025). CrystalFlow: A Flow-Based Generative Model for Crystalline Materials.

[78] MacLeod, B. P., Parlane, F. G. L., Brown, A. K., Hein, J. E., and Berlinguette, C. P. (2021). Flexible automation accelerates materials discovery. *Nat. Mater.*, pages 1–5.

[79] MacLeod, B. P., Parlane, F. G. L., Rupnow, C. C., Dettelbach, K. E., Elliott, M. S., Morrissey, T. D., Haley, T. H., Proskurin, O., Rooney, M. B., Taherimakhsousi, N., Dvorak, D. J., Chiu, H. N., Waizenegger, C. E. B., Ocean, K., Mokhtari, M., and Berlinguette, C. P. (2022). A self-driving laboratory advances the Pareto front for material properties. *Nat Commun*, 13(1):995.

[80] Mehta, K. H., I Made, R., Parkin, I. P., Sankar, G., and Handoko, A. D. (2025). A paradigm shift: From batch processing to flow chemistry. *Small*, page e2411519. Mehta, Kallum Hiten I Made, Riko Parkin, Ivan P Sankar, Gopinathan Handoko, Albertus Denny eng C231218004/A*STAR Horizontal Technology Coordinating Office/ HQ/R25-ARAP210202/A*STAR Research Attachment Programme/ A20G9b0135/A*STAR RIE2020 AME Programmatic Fund/ M22K8a0048/A*STAR RIE2025 MTC IAF-PP/ Germany 2025/05/03 22:25 Small. 2025 May 2:e2411519. doi: 10.1002/smll.202411519.

[81] Merchant, A., Batzner, S., Schoenholz, S. S., Aykol, M., Cheon, G., and Cubuk, E. D. (2023). Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85.

[82] Miller, B. K., Chen, R. T. Q., Sriram, A., and Wood, B. M. (2024). FlowMM: Generating Materials with Riemannian Flow Matching. In *Forty-First International Conference on Machine Learning*.

[83] Mohanty, T., Mehta, M., Sayeed, H. M., Srikumar, V., and Sparks, T. D. (2024). CrysText: A Generative AI Approach for Text-Conditioned Crystal Structure Generation using LLM.

[84] Monticelli, L. and Tieleman, D. P. (2013). *Force Fields for Classical Molecular Dynamics*, pages 197–213. Humana Press, Totowa, NJ.

[85] Mullin, R. (2023). The ethics of ai in the lab. *C&EN Global Enterprise*, 101(31):30–35.

[86] Musil, F., Grisafi, A., Bartók, A. P., Ortner, C., Csányi, G., and Ceriotti, M. (2021). Physics-Inspired Structural Representations for Molecules and Materials. *Chem. Rev.*, 121(16):9759–9815.

[87] Noh, J., Kim, J., Stein, H. S., Sanchez-Lengeling, B., Gregoire, J. M., Aspuru-Guzik, A., and Jung, Y. (2019). Inverse Design of Solid-State Materials via a Continuous Representation. *Matter*, 1(5):1370–1384.

[88] Nong, W., Zhu, R., and Hippalgaonkar, K. (2024). CrySPR: A Python interface for implementation of crystal structure pre-relaxation and prediction using machine-learning interatomic potentials.

[89] Nouira, A., Sokolovska, N., and Crivello, J.-C. (2019). CrystalGAN: Learning to Discover Crystallographic Structures with Generative Adversarial Networks.

[90] Orlando, G., Raimondi, D., Duran-Romaña, R., Moreau, Y., Schymkowitz, J., and Rousseau, F. (2022). Pyuul provides an interface between biological structures and deep learning algorithms. *Nature Communications*, 13(1):961.

[91] Oubari, F., Mathelin, A. d., Décatoire, R., and Mougeot, M. (2021). A binded vae for inorganic material generation. In *NeurIPS 2021 · Thirty-Fifth Annual Conference on Neural Information Processing Systems*, Workshop on Deep Generative Models and Downstream Applications.

[92] Pan, E., Kwon, S., Liu, S., Xie, M., Duan, Y., Prein, T., Sheriff, K., Roman, Y., Moliner, M., Gomez-Bombarelli, R., and Olivetti, E. (2024). A Chemically-Guided Generative Diffusion Model for Materials Synthesis Planning. In *AI for Accelerated Materials Design - NeurIPS 2024*.

[93] Papamakarios, G., Nalisnick, E., Rezende, D. J., Mohamed, S., and Lakshminarayanan, B. (2021). Normalizing Flows for Probabilistic Modeling and Inference. *Journal of Machine Learning Research*, 22(57):1–64.

[94] Park, J., Singh Gill, A. P., Mohamad Moosavi, S., and Kim, J. (2024). Inverse design of porous materials: A diffusion model approach. *Journal of Materials Chemistry A*, 12(11):6507–6514.

[95] Pickard, C. J. and Needs, R. J. (2011). Ab initio random structure searching. *Journal of Physics: Condensed Matter*, 23(5):053201.

[96] Ren, F., Ward, L., Williams, T., Laws, K. J., Wolverton, C., Hattrick-Simpers, J., and Mehta, A. (2018). Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Science Advances*, 4(4):eaaq1566.

[97] Reynard, K., Barry, T., and Reynard, K. (1992). *Standards for Materials Databases - National and International Programmes - Do They Provide for Data Users Needs?*, volume STP1140-EB, page 0. ASTM International.

[98] Röcken, S. and Zavadlav, J. (2024). Accurate machine learning force fields via experimental and simulation data fusion. *npj Comput Mater*, 10(1):69.

[99] Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408. ROSENBLATT, F eng 1958/11/01 Psychol Rev. 1958 Nov;65(6):386-408. doi: 10.1037/h0042519.

[100] Rummukainen, H., Hörhammer, H., Kuusela, P., Kilpi, J., Sirviö, J., and Mäkelä, M. (2024). Traditional or adaptive design of experiments? a pilot-scale comparison on wood delignification. *Heliyon*, 10(2):e24484.

[101] Ruthotto, L. and Haber, E. (2021). An introduction to deep generative modeling. *GAMM-Mitteilungen*, 44(2):e202100008.

[102] Saal, J. E., Kirklin, S., Aykol, M., Meredig, B., and Wolverton, C. (2013). Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM*, 65(11):1501–1509.

[103] Sanchez-Lengeling, B. and Aspuru-Guzik, A. (2018). Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365.

[104] Schneider, G. and Fechner, U. (2005). Computer-based de novo design of drug-like molecules. *Nat Rev Drug Discov*, 4(8):649–663.

[105] Schultz, P. G. and Lerner, R. A. (1995). From molecular diversity to catalysis: Lessons from the immune system. *Science*, 269(5232):1835–1842. Schultz, P G Lerner, R A eng Research Support, Non-U.S. Gov't Research Support, U.S. Gov't, Non-P.H.S. Research Support, U.S. Gov't, P.H.S. Review 1995/09/29 Science. 1995 Sep 29;269(5232):1835-42. doi: 10.1126/science.7569920.

[106] Schultz, P. G. and Xiang, X.-D. (1998). Combinatorial approaches to materials science. *Current Opinion in Solid State and Materials Science*, 3(2):153–158.

[107] Schütt, K., Kindermans, P.-J., Sauceda Felix, H. E., Chmiela, S., Tkatchenko, A., and Müller, K.-R. (2017). SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, volume 30.

[108] Segler, M. H. S., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating Focused Molecule Libraries for Drug Discovery with Recurrent Neural Networks. *ACS Cent. Sci.*, 4(1):120–131.

[109] Simonovsky, M. and Komodakis, N. (2018). GraphVAE: Towards Generation of Small Graphs Using Variational Autoencoders. In *Artificial Neural Networks and Machine Learning – ICANN 2018*, pages 412–422.

[110] Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., and Ganguli, S. (2015). Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265.

[111] Sohn, K., Lee, H., and Yan, X. (2015). Learning Structured Output Representation using Deep Conditional Generative Models. In *Advances in Neural Information Processing Systems*, volume 28.

[112] Sriram, A., Miller, B. K., Chen, R. T., and Wood, B. M. (2024). FlowLLM: Flow Matching for Material Generation with Large Language Models as Base Distributions. *Advances in Neural Information Processing Systems*, 37:46025–46046.

[113] Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N. M., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., and Collins, J. J. (2020). A Deep Learning Approach to Antibiotic Discovery. *Cell*, 180(4):688–702.e13.

[114] Sun, S., Wang, X., Jiang, Y., Lei, Y., Zhang, S., Kumar, S., Zhang, J., Ma, E., Mazzarello, R., Wang, J.-J., and Zhang, W. (2024). High-throughput screening to identify two-dimensional layered phase-change chalcogenides for embedded memory applications. *npj Computational Materials*, 10(1):189.

[115] Team, T. F. (2017). A principled ai discussion in asilomar.

[116] Tempke, R. and Musho, T. (2022). Autonomous design of new chemical reactions using a variational autoencoder. *Commun Chem*, 5(1):1–10.

[117] Tonegawa, S. (1983). Somatic generation of antibody diversity. *Nature*, 302(5909):575–581.

[118] Trask, N., Martinez, C., Shilt, T., Walker, E., Lee, K., Garland, A., Adams, D. P., Curry, J. F., Dugger, M. T., Larson, S. R., and Boyce, B. L. (2024). Unsupervised physics-informed disentanglement of multimodal materials data. *Materials Today*, 80:286–296.

[119] Unke, O. T., Chmiela, S., Sauceda, H. E., Gastegger, M., Poltavsky, I., Schütt, K. T., Tkatchenko, A., and Müller, K.-R. (2021). Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186. Unke, Oliver T Chmiela, Stefan Sauceda, Huziel E Gastegger, Michael Poltavsky, Igor Schutt, Kristof T Tkatchenko, Alexandre Muller, Klaus-Robert eng Research Support, Non-U.S. Gov't 2021/03/12 Chem Rev. 2021 Aug 25;121(16):10142-10186. doi: 10.1021/acs.chemrev.0c01111. Epub 2021 Mar 11.

[120] Vasudevan, R. K., Choudhary, K., Mehta, A., Smith, R., Kusne, G., Tavazza, F., Vlcek, L., Ziatdinov, M., Kalinin, S. V., and Hattrick-Simpers, J. (2019). Materials science in the artificial intelligence age: High-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Communications*, 9(3):821–838.

[121] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ukasz Kaiser, Ł., and Polosukhin, I. (2017). Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30.

[122] Vasylenko, A., Gamon, J., Duff, B. B., Gusev, V. V., Daniels, L. M., Zanella, M., Shin, J. F., Sharp, P. M., Morscher, A., Chen, R., Neale, A. R., Hardwick, L. J., Claridge, J. B., Blanc, F., Gaultois, M. W., Dyer, M. S., and Rosseinsky, M. J. (2021). Element selection for crystalline inorganic solid discovery guided by unsupervised machine learning of experimentally explored chemistry. *Nat Commun*, 12(1):5561.

[123] Wang, J. S., Hyatt, J. S., and Fish, M. (2024). Using conditional normalizing flows to generate material placements in an optimized thermal composite. *International Journal of Heat and Mass Transfer*, 224:125287.

[124] Wang, R., Xu, C., Dong, R., Luo, Z., Zheng, R., and Zhang, X. (2023a). A secured big-data sharing platform for materials genome engineering: State-of-the-art, challenges and architecture. *Future Generation Computer Systems*, 142:59–74.

[125] Wang, Y., Chen, S., Chen, G., Shurberg, E., Liu, H., and Hong, P. (2023b). Motif-based graph representation learning with application to chemical molecules. *Informatics*, 10(1):8.

[126] Weininger, D. (1988). SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28(1):31–36.

[127] Winter, B., Winter, C., Schilling, J., and Bardow, A. (2022). A smile is all you need: Predicting limiting activity coefficients from SMILES with natural language processing. *Digital Discovery*, 1(6):859–869.

[128] Xie, T., Fu, X., Ganea, O.-E., Barzilay, R., and Jaakkola, T. (2022). Crystal Diffusion Variational Autoencoder for Periodic Material Generation.

[129] Xie, T. and Grossman, J. C. (2018). Crystal Graph Convolutional Neural Networks for an Accurate and Interpretable Prediction of Material Properties. *Phys. Rev. Lett.*, 120(14):145301.

[130] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., and Yang, M.-H. (2023). Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.*, 56(4):105:1–105:39.

[131] Yang, Y. and Perdikaris, P. (2018). Physics-informed deep generative models. In *3rd workshop on Bayesian Deep Learning (NeurIPS)*.

[132] Yeung, C., Tsai, R., Pham, B., King, B., Kawagoe, Y., Ho, D., Liang, J., Knight, M. W., and Raman, A. P. (2021). Global Inverse Design across Multiple Photonic Structure Classes Using Generative Deep Learning. *Advanced Optical Materials*, 9(20):2100548.

[133] Yong, A. X. B., Su, T., and Ertekin, E. (2024). Dismai-Bench: Benchmarking and designing generative models using disordered materials and interfaces. *Digital Discovery*, 3(9):1889–1909.

[134] Zagorac, D., Müller, H., Ruehl, S., Zagorac, J., and Rehme, S. (2019). Recent developments in the inorganic crystal structure database: theoretical crystal structure data and related features. *Journal of Applied Crystallography*, 52(5):918–925. Zagorac, D Muller, H Ruehl, S Zagorac, J Rehme, S eng 2019/10/23 J Appl Crystallogr. 2019 Sep 23;52(Pt 5):918-925. doi: 10.1107/S160057671900997X. eCollection 2019 Oct 1.

[135] Zeng, M., Du, Y., Jiang, Q., Kempf, N., Wei, C., Bimrose, M. V., Tanvir, A. N. M., Xu, H., Chen, J., Kirsch, D. J., Martin, J., Wyatt, B. C., Hayashi, T., Saeidi-Javash, M., Sakaue, H., Anasori, B., Jin, L., Mcmurtrey, M. D., and Zhang, Y. (2023). High-throughput printing of combinatorial materials from aerosols. *Nature*, 617(7960):292–298.

[136] Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Shysheya, S., Crabbé, J., Sun, L., Smith, J., Nguyen, B., Schulz, H., Lewis, S., Huang, C.-W., Lu, Z., Zhou, Y., Yang, H., Hao, H., Li, J., Tomioka, R., and Xie, T. (2024). MatterGen: A generative model for inorganic materials design.

[137] Zeni, C., Pinsler, R., Zügner, D., Fowler, A., Horton, M., Fu, X., Wang, Z., Shysheya, A., Crabbé, J., Ueda, S., Sordillo, R., Sun, L., Smith, J., Nguyen, B., Schulz, H., Lewis, S., Huang, C.-W., Lu, Z., Zhou, Y., Yang, H., Hao, H., Li, J., Yang, C., Li, W., Tomioka, R., and Xie, T. (2025). A generative model for inorganic materials design. *Nature*, 639(8055):624–632.

[138] Zhang, H., Fu, H., Zhu, S., Yong, W., and Xie, J. (2021). Machine learning assisted composition effective design for precipitation strengthened copper alloys. *Acta Materialia*, 215:117118.

[139] Zhu, Q., Liu, Z., and Yan, J. (2020). Machine learning for metal additive manufacturing: Predicting temperature and melt pool fluid dynamics using physics-informed neural networks.

[140] Zhu, R., Nong, W., Yamazaki, S., and Hippalgaonkar, K. (2024). WyCryst: Wyckoff inorganic crystal generator framework. *Matter*, 7(10):3469–3488.

[141] Zuo, Y., Qin, M., Chen, C., Ye, W., Li, X., Luo, J., and Ong, S. P. (2021). Accelerating materials discovery with Bayesian optimization and graph deep learning. *Materials Today*, 51:126–135.