

Mapping the Molecular Universe

Daniel Hannah, Yumin Zhang, Stanley Whittingham,* Martin Winter,* Y. Shirley Meng,* Qichao Hu,* and Kang Xu*



Cite This: <https://doi.org/10.1021/jacsau.6c01250>



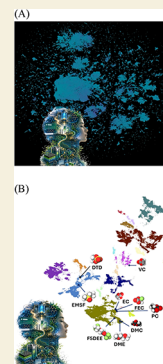
Read Online

ACCESS |

Metrics & More

Article Recommendations

ABSTRACT: The success of large-language-models highlighted the importance of datasets, whose quantity, quality, and diversity dictate the accuracy and creativity of artificial intelligence (AI) systems in learning and inferring. However, in scientific domains, such datasets required for domain-specific AI/machine-learning (ML) models do not exist most of the time. In this perspective, by examining the logics and rationales behind the construction and mapping of a universe for small organic molecules (10^8 – 10^{12}), we explore what new capabilities AI can bring to the materials discoveries. We identify the elusive structure–property relationship as a core challenge for new materials design and explore the possibility AI tools could enable us to design based on targeted properties instead of the structural derivation. In a broader context, we further discuss what challenges and limitations AI-driven material sciences still face and envision possible avenues into the future of AI4Materials.



KEYWORDS: Molecular Universe, AI4Science, Large-language-models, Databases, Computation Chemistry, Materials Discovery, Batteries, Electrolytes, Autonomous Laboratories

1. DATASETS FOR AI4SCIENCE

The ongoing artificial intelligence (AI) revolution rapidly expands the boundaries of machine capabilities. While large-language models (LLMs) have demonstrated to the public how an AI system can remarkably simulate human intelligence by understanding, learning, and inferring knowledge across a wide range of regimes, their success critically relies on the immense high-quality, diverse, and open-sourced data sets, which have been made available by the explosive growth of Internet during the past four decades. By the end of 2025, the entire Internet, consisting of texts, graphs, and videos, was estimated to be >220 zetabytes (> 10^{23} bytes).^{1,2} Together with advanced algorithms (number of parameters) and computing infrastructures (number of graphic-processing units, or GPUs), they dictate how an AI system can learn and infer to generalize across different scenarios.

For scientific domains, unfortunately, such data-sets do not exist for several reasons:

- (1) **Availability:** Most scientific datasets published in peer-reviewed literature are not open-sourced but instead behind paywalls, which sets a high barrier for AI4Science model training. On the other hand, many industry sectors have been accumulating datasets of both high quantity and high quality via their internal R&D, which were never published because commercial interests prevent them from doing so. Only a handful of exceptions exist in selected domains, one prominent example of which is the *Worldwide Protein Data Bank* (wwPDB),³ whose

abundant experimentally determined protein structures directly ensured the success of AlphaFold.⁴

- (2) **Quantity:** Compared with the data size on the order of 10^{23} bytes accessible for LLMs pretrainings, scientific datasets are much smaller in size. Although there have been no quantitative statistics available, a rough analysis of the major databases of science publishers and repositories^{5–8} puts it on the order of 100–1000 petabytes ($10^{17–18}$ bytes), with specific domains such as battery materials on much smaller ($10^{12–13}$ bytes).
- (3) **Quality:** It may sound strange that the scientific datasets in published literatures are not of good quality, but this is a fact that stems from multiple origins, even after we exclude the deliberate fraudulent practices such as data fabrication: (a) Because science explores the unknown, hence more often than not, the data do not reflect the ground truth, but instead require an iterative process of reducing uncertainties and approaching the truth progressively.⁹ In this regard, each scientific article only represents the best knowledge that human can achieve at the moment; (b) Majority of scientific publications only

Received: August 3, 2026

Revised: October 1, 2026

Accepted: October 1, 2026

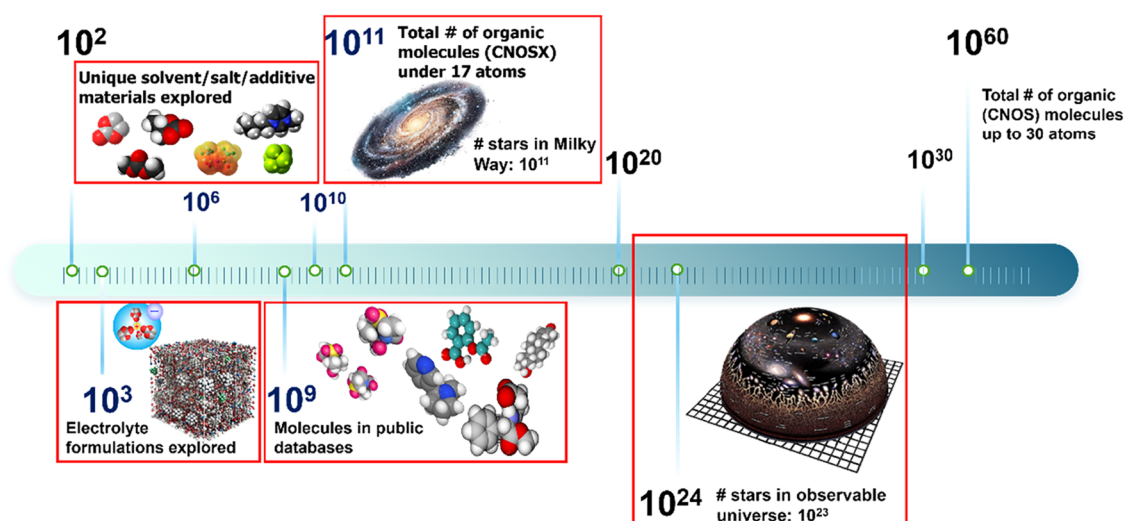


Figure 1. Number of explored electrolyte materials is dwarfed by the whole universe of molecules. Shown at the upper left corner in “ 10^2 order” are popular solvent molecules (ethylene carbonate, propylene carbonate, dimethyl carbonate) and salt anions (hexafluorophosphate, bisfluorosulfonyl imide), while in “ 10^3 ” order are typical electrolyte formulation compositions and structures consisting of lithium-ion (Li^+) being solvated by carbonate molecules. Reconstructed based on Figure 1 in ref 35 with the permission from the Electrochemical Society 2025.

report positive (“successes”) while ignoring negative datasets (“failures”).¹⁰ This bias is driven by the evaluation and merit systems for scientific researchers, as well as the novelty-rewarding incentive of modern publication industry, but it constitutes significant potential risks for AI4Science models because these models must learn both positive and negative results in order to accurately simulate, infer, and predict; (c) Higher complexity arise from battery in particular, which, being a system itself, consist of multiple components. Thus, the electrochemical performances of batteries are the outcome of the summation of physical and chemical interactions among all of these components,¹¹ and any seemingly insignificant perturbation in one of the many parameters involved, be it cathode or anode composition, percentage of additives, or even the process history of these materials, would produce completely different results.^{12,13} Furthermore, among the multiple components within a battery, electrolyte is the most unique one because it has to interface with each and every other component, whose sensitivity to above-mentioned perturbations would be significantly magnified.¹⁴

With AI/machine-learning (ML) models increasingly adopted in diverse science domains, the importance of datasets become ever more pressing. Among them, AI4Batteries are especially data hungry. AI/ML techniques have been applied on various aspects of battery technologies,¹⁵ mostly manufacturing optimization, quality control,^{16,17} or time-series type of predictions on system level.^{18–21} Some efforts were also dedicated to materials development, but they all remained on the level of formulation optimization for targeted properties using existing molecules or materials.^{22–24} Only very limited attempts were made at designing new materials or structures assisted by AI,^{25–29} whose small footprint in the landscape of general AI excitement well echoes with the small chemical design space explored thus far by the battery research community: Based on an internal analysis on 15,000 highly cited battery articles published during 2015–2025, we found that both the unique chemicals (solvent, additive, and salt

anions) and the electrolyte formulations made with those chemicals are on the order of 10^{2-3} (Figure 1).

On the other hand, encouraging achievements have emerged in AI-designed materials, one prominent example of which was the utilization of deep learning in pharmaceutical industry, where a new structural class of antibiotics were discovered, with broad-spectrum and narrow-spectrum activity against some of the world’s most problematic bacterial pathogens.^{30,31}

Amid the rapid rise of AI4Science, it has become increasingly clear that a comprehensive scientific database of high quality remains absent. As a subset of such a database, a molecular database that contains all possible molecules in the universe along with key properties would serve as the firm foundation for the training of AI/ML models that will eventually benefit all industry sectors that rely on small organic molecules, spanning from batteries to catalysts, biomedicines, lubricants, coolants and refrigerants, etc.

2. ENUMERATING THE UNIVERSE OF MOLECULES

Since long, people have been wondering how many molecules that could exist in the universe. Obviously the number must be infinity if one places no limit on the number of atoms in the molecule. More than two decades ago, Bohacek et al. attempted to make such an estimate under constraints to explore possible design space for drugs.³² By applying Lipinski’s rule of five, confining the number of heavy atoms (i.e., any element above hydrogen) under 30, and further limiting the element selections to C, O, N, and S structured in fewer than 4 rings and fewer than 10 branch points, they came up with an astonishing number of 10^{60} , which is the natural consequence of the combinatorial explosion. This number is already far greater than the total number of stars in the entire observable universe (10^{24}).³³ A more accurate number came from a well-established commercial database GDB-17,³⁴ where the number of heavy atoms is further constrained to be under 17, while the elements were expanded to C, O, N, S, and X ($\text{X} = \text{F}, \text{Cl}, \text{Br}, \text{I}$). Under such constraints 1.66×10^{11} unique molecules can be enumerated, which is still more than the number of stars in our galaxy, The Milky Way (Figure 1). Comparison between the total number of possible small

Table 1. Chemistry Valence Laws for Enumerating the Comprehensive Structural Database

Chemistry Valence Grammar	Elementary Building Blocks
Neutral Species	
F: I (monovalent)	
C or Si: IV (tetravalent)	
O: II (bivalent)	
B or N: III (trivalent)	
S: II (bivalent), IV (tetravalent) and VI (hexavalent)	  
P: III (trivalent) and V (pentavalent)	 
Ionic Species (cationic, anionic and zwitterionic)	
C: III (trivalent, e.g., carbocation or carbanion)	 
N: II (bivalent, e.g. imide) or IV (tetravalent, e.g. ammonium)	 
O: III (trivalent, e.g. oxygenium)	
S: III (trivalent, e.g. sulfonium)	
P: IV (tetravalent, e.g. phosphonium)	

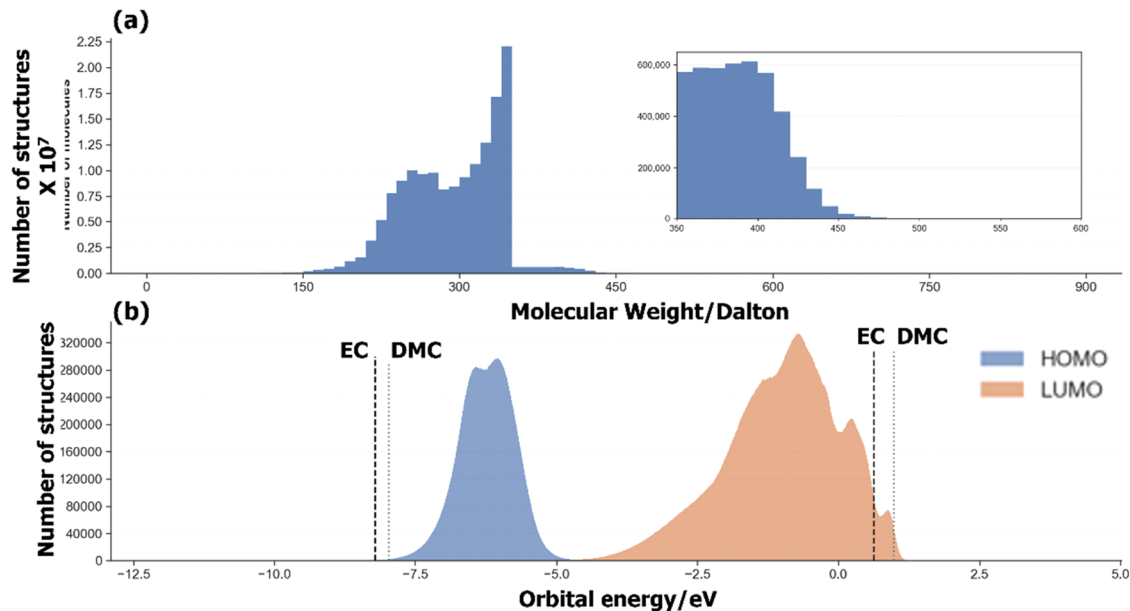


Figure 2. Overview of the molecules in *Molecular Universe*: (a) Distribution of molecular weights. The sharp drop-off at 350 is related to the imposed upper limit of heavy atom number (up to 20) in order to prevent combinatorial explosion. Some molecules with heavy atom numbers >20 but eligible as low-concentration additives are also included, resulting in nonzero counts at higher molecular weights (inset); (b) Distribution of frontier molecular orbital (MO) energies of these molecules. The HOMO and LUMO values for two common solvents, ethylene carbonate (EC) and dimethyl carbonate (DMC) are indicated by dashed vertical lines.

molecules to the vastness of universe is by no means exaggeration, as both involve truly astronomical numbers.

Apparently, the above constraints of both the heavy atom number and element selection are still too restrictive for application scenarios beyond medicine. The organic materials

that support our daily life, i.e., natural and synthetic solvents, additives, ionic compounds, oligomers, and particularly polymers, often consist of much more diverse elements as well as many more heavy atoms. Considering that any relaxation of these constraints will doubtlessly lead to exponential growth of

the molecule numbers, we believe that balancing the comprehensiveness of the databases and the excessively large numbers (such as 10^{60}) beyond our capability to handle is critical, and one must tailor the number of heavy atoms and the selection of elements according to specific needs from each application scenario.

For the domain of battery electrolyte materials, we construct a comprehensive database for small organic molecules,^{35,36} which was designed to contain *all* possible *small* organic molecules in the universe as allowed by the chemistry valence laws. Of course “all” and “small” represent our efforts in making the above-mentioned balance. First, to ensure its comprehensiveness, we enumerate the structural database as the foundation by covering all possible combinations of elements under known chemistry grammar (Table 1), so that all possibly existing molecular structures are exhaustively generated. Second, to prevent combinatorial explosion and ensure its applicability on the domain of battery electrolyte materials, we further modify these two constraints on heavy atom number and element:

- (1) The heavy atom numbers are restricted to be up to and including 20 because all electrolyte components (solvents, additives, salt anions) operate based on the properties brought by the chemical nature of their functional groups, hence as the skeleton to carry these functional groups or their combinations, a molecule does not have to be over 350 Da, which, by average, stay within the atom number of ~ 20 . This judgment is verified from the list of electrolyte materials already explored by human beings in the past three decades.^{11,37–39}
- (2) Prior knowledge of these electrolyte materials' elements indicates that while C, N, O, S, and F are necessary, Cl, Br, and I often introduce electrochemical instability to the molecules containing them because their low oxidation limits often lead to harmful shuttling effects.³⁷ On the other hand, a few more elements on the second or third row of the Periodic Table are frequent components of the electrolyte materials, such as Si, B, and P. Thus, the new element selection tailored for this specific domain should be C, N, O, S, F, Si, P, and B.

The total number of molecules thus enumerated is ca. 2 trillions (2×10^{12}). It is necessary to emphasize here that, while this structural database should exhaustively encompass all possible combinations of the defined elements under the given constraints, the generation of these structures from an algorithm considers only their “mathematic” correctness, while their chemical correctness, such as whether they are chemically or thermodynamically stable, or their feasible synthesizability, was never in the context.

To further reduce the number to a manageable level, we applied a series of structural filters to remove those molecules obviously not qualified as electrolyte building blocks (Figure 2), such as

- (1) The molecule should not be protic; therefore, any linkage consisting of a hydrogen connected with a heteroatom is excluded, such as $-\text{O}-\text{H}$, $-\text{N}-\text{H}$, $-\text{S}-\text{H}$, $-\text{P}-\text{H}$, $-\text{B}-\text{H}$, or $-\text{Si}-\text{H}$.
- (2) The molecule should not display any redox capability itself, hence structures or substructures that can be used in organic cathode or anode materials are excluded, examples of which are quinones, anilines, polyoxygen ($-\text{O}-\text{O}-$), polysulfide ($-\text{S}-\text{S}-$), or other long heteroatom linkages.

- (3) The molecule should possess at least certain molecular polarity so that it can mix with other components, especially ions from the dissociated salts, to form a homogeneous solution; hence, molecules containing a singular element as heavy atom are excluded. This rule mainly eliminates hydrocarbon olefins with no heteroatoms because carbon, among all elements, is the best at connecting with each other and forming long chains, but it also excludes oligomeric silanes, phosphines, and boranes.
- (4) The molecule should not possess unstable substructural units, such as small-membered (3 or 4) rings, and unstable heteroatom combinations such as F directly sitting on O or N. This constraint mainly comes from empirical knowledge accumulated from organic chemistry, but it should be noted that there are rare exceptions in electrolyte additives where 3- or 4-membered ring structures could be allowed as unstable sacrificial components.

Application of these multiple filter layers reduced the number to $\sim 10^9$, which was then subject to a high-throughput workflow that conducted density-function-theory (DFT) calculations on these molecules. DFT calculations were made possible by various acceleration techniques including powerful GPUs and optimized codes and algorithms, whose combination can process up to 9 million structures per day. Such speed and scale was unimaginable only 5 years ago. For example, using a single A100 GPU or H100 GPU, at the same precision, a molecule with approximately 35 atoms (consisting of C, H, O, F, S, N) typically requires only 70–160 s vs $\sim 10^3$ s if the calculation is conducted on a conventional cluster based on central-processing-units (CPUs).⁴⁰ Overall, the higher dollar-per-hour value of GPUs leads to approximately 90% reduction in computation time and cost for most DFT tasks.

Eventually ca. 200 million stable molecules that include both neutral and ionic species were fully calculated, which now serves as the core foundation for the battery materials section of the *Molecular Universe*,³⁶ which is, to the best of our knowledge, the largest DFT database available so far. DFT calculations provided up to ~ 20 basic properties, including the energy levels of the highest-occupied-molecular-orbitals (HOMO) and the lowest-unoccupied-molecular-orbitals (LUMO), maximum and minimum of electrostatic potentials (ESP min/max), electron affinity, ionization potential, bond dissociation energies, defluorination energies, heat of combustion, polarizability, dipole moment and atomic partial charges, etc. The vast databases of these molecular characteristics support the development and training of various AI/ML models now operating in *Molecular Universe*. Majority of the calculated molecules in the *Molecular Universe* are neutral species (ca. 200 million), which are candidates for solvents, additives, and diluents in battery electrolytes, while ca. 1 million are anionic species that can be candidates for the development of new salts after coupling with appropriate working ions (Li^+ , Na^+ , K^+ , Zn^{2+} , etc.).

Figure 2a,2b summarize the distributions of these 200 million calculated molecules and their HOMO/LUMO energy levels.

In *Molecular Universe*, neutral species are loosely classified into 19 distinct groups, with each group dominated by a unique functional substituent.³⁶ One million molecules randomly selected from these neutral species are visualized by a dimension-reduction technique, *Uniform Manifold Approximation and Projection* (UMAP),⁴¹ which groups them into clusters

based on their structural similarity (Figure 3a). Those widely scattered clusters in the UMAP reflect the rich structural

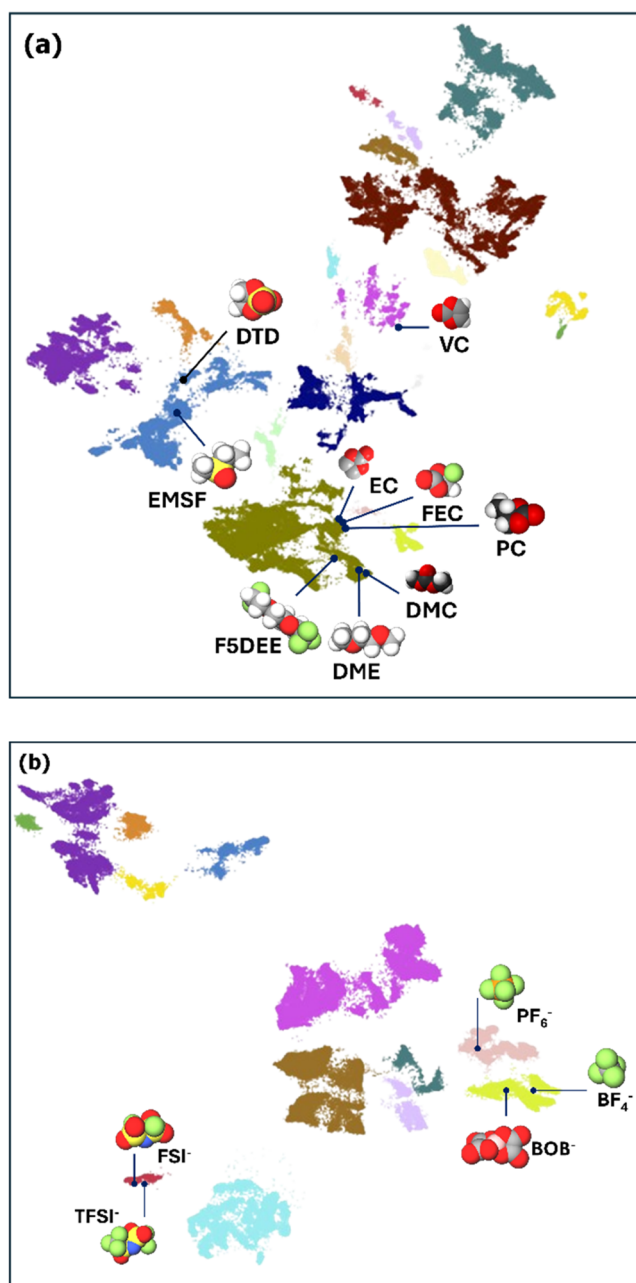


Figure 3. (a) UMAP for 1.0 million neutral molecules randomly sampled from the *Molecular Universe*. Structures of selected known electrolyte solvents and additives are labeled along with dashed lines to indicate their positions on the map. (b) UMAP for the 1.0 million anionic species. Structures of a few known salt anions are labeled along with dashed lines to indicate their positions on the map.

diversity of these molecules. A similar UMAP was also produced for the 1 million anionic species (Figure 3b). Compared with neutral species, both the total number of and the actually explored anionic structures are much smaller.

There is little doubt that such an astronomically sized DFT database would bring tremendous benefits to the materials discovery. Its significance goes far beyond a simple collection of immense molecular structures and properties that have never been available before, but more importantly, it serves as a firm

foundation and rich resource to train, fine-tune, and evaluate various AI/ML models so that one could gain unprecedented insights into the quantitative structure–property relationship that has been at the center of materials design but remains elusive, especially for complicated systems such as battery electrolytes.

It should be noted that the above structural database as well as the chemical filters that led to the 200 million DFT calculations were conducted solely for the application domain of battery electrolytes. While such a database already covers a vast number of molecules that could be used for other domains such as refrigerants, antibiotics, or photoresist agents in semiconductors, a more thorough coverage of molecules for these domains requires reconstruction of the structural database as well as the chemical filters because the structural linkages favored by each specific domain significantly vary.

3. NAVIGATING THE MOLECULAR UNIVERSE: CHEMISTRY-INFORMED LLMs

One of the greatest challenges in managing astronomically sized databases is not the lack of candidate molecules, but the overwhelming abundance of them, and consequently, the difficulty of finding the most relevant structures for a specific application in a fast and accurate manner. The term “relevant structures” here may imply satisfying requirements from many different dimensions, including domain-specific knowledge about which properties dictate that application, the availability of such property data, and the structure–property relationships. In this vast molecular landscape, a chemically informed LLM can serve not merely as a predictor, but as an intelligent navigator that guides exploration toward the most promising regions.

Unlike conventional LLMs trained primarily on natural language, a chemically informed LLM is trained on a multimodal representation of chemistry, including molecular structures represented either by 1D codes like SMILES (Simplified Molecular Input Line Entry System), 2D molecular graphs, or 3D geometries, as well as datasets from DFT and MD simulations, reaction pathways, experimental measurements, patents, and scientific literature, etc. Such a model develops an internal representation of the relationships among the molecular structure, physicochemical properties, synthesis, and function. Instead of searching for molecules in databases through rigid similarity metrics or predefined descriptors, it can reason across multiple levels of chemical knowledge, identifying molecules that satisfy complex and sometimes competing requirements, such as electrochemical stability, ionic conductivity, solubility, synthesizability, cost, safety, and environmental compatibility.

The role of the chemically informed LLM is analogous to that of an experienced explorer navigating an immense, largely unmapped territory.⁴² Instead of exhaustively evaluating every possible molecule, the model interprets scientific objectives, formulates search strategies, prioritizes candidate structures and properties, and explains the rationale behind its recommendations. In effect, it transforms molecular discovery from brute force screening into knowledge-guided exploration.

This navigation capability will become even more powerful when integrated with foundation models, high-throughput simulations, and autonomous laboratories (A-labs).^{43,44} The LLM can orchestrate an iterative discovery cycle by proposing new molecules, selecting computational workflows, interpreting simulation and experimental results, refining design principles, and generating the next generation of candidates. As the molecular database now grows from 10^6 to 10^{12} compounds, the

MU0.5
GPT4.5
DeepSee
Grok-4

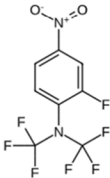
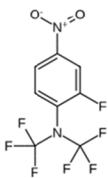
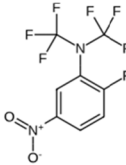
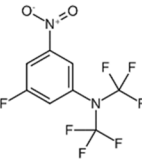
Ground Truth	Grok 4.0	DeepSeek 2.0	GPT 4.5	MU 0.5
additive <chem>O=[N+](=O)c1ccc(N(C(F)(F)F)C(F)(F)F)c(F)c1</chem>	<chem>Fc1cc(N+([O-])cc1N(C(F)(F)F)C(F)(F)F</chem> invalid -1.0	<chem>FC1=C(C=C(C=C1)[N+](=O)[O-])N(C(F)(F)F)C(F)(F)F</chem> 1.0	<chem>FC1=C(C=C(C=C1)[N+](=O)[O-])N(C(F)(F)F)C(F)(F)F</chem> 1.0	<chem>c1c(F)cc([N+](=O)[O-])cc1N(C(F)(F)F)C(F)(F)F</chem> -0.5
				
<chem>O=[N+](=O)c1ccc(N(C(F)(F)F)C(F)(F)F)c(F)c1</chem>	<chem>Fc1cc(N+([O-])cc1N(C(F)(F)F)C(F)(F)F</chem>	<chem>FC1=C(C=C(C=C1)[N+](=O)[O-])N(C(F)(F)F)C(F)(F)F</chem>	<chem>FC1=C(C=C(C=C1)[N+](=O)[O-])N(C(F)(F)F)C(F)(F)F</chem>	<chem>c1c(F)cc([N+](=O)[O-])cc1N(C(F)(F)F)C(F)(F)F</chem>

Figure 4. Chemistry rigor performance displayed by the chemically informed LLM in Molecular Universe (MU 0.5) in comparison with other general LLMs. Upper panel: the overall performance of LLMs in generating SMILES codes for ca. 300 chemicals extracted from the battery literature, where green, yellow, and red marks the correct SMILES, SMILES with minor hallucinations in structures, and completely wrong SMILES that violates chemistry valence laws, respectively; Middle panel: the zoom-in view of the only hallucinated molecular structure by MU-0.5 and its comparison with other LLMs.

chemically informed LLM becomes indispensable and perhaps the only reliable interface between human scientific intent and the immense universe of molecules.

For example, in *Molecular Universe*, such LLMs pretrained with chemistry, materials and battery literature (articles, patents, handbooks, and chemical catalogs) on the scale of 10^7 have shown remarkable chemistry knowledge in the test of “chemical accuracy.” In one test, more than 300 chemicals names extracted from ca. 15,000 battery articles published between 2015 and 2025 were used as input, and the LLMs were asked to generate the correct SMILES. These chemicals have been reported in various battery electrolyte formulations as solvents, cosolvents, additives, diluents, or salt anions, and their nomenclatures almost never follow the standardized IUPAC convention, and certainly never ever SMILES, but instead adopt rather chaotic mixtures of conventions (as electrolyte scientists often do), ranging from commercial names, custom names, trivial names, and acronyms. In comparison with general LLMs including ChatGPT, Grok, and DeepSeek, the SMILE strings generated by the chemically informed LLM in MU performed with the highest accurately, only making a single mistake where a minor hallucination occurred in the relative position of the substitutions (Figure 4).

Besides the above chemistry rigor test, extensive benchmarking was also carried out on *Molecular Universe* against other mainstream LLMs in many aspects, including asking battery or electrolyte-related questions, scoring the inferencing process, and evaluating the generated answers and suggestions. A similar ranking always ensued for all LLMs tested. It was such chemical literature collection, which includes articles, textbooks, chemical handbooks, chemical catalogs, and selected patent literature in the number of more than 17 million, when used as training materials and contents of RAG (retrofit-augmented generation), that ensured the scientific rigor of the chemically informed performance of LLMs and other AI/ML models in the *Molecular Universe*.

4. DESIGN BY PROPERTIES: ELECTROLYTE FOUNDATIONAL MODEL

Looking at the modern history of material discovery, one would find that most of the time people are designing new materials based on existing materials already known to perform by mimicking their structure. Although this trend is not specific in organic material development but universal in all fields, it is especially pronounced in the designing of small organic molecules for diverse application scenarios of drug discovery, organic electronic materials, catalysis, and electrolyte,¹¹ where structural derivation is naturally the most popular avenue adopted by human researchers. Considering that the ultimate goal of developing new materials is better property and performance, not new structures, such structural derivation approach is by no means effective and ideal. The underlying reasons for the popular “structural derivation” approach are diverse, but two major factors stand out:

- (1) Molecular structures, whether represented in 2D chemical formulas or 3D spatial models, are inherently visual-spatial representations that the human brain can process efficiently through dedicated visual and spatial cognition systems, much similar to an analog mode. In contrast, molecular properties are typically represented as abstract numerical, multidimensional, or symbolic variables that lack an intuitive spatial organization, placing greater demands on working memory and abstract reasoning, which are normally much more complicated for the human brain to process.^{45,46}
- (2) Although the structure–property relationship constitutes the center of materials science, quantitative structure–property relationships have been established for only a limited subset of materials and properties.^{47,48} For most complex materials systems, especially those involving multicomponent compositions, hierarchical structures, metastable phases, or dynamic interfaces (all of which are encountered by batteries and especially electrolytes materials), the mapping from structure to properties remains rarely understood and difficult to quantify.

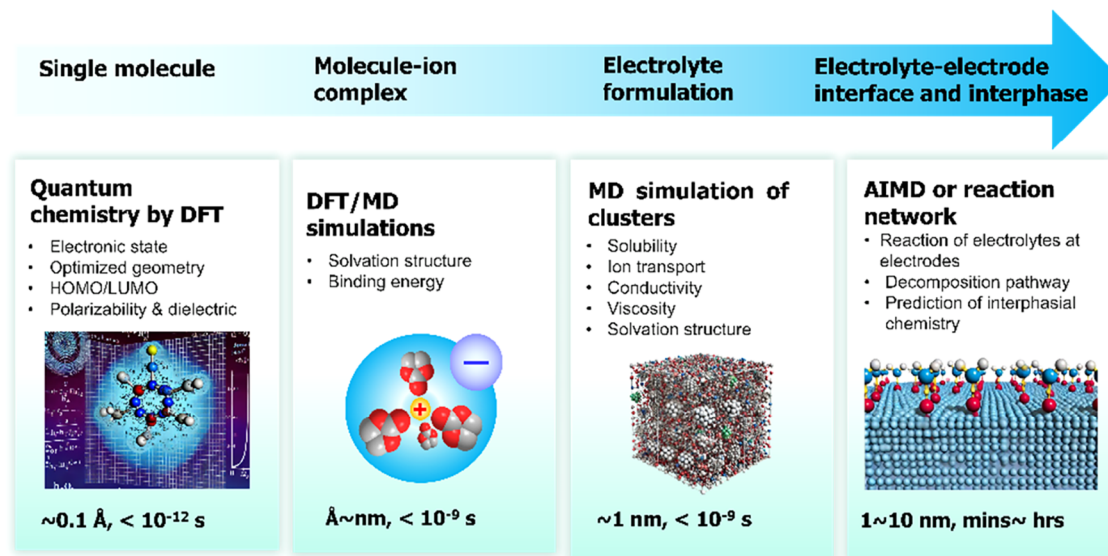


Figure 5. Multiscale *in-silico* prediction of electrolyte properties is still challenging. The state-of-the-art computation and simulation techniques can accurately calculate the properties of single molecules in vacuum, the ion-solvation, and simulate the long-range structure of the large assembly of ions and molecules as consequence of their interactions, and even their distribution and arrangements on a charged interface, but cannot maintain such accuracy once charge-transfer occurs, nor can predict the eventual products and their subsequent evolution in battery environment in long time-scale (10^{-6} s up to hours and days).

Consequently, the concept of “rational design” can only be partially rational.^{49,50} Rather than deriving optimal materials directly from predictive theories, researchers typically rely on a combination of established physical principles, simplified models, empirical correlations, accumulated intuition, and iterative experimentation. In practice, this process frequently involves educated guesses (or “guesstimates”) that are progressively refined through repeated synthesis, characterization, and testing, which was often referred to as “Edisonian”. The success of many historical materials discoveries has therefore depended more on experimental experience and serendipity than first-principles understanding, where direct design targeting certain property goals has been extremely rare.

Under such context, we believe that the success of AI4Materials would be best measured by whether it can allow researchers to design new materials directly by properties, a capability never accessible to us thus far. The feasibility of such a goal depends on efforts in two related and complementary fronts:

- (1) How far can the tools based on first-principles, such as computation and simulations, approach the behaviors of materials in real devices? While nowadays, properties of single molecules or single-crystal lattices can be calculated or simulated with quantum accuracies, what really determines the properties and performances of the actual materials are the physical and chemical behaviors of these molecules or crystal lattices in condensed phases, whose computation and simulation remain challenging because of the complexity in how they interact with each other. Such challenges are especially pronounced for molecules in battery electrolytes, whose performance is dictated not only by their interaction with other components in the electrolyte bulk phase but also by the sacrificial reactions they participate at interfaces to form a passivation component known as interphase.^{11,14} So far, despite the most recent advances in computation and simulation techniques (DFT, molecular dynamics (MD) and *ab*

initio MD (AIMD), machine-learning interatomic potentials, reactive or polarizable force fields, chemical reaction networks, etc.), predicting chemical reactions in long time-scales ($>10^{-6}$ s) remains costly and difficult to scale up (Figure 5). These approaches based on first-principles do not seem practical for high-throughput calculations/simulations in the foreseeable future, but disruptive changes could happen anytime due to the emergence of transformative technologies in both algorithm (advances in computation and force field) and hardware (advances in the computing power of chips), as well as the availability of large and open-sourced data sets accumulated, thanks to the efforts in recent years.

- (2) Where the first-principles tools fail, is there an alternative pathway to circumvent?

If we treat the complicated systems (battery as example) as a “black box,” with molecular information (properties from DFT) or condensed phase information (properties from MD and AIMD simulations) as “Inputs” and cell performance as “Outputs,” then it is possible to construct ML or foundational models that directly link the inputs and the outputs, which is the *de facto* structure–property relationship. Of course, such models require large quantity of high-quality data to pretrain and fine-tune. With the trained model, one does not need to understand the underlying mechanisms of physics and chemistry happening in that black box. The structure–property relationship involved is in fact already encoded, analyzed, and identified in the latent space of the models, although we may never know exactly what the relationships are.

Given the complexity of battery systems, these two approaches are being simultaneously explored by researchers in the AI4Materials domain. Among various efforts, the construction of battery foundation models that directly connect molecular and formulation information with cell performance

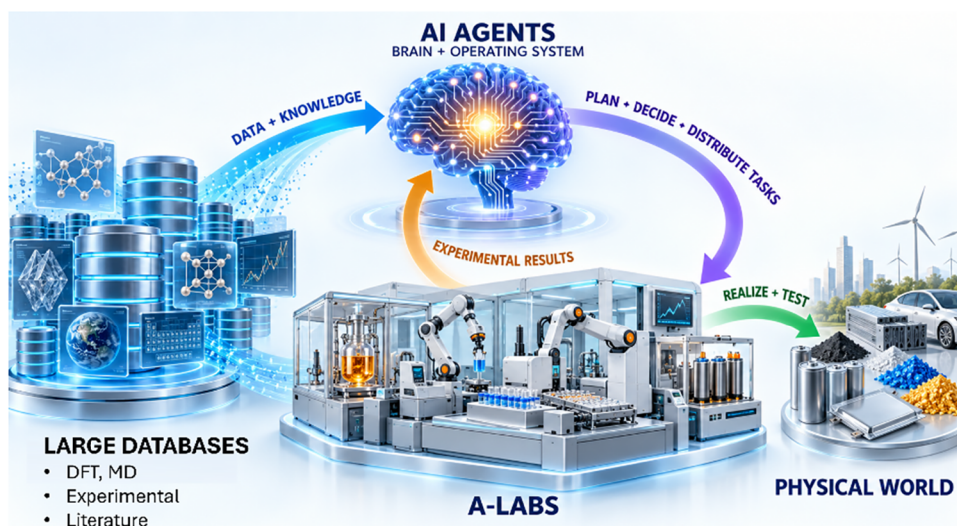


Figure 6. Large databases provide rich data and information to AI agents that function as the brain (operating system) to plan, decide, and distribute tasks, while A-labs serves as the key infrastructure connecting AI to the real physical world by providing key capabilities in performing these tasks and then sending the results back to AI for systematic analysis.

represents a promising way to solve the challenge of designing materials by properties.

Below we use the “Electrolyte Foundational Model” constructed for the *Molecular Universe* as a case study to discuss how “structure-properties” relationship can be established for complex systems such as electrolytes. As discussed above, from properties of single molecules (solvent, additive, diluents and anions etc.), their assemblies in condensed phase (electrolytes, ion-solvation, ion–solvent interaction, ion transport, etc.) and on interfaces (interfacial structure and assembly under electrode potential, adsorption, ion-desolvation under polarization, etc.), to interphases (reactions, products generation and dissipation, chemistry and morphologies of the solid depositions) and finally in devices, there were multiple gaps and dimensions where none of those properties are simple additive of the properties in the previous layer, and computation/simulations based on first-principles can only predict small portion of those at extremely limited scale, spatio- and temporal-lengths. What the foundational models do is to encode the properties of single molecules and electrolyte formulations, producing a surprisingly rich array of physics relevant to device performance. While fine-tuning on specific chemical domains or experimental targets is typically required for design-level predictive accuracy, a well-trained foundation model incorporating the physics of the raw materials can significantly reduce the amount of fine-tuning required for specific downstream prediction tasks. A concrete example is a “Electrolyte Foundation Model” trained in a hierarchical stage:

- (1) The first “stage” of the model, termed “single-molecule embedding,” maps a molecular representation (molecular structure, DFT-derived properties relevant to battery operation) onto a latent space. The encoder is trained by maximizing predictive accuracy for the DFT properties.
- (2) The second “stage” of the model maps the weighted embeddings obtained from the latent space of the previous stage for each electrolyte ingredient onto a single electrolyte-level embedding. The second stage is subsequently encoded and trained in a similar fashion, i.e., maximizing predictive accuracy for solution-level proper-

ties, ideally those properties that can be derived from MD simulations.

The electrolyte-level embeddings then become a starting point for a wide variety of downstream prediction tasks. While immediate prediction of device-level properties is rarely accurate, fine-tuning a predictive model on even just ~ 20 experimental data points can improve predictive accuracy to a substantial degree. For example, a model trained in such a fashion to predict the capacity retention rate can attain R^2 values of ~ 0.9 when comparing model-predicted capacity retention rates to experimentally measured rates for a hold-out set. Absent the physics-informed embedding, which is generated entirely from simulated data, the number of data points required to reach that level of accuracy would be significantly larger. Thus, the main impact of single-molecule and solution-level properties is ultimately to dramatically reduce, if not eliminate, the amount of experimental data required for fine-tuning. While no rational framework for designing models like this yet exists, we find empirically that radically different domains typically require different fine-tuning data sets, one case being that separate fine-tuning exercises are required for different battery cell chemistries (Li-metal vs 12% Si in graphitic anodes, for example). The need for frequent generation of small, per-domain fine-tuning datasets highlights the value of pairing these modeling procedures with an A-labs capable of quickly furnishing data for fine-tuning, which will be discussed in more detail in the next section.

One advantage of this approach is that the early stages of the foundational model can be very broad and encompass a wide range of chemistries, restricting the per-domain retraining to only the last stage of the downstream prediction task. Naturally, as results of this comprehensiveness, the downstream prediction models are not perfectly accurate and are best paired with mechanisms for quantifying uncertainty. Examples include kernels based on molecular similarity metrics (such as Tanimoto similarity) or averaging model predictions across models trained from different random initializations. The authors of this work emphasize that even imperfect models can be incredibly useful for narrowing down a formulation space, as we typically apply such models to rank the top N candidates from a large

formulation space, building in thresholds and cutoffs that allow for some uncertainty to exist in the model predictions. The net result is still a model that allows us to rule out most candidates with a high degree of confidence, assuming that uncertainties are treated with appropriate flexibility.

Together with “Electrolyte Foundational Model,” generative AI (GenAI) models such as diffusion models and reinforcement learning (RL) frameworks pretrained with the databases can design new molecules *de novo* after learning the chemical representation of small organic molecules and their properties. Integrating these advanced AI models can eventually lead us to the goal of designing properties without reliance on exhaustive structural enumerations and DFT calculations.

5. A-LABS: CONNECTING AI TO THE PHYSICAL WORLD AND THE ASSOCIATED CHALLENGES

What distinguishes AI4Science from GenAI models for language, vision, and code processing is its inseparable connection to the physical world. While an LLM can generate text, picture, or video that appears plausibly correct, the success of a scientific hypothesis is ultimately measured not by statistical likelihood but by rigorously obeying laws of physics and chemistry and being creative at the same time. As such, every predicted molecule, material, or reaction must withstand the experimental verification. As such, AI4Science cannot rely solely on virtual results but requires a continuous feedback loop between the computation and reality. Taking *Molecular Universe* as an example, since the majority of the structures have never been touched by human beings, its effectiveness requires massive and repeated experimental validation, which involves synthesis, structural characterization, and electrochemical testing, followed by feeding back into a learning process. This closed-loop interaction transforms AI from a passive and isolated predictor to an active participant in scientific discovery.

It was under such a context that A-labs have emerged as the critical bridge connecting AI to the physical world, which is built on a hybrid human-AI framework and capable of handling the scale and speed required for the new materials design and discoveries.^{44,51–53} A-labs integrate robotics, automated synthesis, high-throughput characterization, and cloud-connected instrumentation, while an agentic AI models serve as their operating systems, which, based on the datasets and information provided as inputs from the databases, makes decisions, plans experiments, and distributes tasks to the different robotic modules in the A-lab (Figure 6). A well-tuned A-lab can execute thousands of experiments with minimal human intervention, and it works around the clock as long as machinery and consumables sustain. More importantly, they generate standardized, reproducible, and richly annotated experimental data, the very resource that has historically limited progress in many scientific disciplines, including AI4Science itself. Instead of relying primarily on historical literature, AI systems can continuously acquire fresh experimental evidence, rapidly validate hypotheses, identify unexpected phenomena, and iteratively refine the predictive models. This self-improving cycle of design, synthesis, characterization, learning, and redesign dramatically accelerates the pace of discovery.

The advantage of A-labs in generating highly standardized experimental data is of especial importance for the battery domain. A battery is never a single material but a system consisting of multiple materials including both active (cathode and anode) and inactive components (electrolyte, separator, binder, conductive additive, cell packaging, etc.), which must

interface with each other to synchronize the functions. The eventual performance of a battery is thus extremely sensitive to any minor disturbances caused by variations in the parameters of these materials, such as composition and formulation, vendor source, history of storing and processing, technology of integration and cell engineering, cell configuration, and testing protocol, etc. For this reason, cell performance reported in the literature, even for identical materials, is often inconsistent and incomparable because of many known and unknown sources of interference induced by different individual researchers. Such data cannot be directly used to train AI models. A-labs are expected to provide the solution to this challenge by standardizing all experiments under highly uniform conditions, whose reproducible, diverse, and well-controlled data sets would be of high value for AI4Materials.

Despite these anticipated benefits, substantial obstacles still exist preventing the immediate and extensive deployment of A-labs. Aside from the manufacturing cost of hardware and their interfacing protocols with various AI models, the most significant is the technology maturity in the construction of high-throughput synthesis modules. While the technology of A-labs for high-throughput optimization rapidly matures, most of these modules still rely on existing materials as starting ingredients, where the design and execution of experiments are actually dealing with high-throughput mixing. The real challenge for A-labs is the materials synthesis, no matter if the target materials are new crystalline lattices (electrode materials, solid electrolytes, etc.) or new organic molecules. There have been encouraging reports on the former, partially because the synthesis protocols of inorganic materials are relatively simpler, whose starting ingredients are often simple inorganic salts, and the related synthetic routes often consist of the three highly uniformed elements of “mixing–pressurizing–calcination.”⁴⁴

The synthesis of small organic molecules, however, is completely at another level of difficulty. First, the synthetic routes for organic structures are highly dynamic. In most instances, even the identical structure could be synthesized via completely different routes and completely different starting materials; Second, the complex structures of most organic molecules require multistep synthesis, with each step requiring separation and purification that involve distillation, recrystallization, sublimation, filtration, etc., which are often conducted in large volumes of solvents. The yield of the overall synthesis process is therefore rather low, and the ratio of product to waste materials is often at horrific odds; Third, gaseous, volatile and/or corrosive agents are often used as starting materials in synthesizing organic molecules, especially when most new electrolyte solvents, anions, and additives favor fluorination, such synthesis hence often requires special conditions, such as high temperature, high pressure, and special reaction vessels; Last but not least, unlike special chemical domains such as drug discovery, bioagents, or dyes, high-throughput synthesis for battery electrolytes must deliver sufficient quantity of products (in gram scales) so that subsequent characterization in coin cells or pouch cells can be supported. The combination of all these factors makes the A-lab module for automated organic synthesis a challenge of nightmarish difficulty. So far, the only example ever reported in the literature is a system for exploring dye-like molecules, where the apparatus driven by ML planning and analysis algorithm iteratively synthesizes various structural derivations of a fixed chemical skeleton, with only three selected properties (absorption on substrate, lipophilicity, and photo-oxidative stability) characterized at the end of loop.⁵³ Obviously,

this capability is still very far from exploring the chemical space on the order of 10^8 , with no constraints on parental structures.

Nevertheless, despite these challenges, the real value of A-Labs lies in the possibility of accessing a vast majority of possible compounds that have never been synthesized or characterized during the several centuries of modern chemistry. The feasibility and reliability of AI-recommended structures will heavily depend on how AI models can assist in evaluating their synthesizability and subsequent design of synthetic routes and conditions. In this sense, the future of AI4Science will not be solely determined by larger models or more powerful computing software and hardware but equally by the ability to connect those models to the physical world through autonomous experimentation, where A-Labs is the essential infrastructure.

6. PERSPECTIVES: CHALLENGES FOR AI4SCIENCE

The current AI frenzy has brought significant anxieties to the general public that artificial general intelligence (AGI) will replace human jobs in the coming years, while a similar perception also increasingly prevails in scientific communities, especially in 2026 which has been considered *annus mirabilis* by some and *annus horribilis* by others because of the rapidly growing capabilities of AI agents, as represented by the events and drama around the solution of millennium mathematic problems.^{54,55} As AI continues to permeate other scientific domains, its future impact will be defined as much by its breakthroughs as by its boundaries. As of today, AI models excel at pattern recognition across vast datasets at amazing speed, scope, and scale, but they not only remain limited by the quality and completeness of training datasets available but also fundamentally confined to those fields where rules, or physical and chemical laws, are well established. Thus, we have seen that AI models can perform such tasks like chess-playing, code-writing and solving mathematic problems at speed and efficiency that a human brain can never rival, but have not yet discovered new, unknown physics or chemistry laws, despite scattered reports where AI recognized or rediscovered certain known laws based on large experimental datasets.^{56–58}

There have been traditional arguments that AI models like LLMs are always trying to predict the next token based on statistics, using the features extracted in high-dimension space without truly understanding what language or science really is. In this sense, the absence of true understanding, and a reliance on interpolation rather than genuine novelty, would prevent AI from discovering new laws that govern the operation of our physical world. The important discoveries made by human scientists were seldom driven by large quantities of data; rather, the most brilliant discoveries were often triggered by sparks of ingenuity and serendipity, even when data sets are scarce, as witnessed by Newton, Maxwell and Einstein. Other people, represented by Hinton, believed that LLMs understand language in a way not fundamentally different from the human brain, as both are based on converting words into high-dimensional feature vectors that interact contextually (“shake hands”) to form meaning and predict the next elements.^{59,60} If this argument stands, it should apply to science, and hence, we cannot exclude the possibility that new scientific laws being discovered by AI in the future.

Nevertheless, for the foreseeable future, we believe that AI4Science is unlikely to eliminate the need for scientific reasoning, experimentation, or human creativity. Instead, they will evolve from predictive tools into scientific partners capable of generating hypotheses, designing experiments, and iteratively

refining knowledge under human supervision, validation, and tuning. The greatest opportunities of AI4Science may not exist in replacing human scientists but in enabling them to explore previously inaccessible regions of scientific space, from billions and trillions of unexplored molecules and materials to entirely new chemistries and physical phenomena. However, given the dynamic pace of the field of AI4Science intersected by scientific domains, AI, and computation technologies (algorithms and hardware), we caution that the above belief could be overthrown at any unexpected moment.

AUTHOR INFORMATION

Corresponding Authors

Stanley Whittingham – Materials Science & Engineering, Binghamton University, Binghamton, New York 13902, United States; orcid.org/0000-0002-5039-9334; Email: stanwhit@gmail.com

Martin Winter – Helmholtz Institute Münster, Forschungszentrum Jülich, 48149 Munster, Germany; MEET Battery Research Center, University of Münster, 48149 Munster, Germany; Email: martin.winter@uni-muenster.de

Y. Shirley Meng – Pritzker School of Molecular Engineering, University of Chicago, Chicago, Illinois 60637, United States; Energy Storage Research Alliance, Argonne National Laboratory, Lemont, Illinois 60439, United States; orcid.org/0000-0001-8936-8845; Email: shirleymeng@uchicago.edu

Qichao Hu – SES AI Corp., Woburn, Massachusetts 01801, United States; Molecular Universe Pte, Singapore 048542, Singapore; Email: qichao@ses.ai

Kang Xu – SES AI Corp., Woburn, Massachusetts 01801, United States; Molecular Universe Pte, Singapore 048542, Singapore; Department of Mechanical and Aerospace Engineering, The Ohio State University, Columbus, Ohio 43210, United States; orcid.org/0000-0002-6946-8635; Email: kang.xu@ses.ai

Authors

Daniel Hannah – SES AI Corp., Woburn, Massachusetts 01801, United States; Molecular Universe Pte, Singapore 048542, Singapore

Yumin Zhang – SES AI Corp., Woburn, Massachusetts 01801, United States; Molecular Universe Pte, Singapore 048542, Singapore

Complete contact information is available at: <https://pubs.acs.org/10.1021/jacsau.6c01250>

Notes

The authors declare the following competing financial interest(s): DH, YZ, QH and KX of SES AI Corp. are developers of the *Molecular Universe* that was referred to as an example in this manuscript. *Molecular Universe* is a commercial AI-platform designed to discover and calculate new molecules, simulate electrolyte formulations without pre-set constraints, and predict behaviors of electrolyte in actual cells and the cell performances such as cycle-life fading.

REFERENCES

- (1) How Big Is the Internet. www.ipxo.com/blog/how-big-is-the-internet (accessed Sept 12, 2026).

- (2) How Much Data Exists in the World?. <https://thebestvpn.com/statistics/how-much-data-exists-in-the-world/> (accessed Sept 12, 2026).
- (3) Protein Data Bank. www.rcsb.org (accessed Sept 12, 2026).
- (4) AlphaFold. <https://alphafold.com/> (accessed Sept 12, 2026).
- (5) PubMed. <https://pubmed.ncbi.nlm.nih.gov/> (accessed Sept 12, 2026).
- (6) Web of Science. www.webofscience.com (accessed Sept 12, 2026).
- (7) Scopus. www.scopus.com (accessed Sept 12, 2026).
- (8) arXiv. <https://arxiv.org/> (accessed Sept 12, 2026).
- (9) Popper, K. *The Logic of Scientific Discovery*; Routledge, 2005.
- (10) Ioannidis, J. P. A. Why Most Published Research Findings Are False. *PLoS Med.* **2005**, *2*, No. e124.
- (11) Xu, K. *Electrolytes, Interfaces and Interphases: Fundamentals and Applications in Batteries*; Royal Society of Chemistry Press, 2023.
- (12) Bird, L.; Shao, Y.; Pang, B.; Robinson, J.; Shearing, P. Establishing a living systematic review of characterisation and parameter reporting in lithium-ion and lithium-sulfur cathode research. *Discov. Electrochem.* **2026**, *3*, No. 35.
- (13) Puls, S.; Nazmutdinova, E.; Kalyk, F.; et al. Benchmarking the reproducibility of all-solid-state battery cell performance. *Nat. Energy* **2024**, *9*, 1310.
- (14) Meng, Y. S.; Srinivasan, V.; Xu, K. Designing better electrolytes. *Science* **2022**, *378*, No. eabq3750.
- (15) Lombardo, T.; Duquesnoy, M.; El-Bouysidy, H.; Årén, F.; Gallo-Bueno, A.; Jørgensen, P. B.; Bhowmik, A.; Demortière, A.; Ayerbe, E.; Alcaide, F.; Reynaud, M.; Carrasco, J.; Grimaud, A.; Zhang, C.; Vegge, T.; Johansson, P.; Franco, A. A. Artificial Intelligence Applied to Battery Research: Hype or Reality? *Chem. Rev.* **2022**, *122*, 10899–10969.
- (16) Liu, K.; Wei, Z.; Zhang, C.; Shang, Y.; Teodorescu, R.; Han, Q.-L. Towards Long Lifetime Battery: AI-Based Manufacturing and Management. *IEEE/CAA J. Autom. Sin.* **2022**, *9*, 1139–1165.
- (17) Niri, M. F.; Liu, K.; Apachitei, G.; Román-Ramírez, L. A. A.; Lain, M.; Widanage, D.; Marco, J. Quantifying key factors for optimized manufacturing of Li-ion battery anode and cathode via artificial intelligence. *Energy AI* **2022**, *7*, No. 100129.
- (18) Liu, X.; Zhang, X.-Q.; Chen, X.; Zhu, G.-L.; Yan, C.; Huang, J.-Q.; Peng, H.-J. A generalizable, data-driven online approach to forecast capacity degradation trajectory of lithium batteries. *J. Energy Chem.* **2022**, *68*, 548–555.
- (19) Wang, S.; Zhou, R.; Ren, Y.; Jiao, M.; Liu, H.; Lian, C. Advanced data-driven techniques in AI for predicting lithium-ion battery remaining useful life: A comprehensive review. *Green Chem. Eng.* **2025**, *6*, 139–153, DOI: 10.1016/j.gce.2024.09.001.
- (20) Severson, K. A.; Attia, P. M.; Jin, N.; Perkins, N.; Jiang, B.; Yang, Z.; Chen, M. H.; Aykol, M.; Herring, P. K.; Fraggadakis, D.; Bazant, M. Z.; Harris, S. J.; Chueh, W. C.; Braatz, R. D. Data-driven prediction of battery cycle life before capacity degradation. *Nat. Energy* **2019**, *4*, 383–391.
- (21) Wu, B.; Widanage, W. D.; Yang, S.; Liu, X. Battery digital twins: Perspectives on the fusion of models, data and artificial intelligence for smart battery management systems. *Energy AI* **2020**, *1*, No. 100016.
- (22) Narayanan Krishnamoorthy, A.; Wolke, C.; Diddens, D.; Maiti, M.; Mabrouk, Y.; Yan, P.; Grunebaum, M.; Winter, M.; Heuer, A.; Cekic-Laskovic, I. Data-Driven Analysis of High-Throughput Experiments on Liquid Battery Electrolyte Formulations: Unraveling the Impact of Composition on Conductivity. *Chem.-Methods* **2022**, *2*, No. e202200008.
- (23) Dave, A.; Mitchell, J.; Burke, S.; Lin, H.; Whitacre, J.; Viswanathan, V. Autonomous optimization of non-aqueous Li-ion battery electrolytes via robotic experimentation and machine learning coupling. *Nat. Commun.* **2022**, *13*, No. 5454.
- (24) Zhu, S.; Ramsundar, B.; Annevelink, E.; Lin, H.; Dave, A.; Guan, P.-W.; Gering, K.; Viswanathan, V. Differentiable modeling and optimization of non-aqueous Li-based battery electrolyte solutions using geometric deep learning. *Nat. Commun.* **2024**, *15*, No. 8649.
- (25) Ahmad, Z.; Xie, T.; Maheshwari, C.; Grossman, J. C.; Viswanathan, V. Machine Learning Enabled Computational Screening of Inorganic Solid Electrolytes for Suppression of Dendrite Formation in Lithium Metal Anodes. *ACS Cent. Sci.* **2018**, *4*, 996–1006.
- (26) Zhang, Q.; Zhou, C.; Zhang, D.; Kramer, D.; Peng, C.; Xue, D. Data-driven discovery and intelligent design of artificial hybrid interphase layer for stabilizing lithium-metal anode. *Matter* **2023**, *6*, 2950–2962.
- (27) Kim, S. C.; Oyakhire, S. T.; Athanitis, C.; Wang, J.; Zhang, Z.; Zhang, W.; Boyle, D. T.; Kim, M. S.; Yu, Z.; Gao, X.; Sogade, T.; Wu, E.; Qin, J.; Bao, Z.; Bent, S. F.; Cui, Y. Data-driven electrolyte design for lithium metal anodes. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, No. e2214357120.
- (28) Gao, Y.-C.; Yao, N.; Chen, X.; Yu, L.; Zhang, R.; Zhang, Q. Data-Driven Insight into the Reductive Stability of Ion–Solvent Complexes in Lithium Battery Electrolytes. *J. Am. Chem. Soc.* **2023**, *145*, 23764–23770.
- (29) Szymanski, N. J.; Rendy, B.; Fei, Y.; Kumar, R. E.; He, T.; Milsted, D.; McDermott, M. J.; Gallant, M.; Cubuk, E. D.; Merchant, A.; Kim, H.; Jain, A.; Bartel, C. J.; Persson, K.; Zeng, Y.; Ceder, G. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624*, 86–91.
- (30) Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackerman, 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.; Collins, J. J. A deep learning approach to antibiotic discovery. *Cell* **2020**, *180*, 688–702.
- (31) Liu, G.; Catacutan, D. B.; Rathod, K.; Swanson, K.; Jin, W.; Mohammed, J. C.; Chiappino-Pepe, A.; Syed, S. A.; Fraggis, M.; Rachwalski, K.; Magolan, J.; Surette, M. G.; Coombes, B. K.; Jaakkola, T.; Barzilay, R.; Collins, J. J.; Stokes, J. M. Deep learning-guided discovery of an antibiotic targeting *Acinetobacter baumannii*. *Nat. Chem. Biol.* **2023**, *19*, 1342–1350.
- (32) Bohacek, R. S.; McMartin, C.; Guida, W. C. The art and practice of structure-based drug design: A molecular modeling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
- (33) NASA. Star Basics. <https://science.nasa.gov/universe/stars/> (accessed Sept 12, 2026).
- (34) Ruddigkeit, L.; van Deursen, R.; Blum, L. C.; Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **2012**, *52*, 2864–2875.
- (35) Hannah, D.; Zhang, Y.; Lig, X.; et al. Search for ideal electrolytes in the Molecular Universe. *Electrochem. Soc. Interface* **2025**, *34*, 35.
- (36) The Molecular Universe. www.molecular-universe.com/map (accessed Sept 12, 2026).
- (37) Xu, K. Nonaqueous liquid electrolytes for lithium-based rechargeable batteries. *Chem. Rev.* **2004**, *104*, 4303.
- (38) Xu, K. Electrolytes and interphases in Li-ion batteries and beyond. *Chem. Rev.* **2014**, *114*, 11503.
- (39) Kim, S. C.; Oyakhire, S. T.; Athanitis, C.; et al. Data-driven electrolyte design for lithium metal anodes. *Proc. Natl. Acad. Sci. U.S.A.* **2023**, *120*, No. e2214357120.
- (40) Wu, X.; Sun, Q.; Pu, Z.; Zheng, T.; Ma, W.; Yan, W.; Yu, X.; Wu, Z.; Huo, M.; Li, X.; Ren, W.; Gong, S.; Zhang, Y.; Gao, W. Enhancing GPU-Acceleration in the Python-Based Simulations of Chemistry Framework. 2024, arXiv:2404.09452. arXiv.org e-Print archive. <https://arxiv.org/abs/2404.09452>.
- (41) McInnes, L.; Healy, J.; Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. 2018, arXiv:1802.03426. arXiv.org e-Print archive. <https://arxiv.org/abs/1802.03426>.
- (42) Zhang, D.; Liu, W.; Tan, Q. et al. ChemLLM: A Chemical Large Language Model. 2024, arXiv:2402.06852. arXiv.org e-Print archive. <https://arxiv.org/abs/2402.06852>.
- (43) Liu, S.; Dong, D.; Xiao, Z.; et al. Machine learning in electrolyte design: Accelerating discovery and multidimensional optimization for next-generation batteries. *Chem* **2026**, *12*, No. 103024.
- (44) Szymanski, N. J.; Rendy, B.; Fei, Y.; et al. An autonomous laboratory for the accelerated synthesis of novel materials. *Nature* **2023**, *624*, 86.

- (45) Kravitz, D. J.; Saleem, K. S.; Baker, C. I.; et al. A new neural framework for visuospatial processing. *Nat. Rev. Neurosci.* **2011**, *12*, 217.
- (46) Nieder, A. The calculating brain. *Physiol. Rev.* **2025**, *105*, 267.
- (47) Callister, W. D.; Rethwisch, D. G. *Materials Science and Engineering: An Introduction*, 9th ed.; Wiley, 2014.
- (48) Miller, K. D.; Campbell, M.; Puggioni, D.; et al. Decoratypes: An extensible crystal taxonomy for machine learning-guided materials discovery. *Phys. Rev. Mater.* **2025**, *9*, No. 104404.
- (49) Yu, Z.; Wang, H.; Kong, X.; Huang, W.; Tsao, Y.; Mackanic, D. G.; Wang, K.; Wang, X.; Huang, W.; Choudhury, S.; Zheng, Y.; Amanchukwu, C. V.; Hung, S. T.; Ma, Y.; Lomeli, E. G.; Qin, J.; Cui, Y.; Bao, Z. Molecular design for electrolyte solvents enabling energy-dense and long-cycling lithium metal batteries. *Nat. Energy* **2020**, *5*, 526–533.
- (50) Xu, J.; Zhang, J.; Pollard, T. P.; Li, Q.; Tan, S.; Hou, S.; Wan, H.; Chen, F.; He, H.; Hu, E.; Xu, K.; Yang, X.-Q.; Borodin, O.; Wang, C. Electrolyte design for Li-ion batteries under extreme operating conditions. *Nature* **2023**, *614*, 694–700.
- (51) Fei, Y.; Rendy, B.; Kumar, R.; et al. AlabOS: a Python-based reconfigurable workflow management framework for autonomous laboratories. *Digital Discovery* **2024**, *3*, 2275.
- (52) Abolhasani, M.; Kumacheva, E. The rise of self-driving labs in chemical and materials sciences. *Nat. Synth.* **2023**, *2*, 483.
- (53) Koscher, B. A.; Canty, R. B.; McDonald, M. A.; et al. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science* **2023**, *382* (6677), No. eadi1407.
- (54) On the Navier–Stokes Millennium Prize Problem. <https://openai.com/index/navier-stokes-solution/> (accessed Sept 12, 2026).
- (55) 25 Fields Medal Winners Including Terence Tao Sign Declaration Saying Rapid AI Proofs Are Harming Math in Severe Misalignment. <https://officechai.com/ai/25-fields-medal-winners-including-terence-tao-sign-declaration-saying-rapid-ai-proofs-are-harming-math-in-severe-misalignment/> (accessed Sept 12, 2026).
- (56) Fang, Y.-L.; Jian, D. S.; Li, X. et al. AI-Newton: A Concept-Driven Physical Law Discovery System Without Prior Physical Knowledge. 2025, arXiv:2504.01538. arXiv.org e-Print archive. <https://arxiv.org/abs/2504.01538>.
- (57) Srivastava, K.; Dash, S.; Cory-Wright, R. et al. Bridging the Gap Between Scientific Laws Derived by AI Systems and Canonical Knowledge via Abductive Inference with AI-Noether. 2025, arXiv:2509.23004. arXiv.org e-Print archive. <https://arxiv.org/abs/2509.23004>.
- (58) Yu, W.; Abdelaleem, E.; Nemenman, I.; et al. Physics-tailored machine learning reveals unexpected physics in dusty plasmas. *Proc. Natl. Acad. Sci. U.S.A.* **2025**, *122*, No. e2505725122.
- (59) Hinton, G. Digital Intelligence vs. Biological Intelligence, Speech at Royal Institute. <https://youtu.be/B-SOMe-FGTs?si=ndadM2lvQx8ltWeA>.
- (60) Mischler, G.; Li, Y. A.; Bickel, S.; et al. Contextual feature extraction hierarchies converge in large language models and the brain. *Nat. Mach. Intell.* **2024**, *6*, 1467.