A Foundational Molecular Formulation Database for High-Precision Prediction of Mixture Properties

Anonymous Author(s)
Affiliation
Address
email

Abstract

Predicting the properties of multi-component mixtures is a fundamental challenge in chemistry and materials science. Unlike single-molecule systems, mixture behavior emerges from nonlinear interactions and excess properties, making additive rules ineffective. Existing datasets are sparse, fragmented, and lack negative results or standardized metadata, limiting machine learning (ML) models that generalize across formulation spaces. We propose the **Foundational Molecular Formulation Database**, an open dataset generated using modular self-driving laboratories (SDLs) for automated, high-throughput experimentation. Spanning four domains—battery electrolytes, thermofluids, fragrances, and solution-processed semiconductors—the dataset captures key functional properties (e.g., ionic conductivity, volatility, stability) with structured metadata. It enables ML benchmarks in **property prediction, generative design, active learning, and cross-domain transfer**, establishing a foundation for data-driven formulation science analogous to ImageNet or AlphaFold in their fields.

1 Introduction

ML for molecular design has transformed domains such as proteins [1] and drug formulations [2], but chemical mixtures remain a bottleneck. Mixture properties—conductivity, volatility, viscosity, film stability—arise from nonlinear interactions and cannot be inferred from single components. Existing resources are sparse and inconsistent, while the combinatorial design space ($\sim 10^{55}$ binary mixtures) makes systematic exploration infeasible without automation.

Recent advances in **self-driving laboratories (SDLs)**—robotic platforms guided by ML—have enabled autonomous electrolyte optimization [3, 4] and are broadly transforming chemistry [5]. We leverage this paradigm to build a large, standardized mixture dataset across critical domains.

2 Motivation and Impact

Electrolytes require mixture and electrode-interface properties rather than only molecular data [6]. Thermofluids suffer from limited characterization (only 8 refrigerants and 4 HFO blends) despite urgent low-GWP needs [7, 8, 9]. Fragrance datasets ($\sim$5k molecules) are subjective, inconsistent, and lack physicochemical metadata [10, 11, 12]. Semiconductor additive studies remain piecemeal, with little negative data and no standardized reporting [13, 14]. The proposed dataset will enable researchers to predict the functional, perceptual, and performance properties of multi-component mixtures from their compositions and constituent properties. Mapping a given formulation to emergent macroscopic behavior remains a pertinent unsolved problem across disciplines — from electrochemistry and materials science to psychophysics. The key challenge is all the functionality is determined by the excess property, the deviation from linear mixing. Most mixtures exhibit highly

Submitted to AI for Science workshop (NeurIPS 2025)." Do not distribute.

non-linear response: for example, the conductivity of an electrolyte, volatility of a fragrance accord, or viscosity of a thermofluid. To approximate this nonlinear, often discontinuous, manifold of mixture properties machine learning models need many examples across a wide parameter range, which necessitates an expansive and dense dataset.

Property prediction surrogates trained on this dataset could be used in high-throughput screening pipelines accelerating the exploration of high-dimensional, combinatorial mixture space. Surrogate-driven screening can simultaneously evaluate multiple objectives and map Pareto fronts, enabling informed design decisions to be made in an accelerated and cost effective manner. Additionally, probing clusters in the embedding space of models trained on large datasets make it easier to find safer or cheaper substitutions for existing formulations.

3 Dataset Description

Domains: (a) Electrolytes—conductivity, viscosity, stability; (b) Thermofluids—thermal conductivity, vapor pressure; (c) Fragrances—volatility, odor descriptors; (d) Semiconductors—film conductivity, mechanical stability. **Collection:** Flow-through SDLs share a backbone (liquid handling, balances, temperature control) with domain-specific modules.

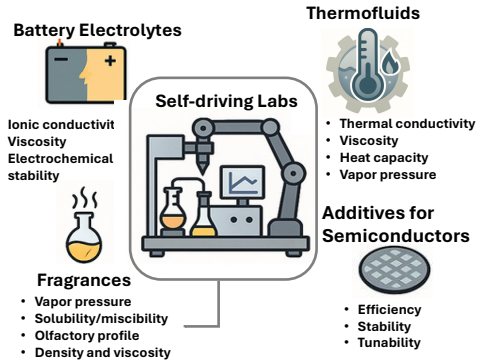

Figure 1: Modular self-driving lab for foundational mixture dataset collection.

4 Tasks and Benchmarks

The mixture dataset collected through our SDLs will supports four core ML tasks:

1. Property Prediction: Predict emergent properties (e.g., conductivity, viscosity, volatility) from composition and metadata. Metrics: RMSE, MAE, ranking for multi-objective scenarios.

2. Generative Design: Inverse design of mixtures to meet target properties under constraints (e.g., maximize conductivity, minimize cost). Metrics: success rate, diversity, computational efficiency.

3. Active Learning: Propose experiments to efficiently explore combinatorial space under budget constraints. Metrics: error reduction vs. iterations, sample efficiency.

4. Cross-Domain Transfer: Benchmark zero-shot and fine-tuned performance across domains. Metrics: transfer accuracy, adaptation speed.

Standardized splits, baseline models, and leaderboards will ensure reproducibility.

5 Ethical Considerations

Data is generated in controlled environments with open licensing (CC-BY 4.0). No human or sensitive data is involved.

6 Limitations

The dataset initially focuses on SDLs for liquid-phase mixtures and may underrepresent rare additive chemistries. Future work will expand to more domains and testing conditions.

A Data Card

Feature	Description
Domains	Electrolytes, Thermofluids, Fragrances, Semiconductors
Properties	Conductivity, viscosity, volatility, stability, spectra
Format	CSV with metadata JSON
License	CC-BY 4.0
Collection Method	Automated self-driving lab workflows

B Detailed Description of Domains

B.1 Battery Electrolytes

Electrolytes are central to battery performance, yet their design remains challenging. Unlike single-molecule systems, properties depend on **mixture interactions** and **electrode interfaces**, not molecular data alone [6]. Linear mixing rules fail because key behaviors arise from *excess properties* and cross-term effects. Recent work combining differentiable mixture models (e.g., DiffMix [4]) with robotic experimentation shows small changes in the formulation can dramatically shift ionic conductivity [6]. With an astronomical design space (10^{55} binary mixtures), systematic exploration is infeasible without standardized datasets and ML. A foundational database will enable predictive design of electrolytes optimized for conductivity, stability, safety, and cost.

B.2 Thermofluids

Next-generation HVAC systems require **low-GWP refrigerants**, yet no single fluid meets both performance and regulatory demands [8]. Mixtures offer a path forward, but current databases cover only eight fully characterized fluids and four HFO blends [7, 9]. This lack of systematic data across composition space slows innovation and deployment. A comprehensive thermofluid dataset will close this gap and accelerate sustainable refrigerant design [15].

B.3 Fragrances

Predicting odor from structure remains unsolved across chemistry and neuroscience. Odor perception is nonlinear: small chemical changes can dramatically alter scent, while unrelated molecules may smell alike [11]. Existing datasets ($\sim$5k molecules) are subjective, inconsistent, and lack physicochemical metadata [10, 12]. A large, standardized dataset with chemical coverage and structured descriptors will enable ML for predictive olfaction and electronic nose applications in health, food, and environment [16].

B.4 Semiconductors

Solution-processed semiconductors—metal oxides, conjugated polymers, perovskites—benefit from additives that enhance performance and stability in OLEDs, OFETs, OECTs, and solar cells [17, 18, 19]. Yet progress is hindered by fragmented studies, lack of negative data, and no standardized metadata. Using a self-driving lab, we will build a large, high-quality dataset of additive/semiconductor mixtures spanning broad parameter space. ML-guided sampling will uncover principles governing additive effects, accelerating design of next-generation semiconductors with tunable properties.

C Market Opportunity of the Four Domains

Domain	2023–2024 Value	Projected 2030–2032 Value	CAGR
Battery Electrolytes [20]	\$12.1 B	\$35.0 B	$\sim$14.2%
Thermofluids [21]	\$11.1 B	\$14.2 B	$\sim$3.7%
Fragrances [22]	\$23.4 B	\$35.0 B	$\sim$6.8%
Semiconductor Additives [23]	\$11.2 B	\$17.6 B	$\sim$6.7%

D Required Instrumentation List for SDLs Development

Battery Electrolytes: viscometer, balance, potentiostat.
Thermofluids: viscometer, thermal conductivity probe, vapor pressure sensor.
Fragrances: microbalance (for volatility), refractometer
Semiconductor Additives: spin coater, semiconductor parameter analyzer, contact angle goniometer, FTIR spectrometer, viscometer

E Estimation of the cost per data point

As stated in the main text, SDLs enable high-throughput experimentation, and ML algorithms enable scientists to optimize experiments that include multiple parameters and conditions. However, ML models require very large experimental datasets in order to be trained. Hence, in order to assess the feasibility of acquiring large mixture datasets with SDLs it is important to estimate the cost involved in the process.

In this section, we take as an example the field of solution-processed semiconductors, and estimate the cost per data point by doing back-of-the-envelope calculations. We use a typical semiconductor (C14-PBTTT) and a typical additive (F4-TCNQ), dissolved in a typical solvent (1,2-Dichlorobenzene), with microscope slides used as substrate. The costs per unit were calculated using the prices listed in mainstream vendors (Merck, TCI Chemicals, Fisher). As shown in the following table, even with a conservative estimate of 10 samples per experimental session, the cost of consumables per sample is less than a dollar. This is because of the small quantities of materials required to fabricate an individual sample. In order to deal with integers, we round this number up to ($\$0.25$). Assuming a budget of $\$100,000$, this amounts to $400,000$ samples.

Component	Cost/unit	Quantity/exp	Samples/exp	Quantity/sample	Cost/sample
Semiconductor	\$2/mg	500 ug	10	50 ug	\$0.1
Additive	\$1.5/mg	10 ug	10	1 ug	\$0.0015
Solvent	\$1/mL	500 uL	10	50 uL	\$0.05
Substrate	\$0.65/slide	1	10	10 samples/slide	\$0.065

The cost per data point can be approximated if we account for the number of properties to be measured and the number of measurements per sample. Assuming that 4 properties will be measured and each property measurement will be performed a total of 3 times for reproducibility reasons, this results in 12 measurements per sample, and hence in $4,800,000$ measurements, whose square root (since we consider binary mixtures) results in approximately 2191 data points. Dividing the budget with this number results in an approximate cost of $\$45$ per data point.

References

[1] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[2] Pauric Bannigan, Matteo Aldeghi, Zeqing Bao, Florian Häse, Alán Aspuru-Guzik, and Christine Allen. Machine learning directed drug formulation development. *Advanced Drug Delivery Reviews*, 175:113806, 2021.

[3] Adarsh Dave, Jared Mitchell, Sven Burke, Hongyi Lin, Jay Whitacre, and Venkatasubramanian Viswanathan. Autonomous optimization of non-aqueous Li-ion battery electrolytes via robotic experimentation and machine learning coupling. *Nature Communications*, 13(1):5454, September 2022. Publisher: Nature Publishing Group.

[4] Shang Zhu, Bharath Ramsundar, Emil Annevelink, Hongyi Lin, Adarsh Dave, Pin-Wen Guan, Kevin Gering, and Venkatasubramanian Viswanathan. Differentiable modeling and optimization of non-aqueous Li-based battery electrolyte solutions using geometric deep learning. *Nature Communications*, 15(1):8649, October 2024. Publisher: Nature Publishing Group.

[5] Connor W. Coley, Florian Häse, Klavs F. Jensen, and Alán Aspuru-Guzik. Self-driving laboratories for chemistry and materials science. *Chemical Reviews*, 124(12):6467–6536, 2024.

[6] Austin D. Sendek and Venkatasubramanian Viswanathan. Artificial intelligence for electrolyte design: Going beyond the molecular paradigm. *Electrochemical Society Interface*, 34(2):39–43, 2025.

[7] Laura Fedele, Gerard Lombardo, Ilaria Greselin, Davide Menegazzo, and Sergio Bobbo. Thermophysical properties of low gwp refrigerants: An update. *International Journal of Thermophysics*, 44(Article 80), 2023.

[8] James M. Calm. The next generation of refrigerants – historical review, considerations, and outlook. *International Journal of Refrigeration*, 38:18–43, 2013.

[9] Ian Bell, Demian Riccardi, Ala Bazyleva, and Mark O. McLinden. Survey of data and models for refrigerant mixtures containing halogenated olefins. *Journal of Chemical & Engineering Data*, 66, 2021.

[10] Andreas Keller, Richard C. Gerkin, Yuanfang Guan, Amit Dhurandhar, Gabor Turu, Bence Szalai, Joel D. Mainland, Yusuke Ihara, Chung Wen Yu, Russ Wolfinger, Celine Vens, Leander Schietgat, Kurt De Grave, Raquel Norel, Gustavo Stolovitzky, Guillermo A. Cecchi, Leslie B. Vosshall, and Pablo Meyer. Predicting human olfactory perception from chemical features of odor molecules. *Science*, 355(6327):820–826, feb 2017.

[11] Brian K. Lee, Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie A. Little, Matthew A. Andres, Britney B. Nguyen, Theresa Moloy, Jacob Yasonik, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko. A principal odor map unifies diverse tasks in olfactory perception. *Science*, 381(6661):999–1006, 2023.

[12] E. Darío Gutiérrez, Amit Dhurandhar, Andreas Keller, Pablo Meyer, and Guillermo A. Cecchi. Predicting natural language descriptions of mono-molecular odorants. *Nature Communications*, 9(1):4979, 2018.

[13] Yuan Cai, Jian Cui, Ming Chen, Miaomiao Zhang, Yu Han, Fang Qian, Huan Zhao, Shaomin Yang, Zhou Yang, Hongtao Bian, Tao Wang, Kunpeng Guo, Molang Cai, Songyuan Dai, Zhike Liu, and Shengzhong (Frank) Liu. Multifunctional Enhancement for Highly Stable and Efficient Perovskite Solar Cells. *Advanced Functional Materials*, 31(7):2005776, 2021.

[14] Jingfu Chen, Jiefeng Luo, Enlong Hou, Peiquan Song, Yuqing Li, Chao Sun, Wenjing Feng, Shuo Cheng, Hui Zhang, Liqiang Xie, Chengbo Tian, and Zhanhua Wei. Efficient tin-based perovskite solar cells with trans-isomeric fulleropyrrolidine additives. *Nature Photonics*, 18(5):464–470, 2024.

[15] E. W. Lemmon, I. H. Bell, M. L. Huber, and M. O. McLinden. Nist standard reference database 23: Reference fluid thermodynamic and transport properties–refprop, version 10.0, 2018.

[16] Rinu Chacko, Deepak Jain, Manasi Patwardhan, Abhishek Puri, Shirish Karande, and Beena Rai. Data based predictive models for odor perception. *Scientific Reports*, 10(1):17136, 2020.

[17] Mark Nikolka, Katharina Broch, John Armitage, David Hanifi, Peer J. Nowack, Deepak Venkateshvaran, Aditya Sadhanala, Jan Saska, Mark Mascal, Seok-Heon Jung, Jin-Kyun Lee, Iain McCulloch, Alberto Salleo, and Henning Sirringhaus. High-mobility, trap-free charge transport in conjugated polymer diodes. *Nature Communications*, 10(1):2122, 2019.

[18] Pegah Ghamari, Muhammad Rizwan Niazi, and Dmytro F. Perepichka. Improving Environmental and Operational Stability of Polymer Field-Effect Transistors by Doping with Tetranitrofluorenone. *ACS Applied Materials & Interfaces*, 15(15):19290–19299, 2023.

[19] David Ohayon, Lucas Q. Flagg, Andrea Giugni, Shofarul Wustoni, Ruipeng Li, Tania C. Hidalgo Castillo, Abdul-Hamid Emwas, Rajendar Sheelamanthula, Iain McCulloch, Lee J. Richter, and Sahika Inal. Salts as Additives: A Route to Improve Performance and Stability of n-Type Organic Electrochemical Transistors. *ACS Materials Au*, 3(3):242–254, 2023.

[20] MarketsandMarkets. Battery electrolyte market size, share and trends analysis report, 2024.

[21] Allied Market Research. Heat transfer fluids market size, share and forecast, 2024.

[22] Fortune Business Insights. Fragrance market size, share and industry forecast, 2024.

[23] TechNavio. Semiconductor process chemicals market analysis report, 2024.