Foundation Models for Discovery and Exploration in Chemical Space

Anoushka Bhutani
Alexius Wadell (Mentor) / Venkat Viswanathan (Advisor)
abhutani@umich.edu

SC26 Chicago, IL | hpc unites.

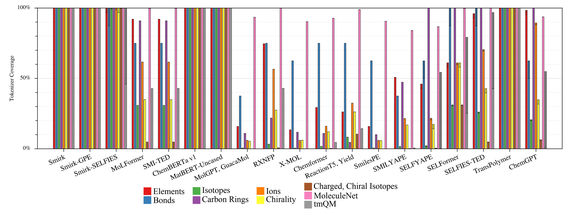
1 Rethinking Tokenization for Molecular Foundation Models

/BattModels/smirk JCIM Paper

Text-based foundation models become an important part of scientific discovery, with molecular foundation models accelerating advancements in material science and molecular design. However, existing models are constrained by closed-vocabulary tokenizers that capture only a fraction of molecular space. To overcome the limitations of existing tokenizers, we propose two new tokenizers—Smirk and Smirk-GPE—with full coverage of the OpenSMILES specification. The proposed tokenizers systematically integrate nuclear, electronic, and geometric degrees of freedom, facilitating applications in pharmacology, agriculture, biology, and energy storage.

Why did we need a new tokenizer?

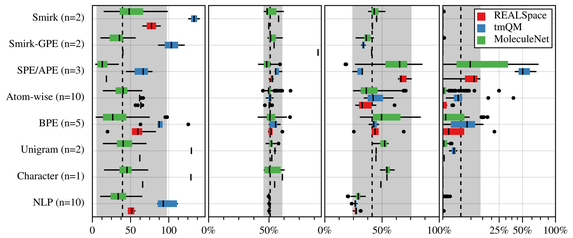
We systematically evaluate 35 tokenizers, including 20 chemistry-specific tokenizers, and reveal significant gaps in their coverage of the SMILES molecular representation.



Fraction of molecules tokenized without `[UNK]` across tokenizers and datasets. Atom-wise tokenization performs well on common benchmarks (MoleculeNet), but fails on datasets rich in bracketed atoms (tmQM). Because OpenSMILES admits over 28 trillion bracketed atoms, vocabulary growth is not a viable fix for atom-wise tokenizers.

Evaluating Tokenizer Quality before Pre-training

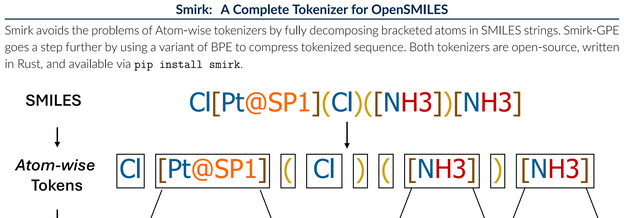
Full pre-training and downstream evaluation are expensive. That makes extrinsic tokenizer comparisons hard to run at scale. Intrinsic metrics aim to predict a tokenizer's impact on overall performance before training.



Intrinsic metrics were designed for open-vocabulary tokenizers and therefore miss the cost of `[UNK]`. On tmQM, for example, SPE/APE achieve low fertility (avg. tokens per molecule), but half of those tokens are `[UNK]`. Preferred directions: fertility (↓), imbalance (↓), η (↑), and `[UNK]` frequency (↓).

Smirk: A Complete Tokenizer for OpenSMILES

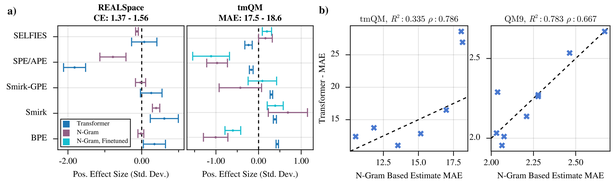
Smirk avoids the problems of Atom-wise tokenizers by fully decomposing bracketed atoms in SMILES strings. Smirk-GPE goes a step further by using a variant of BPE to compress tokenized sequence. Both tokenizers are open-source, written in Rust, and available via `pip install smirk`.



Smirk is effectively a character-level tokenizer over the OpenSMILES glyph set rather than Unicode. Its 165-token vocabulary requires no training and, by construction, can faithfully tokenize any OpenSMILES-encoded molecule.

N-Grams Metrics Are Prognostics

- Intrinsic metrics miss the cost of `[UNK]`, and full transformer evaluations are costly.
- We use N-grams as a cheap proxy for extrinsic performance.
- N-grams predict the next token from the previous $n-1$ tokens.
- They train cheaply, parallelize well, and preserve the same next-token prediction task as transformers.
- Using linear fixed-effects models we found N-gram metrics predictive of the impact of tokenizer choice and molecular encoding on downstream performance.



Standardized effect sizes for tokenizers and molecular encodings, where ↑ is better. Smirk matches or exceeds the performance of existing tokenizers, while retaining chemically interpretable tokens. (b) N-gram metrics (cross-entropy, information loss) linearly predict downstream performance. They complement expensive extrinsic evaluations and low-fidelity intrinsic metrics for tokenizer evaluation.

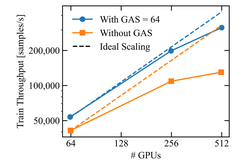
2 Training a Billion Parameter Foundation Model

/BattModels/mist Preprint

MIST: Molecular Insight SMILES Transformer



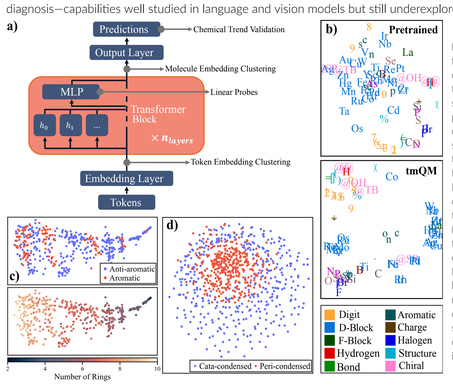
- MIST is a family of molecular foundation models which learn transferable chemical representations from molecular structures.
- MIST models match or exceed state-of-the-art performance across benchmarks spanning physiology to quantum chemistry.
- They are encoder-only transformers trained with masked language modeling on SMILES sequences tokenized with Smirk.
- The largest model has 1.8B parameters and was pretrained on 2B molecules.
- It was trained on 40 NVIDIA A100 GPUs for ~7 days using Distributed Data Parallelism.
- During fine-tuning the molecular embeddings are processed by small task networks to predict properties.



Enabling Gradient Accumulation Steps (GAS) was key to efficiently scaling MIST to 128 Polaris nodes or 512 GPUs, by reducing frequency of collective gradient accumulations. Without GAS, scaling beyond 16 nodes (64 GPUs) incurred a significant performance penalty. Our production run used 40 GPUs due to the inefficiency of large batch optimization.

Meaningful Encoding of Chemical Space

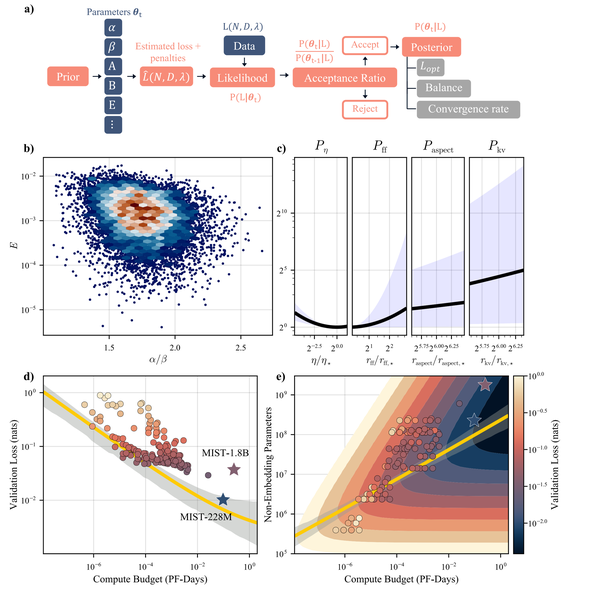
SciFMs, like MIST, need interpretable and steerable features that support hypothesis generation and failure diagnosis—capabilities well studied in language and vision models but still underexplored in scientific contexts.



(a) Interpretable chemical features were extracted from every stage of MIST models — token embeddings to downstream predictions. (b) A t-SNE projection of token embeddings shows dataset-specific structure. All datasets cluster the halogens and digits. The transition metal Quantum Mechanics dataset forms clusters for the left and right transition metal block. (c) & (d) UMAP projections reveal that the pretrained embeddings encode ring aromaticity and Polycyclic Aromatic Hydrocarbons classes, despite the model never training on labelled examples of these classes, suggesting the existence of chemically coherent structures in MIST's embedding space.

Compute-optimal Training

We developed hyperparameter-aware Bayesian scaling laws that eliminate per-scale hyperparameter sweeps. Fitting them required 138 MIST models spanning 393K–603M parameters and 4,760 GPU-hours; we estimate the standard exhaustive grid-search approach would require at least 10× more compute.

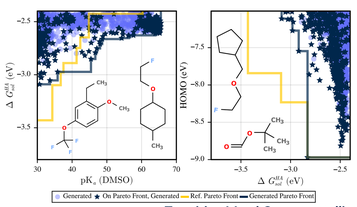


(a) The scaling laws were parameterized using MCMC, with penalty terms accounting for off-optimal hyperparameter selection. (b) Covariance plot of posterior samples for E and α/β after fitting penalized scaling laws shows low variance in α/β but higher variance in E. (c) Fitted penalty terms (with 95% credible interval) for learning rate (P_η) and encoder shape (P_{ff}, P_{nspect} & P_{kv}), showing estimated impact of off-optimal hyperparameters on the pretraining loss (L). (d) 95% credible interval for the compute-optimal frontier. (e) Posterior mean estimate of the scaling law and 95% credible interval for the optimal frotier minimum loss (L_{opt}). Overlaid scatter plots in d & e show the minimum validation loss of the model used to fit the scaling law.

3 MIST Rapidly Explores Chemical Space

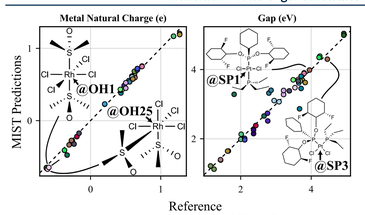
MIST models have been fine-tuned to predict more than 400 structure–property relationships and have been shown to match or exceed state-of-the-art performance across diverse benchmarks, from physiology to electrochemistry and quantum chemistry. We demonstrate the ability of these models to solve real-world problems across chemical space — multi-objective electrolyte solvent screening, isotope half-life prediction, stereochemical reasoning for chiral organometallic compounds and mixture property prediction.

Metal Air Batteries



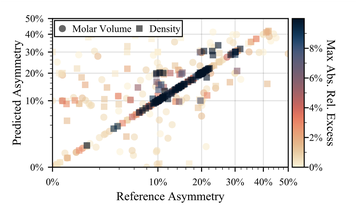
Using fine-tuned MIST-28M models, we identified 390 solvent molecules for lithium-air batteries on the Pareto front, improving stability against hydrogen abstraction and encouraging solution-mediated peroxide growth relative to our reference set.

Transition Metal Organometallics



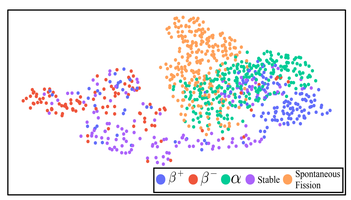
Parity plots for MIST-28M models, fine-tuned on transition metal organometallic complexes, with annotated examples of stereoisomer pairs. Stereochemical nuances are accurately captured by the Smirk tokenizer, enabling MIST to differentiate between isomers.

Additive Design



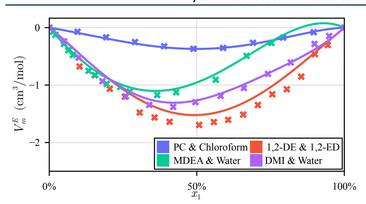
Identifying additives with high relative excess at small concentrations is critical during electrolyte formulation. MIST models accurately estimate the composition at the excess extrema.

Radioisotope Decay



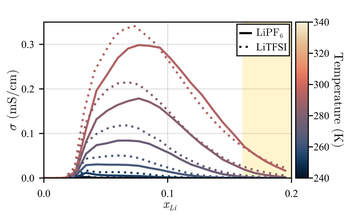
A Uniform Manifold Approximation and Projection (UMAP) projection of embedding vectors from a MIST-28M variant fine-tuned on the half-lives of radioisotopes reveals nuclide clusters organized by dominant decay channels.

Beyond Molecules — Mixtures



MIST mixture property models accurately predict thermodynamic excess properties of mixtures, including for challenging mixtures not captured by similarity-based descriptors.

Physically Valid Extrapolation

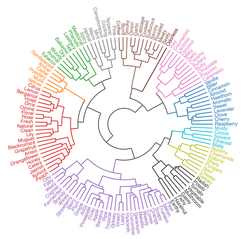


MIST models learn physically consistent trends in electrolyte ionic conductivity with salt mole fraction, correctly capturing the expected decay outside the training data range.

4 Olfaction: A Challenging Problem

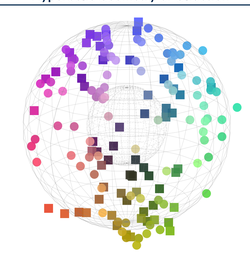
- **Difficult out-of-domain test for MIST:** the structure-to-odour map is nonlinear, discontinuous, and person-dependent.
- **Discontinuous mapping:** small molecular edits can cause large perceptual shifts.

Human Interpretable Scent Clusters



We observe that the model learns a hierarchical clustering of scents consistent with human perceptual organization; for example, "beefy", "meaty", "roasted", "savory" all form a cluster, fruity scents (such as "strawberry", "berry", "plum" and "grape") form another and notably "odourless" molecules are clustered alone.

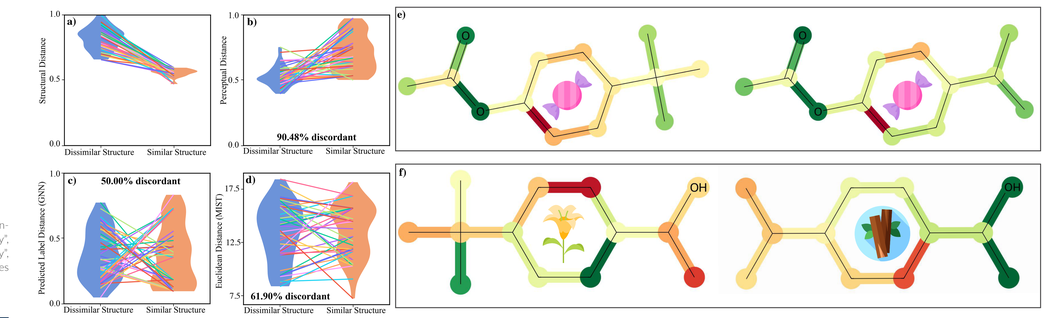
Hyperbolic Geometry of Olfaction



The existence of this hierarchical structure is consistent with the model having learned the hyperbolic geometry of the olfactory space. We modeled the olfactory profiles in hyperbolic and Euclidean space and compared the empirical Betti curves, showing the learned representation is statistically consistent with a low-dimensional hyperbolic geometry — notably, studies in neuroscience suggest the human brain also represents scent in a hyperbolic space.

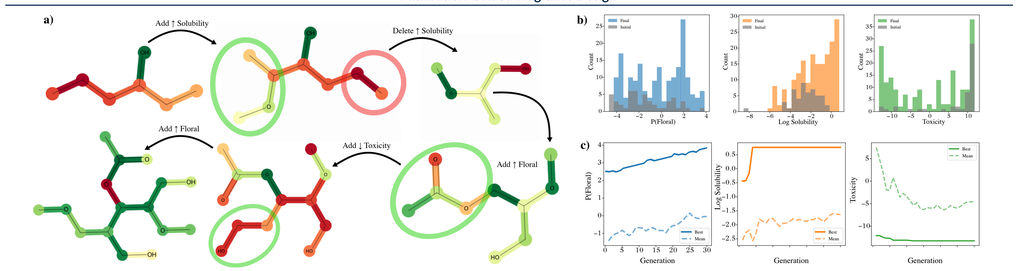
Explaining Discordance in Structure–Property Mappings

Discordant triplets are sets of three odourants where the structurally similar pair is not the pair which smells similar. This discrepancy between structural and perceptual similarities makes identifying discordant triplets a difficult prediction task. MIST prospectively resolved experimentally validated structure–odour discordance.



(a–b) Structure alone often fails to explain odor similarity: across 41 odor-discordant triplets, the experimentally closer-smelling molecule is frequently not the structurally closest one. (c) The prior state of the art resolves 50.0% of these discordances (d) while MIST resolves 61.9%. (e) Token-level attributions reveal why the same local structural change can alter olfactory profile in one molecular context (f) but not another.

Attribution-Guided Fragrance Design



Fragrance design requires jointly optimizing perceptual quality, physicochemical properties, and safety, posing a challenging multi-objective molecular design problem. Our attribution-guided NSGA-II framework uses atom-level importance scores to direct fragment additions and deletions toward Pareto-improving candidates, producing molecules with stronger floral character, improved solubility, and reduced predicted toxicity.

Use MIST /BattModels/mist-demo

MIST Models /mist-models