Scaling Foundation Models for Molecular Chemistry

Alexius Wadell[*][a], **Anoushka Bhutani**[*][a], **Venkat Viswanathan**[b]

[a] *Department of Mechanical Engineering, University of Michigan, Ann Arbor, Michigan*

[b] *Department of Aerospace Engineering, University of Michigan, Ann Arbor, Michigan* venkvis@umich.edu

1. Introduction

Current state-of-the-art molecular property prediction models require labeled training data generated using expensive wet-lab experiments or *ab initio* calculations [1, 2, 3, 4]. Their utility is limited by the scarcity and heterogeneity of labeled materials datasets. Foundation models (FMs) offer a solution to this: they use self-supervised pre-training strategies to leverage unlabeled datasets and learn representations of data that can be applied to downstream tasks. Prior attempts to train FMs for molecular property prediction demonstrate promise; however, supervised equivariant geometric models are still more accurate [5, 6, 7]. This can be attributed to the fact that FM are computationally expensive to train and difficult to interpret [8]. Our work addresses these challenges on three fronts: (1) we train a 228M parameter molecular FM on 6B molecules achieving state-of-the-art performance across various tasks (2) we develop Bayesian neural scaling laws enabling compute-optimal molecular FMs and (3) we probe the model to uncover chemical concepts learnt by it.

2. Model and Performance

		MIST-228M	MoLFormer [9]	SELFormer [10]
QM8	↓ Avg. MAE	0.0104 ± 0.0002	$\mathbf{0.0102}^{*}$	–
QM9	↓ Avg. MAE	2.4380 ± 0.0260	$\mathbf{1.5894}^{*}$	–
Tox21	↑ AUROC	$\mathbf{83.90 \pm 1.25}$	84.7	–
ToxCast	↑ AUROC	$\mathbf{84.35 \pm 0.84}$	–	–
ClinTox	↑ AUROC	$\mathbf{98.77 \pm 0.99}$	94.8	–
BACE	↑ AUROC	87.06 ± 2.82	**88.2**	83.2
BBBP	↑ AUROC	92.60 ± 3.10	**93.7**	86.3

Table 1: MIST-228M performance on MoleculeNet [11] compared to SOTA. $*$ indicates multiple models were used for a multitask benchmark.

MIST (Molecular Insight SMILES Transformer) is an encoder only pre-layer norm [12] transformer model based on the BERT architecture [13] pre-trained using Masked Language Modeling (MLM). The current variant model (MIST-228M) has 228M parameters and was trained on 6.14B molecules from Enamine's REAL Space [14]. It was trained using data distributed parallelism on 32 NVIDIA A100 GPUs with a throughput of $\approx 8k$ molecules per second. We used gradient accumulation to reduce communication latency resulting in large effective batch sizes. To enable data efficient training at large batch sizes we use the LAMB (Layerwise Adaptive Large Batch) algorithm [15]. For tokenization we used Smirk, a novel chemically meaningful tokenization strategy developed to train MIST; Smirk improves on the widely used "Atom-wise" tokenizers which often mask out critical chemical information [16].

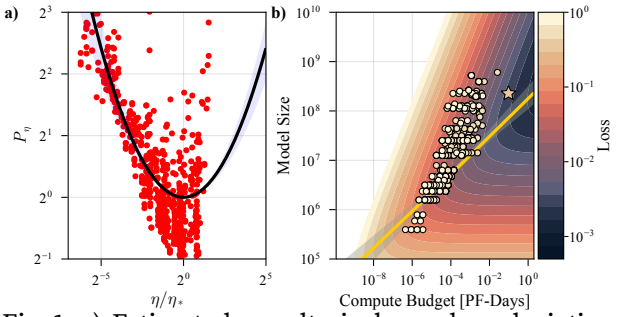

Fig. 1: a) Estimated penalty in loss when deviating from the ideal η_* learning rate. b) Bayesian neural scaling laws used to select MIST-228M hyperparameters ($\star$).

3. Bayesian Scaling Laws

Training FMs is an extremely computationally expensive endeavor. This is exacerbated when the dataset and model size needed to achieve the desired accuracy are not known. We developed neural scaling laws to guide the training of MIST and ensure a compute-optimal trade-off between model and dataset size [17]. Neural scaling laws for NLP have been widely studied, however, their application to scientific FMs is limited [18, 19]. Neural scaling laws estimate the eventual loss of the model L from the number of non-embedding parameters N, the dataset size D: $L(N, D) = \left(\frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + E\right) \prod P$. Where A, α, B, β and E are fitted coefficients typically found via non-linear optimization[20, 17]. We added penalty terms $\prod P$ to model non-optimal hyperparameter selection, extending the insight of neural scaling laws beyond the data/model size trade-off. By introducing penalty terms $P(x)$ with a single global minimum x^* such that $P(x^*) = 0$, we can directly model the impact of sub-optimal hyperparameter selection. The fitted scaling laws informed the size, shape and optimizer settings for MIST-228M (fig. 1).

4. Downstream Applications

MIST has been applied to various electrochemical materials design problems. Mixture property prediction is discussed as an example application here. Accurate mixture property prediction is difficult due to the complexity of mixture chemistry and data is limited [21]. We have sucessfully extended MIST's capabilities from single molecule property prediction

[*]These authors contributed equally to this work

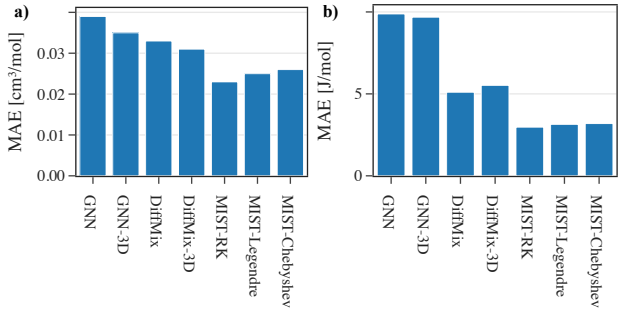

Fig. 2: MIST outperforms models designed specifically for mixture property prediction[22], some of which use 3D molecular geometry, on both mixture molar volume (a) and enthalpy (b) prediction.

to mixtures property prediction using a downstream task network informed by chemical thermodynamics. A mixture property can be decomposed into a linear mixing and an excess term, which quantifies the non-ideal behavior of real mixtures due to component interactions. The excess term depends on the mixture composition and is commonly modeled using a Redlich-Kister (R-K) polynomial basis. In MIST, this is modeled by passing the hidden states to a single component property and polynomial coefficient prediction network. The mixture property is then calculated using the method show in section E.

5. Interpretability

Several molecular FMs demonstrate near state-of-the-art performance on property prediction [23, 10, 9]; however, the vast number of non-physical parameters make interpretation difficult. Recent works have explored using attention maps to visualize how the model evaluates a molecule[24, 9, 25]. We apply similar methods to deduce insights from MIST's attention matrices (fig. A1) and embedding vectors (fig. 4). Additionally, we propose a novel quantitative analysis of model interpretability based on synthetic accessibility and molecular assembly index.

Practical synthesis of novel compounds is a pertinent problem and a barrier to realizing inverse molecular design [26]; Synthetic Accessibility (SA) metrics estimate the difficulty of synthesis using hand-crafted features[27, 28], retrosynthetic planning[29] or data-driven methods[30, 31]; Unfortunately, the "ease of synthesis" remains highly subjective with limited agreement between chemists [32, 27, 33]. Parallelly, the Molecular Assembly Index (MA), evaluates molecular complexity from an Assembly Theory perspective [34]. Assembly theory proposes that molecules with a higher MA are more difficult to compose and hence their occurrence is less likely. We propose framing SA and MA as a measure of a chemist's surprise for a molecule's existence. This framing suggests that molecular FMs learn synthetic accessibility and the assembly index implicitly during the pre-training process.

To validate this, we compared the log-probabilities predicted by MIST and MoL-

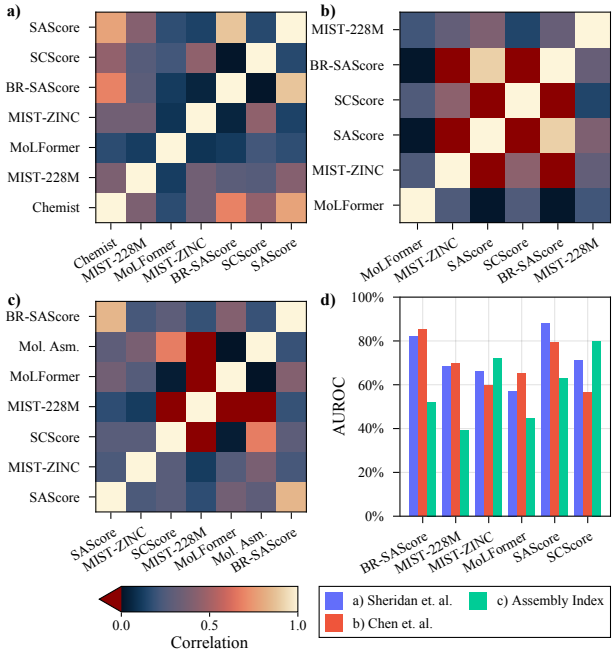

Fig. 3: R^2 correlation between different metrics on datasets for Synthetic Accessibility (a&b) and Molecular Assembly (c). AUROC scores for each metric evaluated on the datasets are shown in (d).

Former to three benchmark datasets for synthetic accessibly[9, 32, 28]. We found the log-probabilities predicted by both models are strongly correlated with existing metrics for measuring synthetic accessibility (fig. 3).

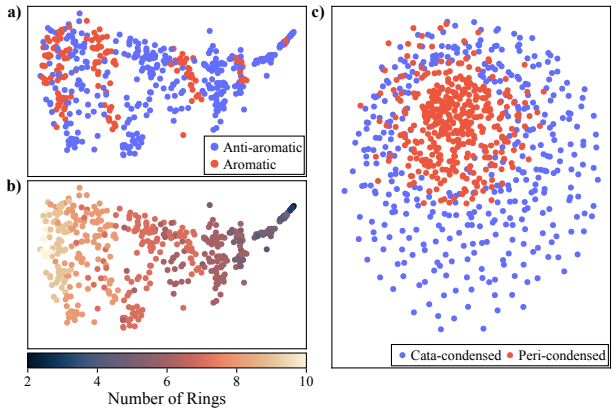

Fig. 4: a) t-SNE of MIST's pre-trained embeddings shows 'bands' of aromatic and anti-aromatic molecules. b) One t-SNE component aligns with the number of benzene rings in the molecule. c) UMAP of MIST's embeddings shows concentric rings differentiating between cata and peri-condensed polybenzenoid hydrocarbons.

Additionally, we used MIST's pre-trained embeddings to identify chemical rules encoded by the model. We probed if the model can differentiate aromatic from anti-aromatic molecules (figs. 4a and 4b) and cata from peri-condensed polybenzenoid hydrocarbons (fig. 4c) in the COMPAS datasets [35, 36, 37].

Acknowledgments

Computational resources for this work were provided by an award for computer time was provided by the U.S. Department of Energy's (DOE) Innovative and Novel Computational Impact on Theory and Experiment (INCITE) Program [38]. This research used resources from the Argonne Leadership Computing Facility, a U.S. DOE Office of Science user facility at Argonne National Laboratory, which is supported by the Office of Science of the U.S. DOE under Contract No. DE-AC02-06CH11357. The authors acknowledge the National Artificial Intelligence Research Resource (NAIRR) Pilot and NVIDIA DGX Cloud for contributing to this research result. This work was supported by Los Alamos National Laboratory under the project "Algorithm/Software/Hardware Co-design for High Energy Density applications" at the University of Michigan. This research was supported in part through computational resources and services provided by Advanced Research Computing at the University of Michigan, Ann Arbor. This work used the Delta system at National Center for Supercomputing Applications through allocation CTS180061 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Additionally, AW is grateful for the support of Meta's AR/AV Battery Research Fellowship. AB is supported by a Catalyst grant from the Michigan Institute for Computational Discovery and Engineering at the University of Michigan.

References

[1] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1):2453, May 2022.

[2] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs.

[3] Guillem Simeon and prefix=de useprefix=false family=Fabritiis, given=Gianni. Tensornet : Cartesian tensor representations for efficient learning of molecular potentials.

[4] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2 : Improved equivariant transformer for scaling to higher-degree representations.

[5] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *CoRR*, abs/2010.09885, 2020.

[6] Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. 2022.

[7] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. 2022.

[8] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ B. Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri S. Chatterji, Annie S. Chen, Kathleen Creel, Jared Quincy Davis, Dorottya Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah D. Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark S. Krass, Ranjay Krishna, Rohith Kuditipudi, and et al. On the opportunities and risks of foundation models. *CoRR*, abs/2108.07258, 2021.

[9] Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. 4(12):1256–1264.

[10] Atakan Yüksel, Erva Ulusoy, Atabey Ünlü, and Tunca Doğan. SELFormer: Molecular representation learning via SELFIES language models. *Machine Learning: Science and Technology*, 4(2):025035, June 2023.

[11] Zhenqin Wu, Bharath Ramsundar, Evan N. Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S. Pappu, Karl Leswing, and Vijay Pande. MoleculeNet: A Benchmark for Molecular Machine Learning, October 2018.

[12] Sho Takase, Shun Kiyono, Sosuke Kobayashi, and Jun Suzuki. Spike no more : Stabilizing the pre-training of large language models.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018.

[14] Enamine Ltd. REAL Space, 2024.

[15] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning : Training bert in 76 minutes.

[16] Alexius Wadell, Anoushka Bhutani, and Venkatasubramanian Viswanathan. Smirk: An Atomically Complete Tokenizer for Molecular Foundation Models, September 2024.

[17] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022.

[18] Shashank Subramanian, Peter Harrington, Kurt Keutzer, Wahid Bhimji, Dmitriy Morozov, Michael Mahoney, and Amir Gholami. Towards foundation models for scientific machine learning : Characterizing scaling and transfer behavior.

[19] Nathan C. Frey, Ryan Soklaski, Simon Axelrod, Siddharth Samsi, Rafael Gómez-Bombarelli, Connor W. Coley, and Vijay Gadepally. Neural scaling of deep chemical models. pages 1–9.

[20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020.

[21] Adarsh Dave, Jared Mitchell, Kirthevasan Kandasamy, Han Wang, Sven Burke, Biswajit Paria, Barnabás Póczos, Jay Whitacre, and Venkatasubramanian Viswanathan. Autonomous discovery of battery electrolytes with robotic experimentation and machine learning. 1(12):100264.

[22] 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. 15(1):8649.

[23] Eduardo Soares, Victor Shirasuna, Emilio Vital Brazil, Renato Cerqueira, Dmitry Zubarev, and Kristin Schmidt. A Large Encoder-Decoder Family of Foundation Models For Chemical Language, July 2024.

[24] Philippe Schwaller, Daniel Probst, Alain C. Vaucher, Vishnu H. Nair, David Kreutter, Teodoro Laino, and Jean-Louis Reymond. Mapping the space of chemical reactions using attention-based neural networks. *Nature Machine Intelligence*, 3(2):144–152, January 2021.

[25] Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction, October 2020.

[26] Filip T. Szczypiński, Steven Bennett, and Kim E. Jelfs. Can we predict materials that can be synthesised? *Chemical Science*, 12(3):830–840, 2021.

[27] Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of Cheminformatics*, 1(1):8, June 2009.

[28] Shuan Chen and Yousung Jung. Estimating the synthetic accessibility of molecules with building block and reaction-aware SAScore. *Journal of Cheminformatics*, 16(1):83, July 2024.

[29] Qi Huang, Lin-Li Li, and Sheng-Yong Yang. RASA: A Rapid Retrosynthesis-Based Scoring Method for the Assessment of Synthetic Accessibility of Drug-like Molecules. *Journal of Chemical Information and Modeling*, 51(10):2768–2777, October 2011.

[30] Connor W. Coley, Luke Rogers, William H. Green, and Klavs F. Jensen. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *Journal of Chemical Information and Modeling*, 58(2):252–261, February 2018.

[31] Milan Voršilák, Michal Kolář, Ivan Čmelo, and Daniel Svozil. SYBA: Bayesian estimation of synthetic accessibility of organic compounds. *Journal of Cheminformatics*, 12(1):35, May 2020.

[32] Robert P. Sheridan, Nicolas Zorn, Edward C. Sherer, Louis-Charles Campeau, Charlie (Zhenyu) Chang, Jared Cumming, Matthew L. Maddess, Philippe G. Nantermet, Christopher J. Sinz, and Paul D. O'Shea. Modeling a Crowdsourced Definition of Molecular Complexity. *Journal of Chemical Information and Modeling*, 54(6):1604–1616, June 2014.

[33] Yuji Takaoka, Yutaka Endo, Susumu Yamanobe, Hiroyuki Kakinuma, Taketoshi Okubo, Youichi Shimazaki, Tomomi Ota, Shigeyuki Sumiya, and Kensei Yoshikawa. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition. *Journal of Chemical Information and Computer Sciences*, 43(4):1269–1275, July 2003.

[34] Abhishek Sharma, Dániel Czégel, Michael Lachmann, Christopher P. Kempes, Sara I. Walker, and Leroy Cronin. Assembly theory explains and quantifies selection and evolution. 622(7982):321–328.

[35] Alexandra Wahab, Lara Pfuderer, Eno Paenurk, and Renana Gershoni-Poranne. The compas project : A computational database of polycyclic aromatic systems . phase 1: Cata - condensed polybenzenoid hydrocarbons. 62(16):3704–3713.

[36] Eduardo Mayo Yanes, Sabyasachi Chakraborty, and Renana Gershoni-Poranne. Compas-2 : A dataset of cata-condensed hetero-polycyclic aromatic systems. 11(1):97.

[37] Alexandra Wahab and Renana Gershoni-Poranne. Compas-3 : A data set of peri-condensed polybenzenoid hydrocarbons.

[38] Office of Science U.S. Department of Energy. 2024 incite fact sheet .pdf.

[39] Huggingface/tokenizers: Fast State-of-the-Art Tokenizers optimized for Research and Production, February 2024.

[40] Microsoft/DeepSpeed. Microsoft, May 2024.

[41] Jerry Ma and Denis Yarats. On the adequacy of untuned warmup for adaptive optimization.

[42] Elad Hoffer, Itay Hubara, and Daniel Soudry. Train longer, generalize better: Closing the generalization gap in large batch training of neural networks.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. *arXiv:1706.03762 [cs]*, December 2017.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need.

[45] Michael Jirasek, Abhishek Sharma, Jessica R. Bame, S. Hessam M. Mehr, Nicola Bell, Stuart M. Marshall, Cole Mathis, Alasdair MacLeod, Geoffrey J. T. Cooper, Marcel Swart, Rosa Mollfulleda, and Leroy Cronin. Investigating and quantifying molecular complexity using assembly theory and spectroscopy. 10(5):1054–1064.

[46] Celia Kelly, Emil Annevelink, Adarsh Dave, and Venkatasubramanian Viswanathan. Excess density as a descriptor for electrolyte solvent design.

[47] Arthur D. Pelton and Christopher W. Bale. Legendre polynomial expansions of thermodynamic properties of binary solutions. 17(6):1057–1063.

Appendix A. Model Architecture and Training

The MIST-228M model is based on the `RoBERTa-PreLayerNorm` implementation in HuggingFace [39] with absolute position embeddings, 16 attention heads, 18 hidden layers, a hidden size of 1,024, intermediate size of 4,096 and a maximum sequence length of 512. The DeepSpeed implementation of the FusedLAMB optimizer [40] was used to train the model. A linear warm-up [41] and cosine decay learning rate schedule [17] was used with a maximum learning rate of $2.5e-4$. The learning rate was selected based on the fitted optimal learning rate as determined our fitted Bayesian neural-scaling laws fig. 1. During finetuning, a task network consisting of two feedforward layers with GeLU activations was attached to the network. All model weights were updated during finetuning.

Appendix B. Bayesian Neural Scaling Laws

Neural scaling laws are typically fit via non-linear optimization to $L(N, D)$ as a function of A, α, B, β and E [20, 17]. Once fit they can be used identify the "compute-optimal" model size, or the model size N_{opt} that minimizes $L(N, D)$ for a given compute budget $C \approx 6ND$ [17]. A "compute-optimal" D_{opt} dataset size can similarly be defined.

$$N_{opt} = G\left(\frac{C}{6}\right)^{a} D_{opt} = G^{-1}\left(\frac{C}{6}\right)^{b} \tag{A1}$$

$$G = \left(\frac{\alpha A}{\beta B}\right)^{\frac{1}{\alpha+\beta}} a = \frac{\beta}{\alpha+\beta} \qquad b = \frac{\alpha}{\alpha+\beta}$$

2.1 Optimal Learning Rate Scaling

We model the optimal learning rate $\eta_*(N, B) = \eta_0 N^{\gamma} B^{\delta}$; where η_0, γ and δ are fitted coefficients and B is the effective batch size (Global batch size times gradient accumulation steps). By using a Bayesian framework, we can incorporate our prior-belief that $\eta_* \approx 1.4 \times 10^{-4}$ for a batch size of ≈ 1024 [9]. As well as our expectation that $\delta \approx 0.5$, corresponding square-root scaling with batch size [42]. Prior NLP models have recommended decreasing η with model size[20, 43]; however similar recommendations are absent from the molecular foundation model literature. We elected to use an uninformative normal prior with mean of 0, to reflect our uncertainty here.

Appendix C. Attention Map Interpretability

A transformer's attention matrix provides a distribution over attended-to input units, and can be interpreted as communicating the relative importance of inputs [44]. MIST is able to capture spatial relations between atomic tokens that are not necessarily neighbors in the SMILES sequence in it's attention matrices. The attention map exhibits awareness of bond connectivity and interatomic distances in 3D space. This suggests the model is able to learn 3D geometric information about molecular structure not explicitly present in SMILES strings.

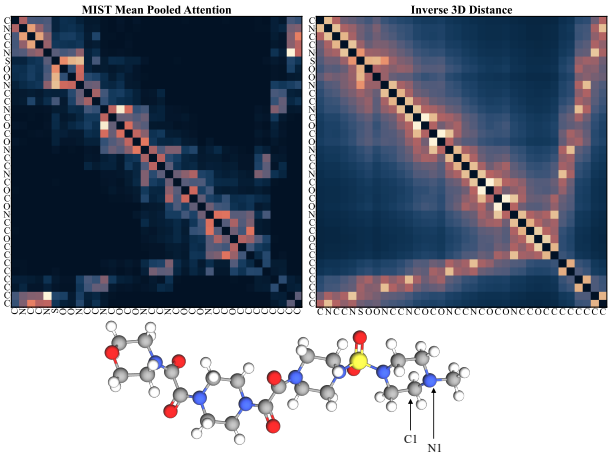

Fig. A1: Visualization of the learned attention map and corresponding molecular structure (inverse 3D distance) for a molecule in the pre-training validation set. The mean pooled attention (across 16 attention heads) in layer 9 of the encoder is visualized. The molecule visualized is represented by the SMILES string `C*N1*CCN(S(=O)(=O)N2CCN(C(=O)C(=O)N3CCN(C(=O)C(=O)N4CCOCC4)CC3)CC2)C*C1*`. From the plot we can see that the model assigns high attention scores to pairs of Carbon and Nitrogen atoms which appear next to each other in rings (for e.g `N1` and `C1`) even though they are far apart in the SMILES sequence.

Appendix D. Synthetic Accessibility and Molecular Assembly

4.1 Synthetic Accessibility

We used two datasets to assess the ability of molecular foundation models to predict a molecule's "ease of synthesis." The first dataset (fig. 3a) consists of 1,775 molecules scored on a scale of 1 - 5 as a part of a crowdsourced evaluation of the "complexity" of a molecule. Participants were asked to score a random selection of 5 molecules on a scale of 1 to 5; participants were encouraged to evaluate multiple sets. Each molecule was evaluated by 41 ± 16 chemists from Merck with an overall average score of 2.85 ± 0.78. The authors noted that "complexity" was highly subjective; on average molecules, the standard deviation of a molecule's complexity was 0.77 ± 0.14. The "Chemist" shown in fig. 3a is the `meanComplexity` as reported by the dataset. We used 2.85, the average complexity score, as the threshold for evaluating the AUROC of other metrics (fig. 3d). That is fig. 3d, shows evaluates the ability of metrics to predict `meanComplexity` > 2.85.

The second dataset (fig. 3b) of 5000 molecules was curated by Chen et al. from multiple public datasets and labeled as either easy or hard to synthesize based on the source dataset [28]. We are using the test split, as released, as the other splits were not released. Notable, this is the dataset used by Chen et al. for the development of BR-SAscore, one of the metrics shown in fig. 3. In fig. 3d, we evaluated each metrics' ability to discriminate "hard" from "easy" to synthesize molecules.

4.2 Molecular Assembly

The Molecular Assembly Index (MA) is defined as the number of steps on a recursive minimal path to produce the molecular graph. Molecules with an assembly index of above 15 are considered biosignatures [34]. For our analysis (fig. 3c), we used a dataset on 450 molecules from Ref. [45]. The dataset contains MA values computed using an algorithmic search as well as values computed using experimentally collected descriptors such as tandem mass spectrometry and infrared spectroscopy. In fig. 3d, we evaluated each metrics' ability to correctly classify molecules as biosignatures (MA > 15). We compare the algorithmically computed the MA values calculate correlations with other metrics.

Appendix E. Mixture Property Prediction

MIST predicts properties of mixtures using a downstream task network informed by chemical thermodynamics. The architecture of this task network is informed by the knowledge that mixture properties can be decomposed into a linear mixing and an excess term. The excess term quantifies the non-ideal behavior of real mixtures due to component interactions. This term is commonly modeled as a function of mole fractions using a Redlich-Kister (R-K) polynomial basis [46]. The architecture used to predicted mixture properties using MIST is shown in fig. A2. The mixture embeddings are constructed using the hidden states extracted from the final hidden layer of the transformer encoder mean-pooled across the heads of the transformer. The hidden states are then passed to a single component property prediction network and polynomial coefficient prediction network. The mixture property, for a binary mixture, is then calculated using a polynomial:

$$
\begin{aligned}
x &= x_1 - x_2 \rightarrow 1 - 2x_2 \\
P_{mix} &= \underbrace{x_1 P_1 + x_2 P_2}_{\text{linear mixing}} \\
&\quad \underbrace{+ x(1-x) \sum_{i=0}^{n} \Omega_i \cdot f_i(1-2x)}_{\text{excess term}}
\end{aligned}
$$

where x_1 and x_2 are the mole fractions of the two components in the mixture, P_1 and P_2 are single component properties predicted by the network, f_i is the i^{th} R-K, Legendre or Chebyshev polynomial and Ω_i is the predicted coefficients. We use Legendre and Chebyshev polynomials in addition to the widely used R-K polynomials to test if the prediction accuracy improves when a orthonormal polynomial ba-

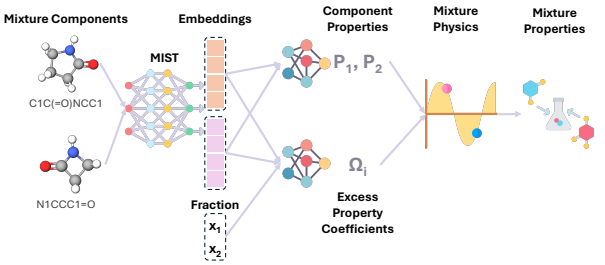

Fig. A2: Mixture property prediction workflow: SMILES representations of the two molecules in the mixture are individually passed to the pre-trained MIST encoder in order to obtain their embedding vectors. The individual embedding vectors are passed to a component property prediction network (a MLP - multi-layer perceptron - with two layers). This predicts individual component properties (P_1 and P_2). The two embeddings are concatenated together and passed to a excess property prediction network (also a three layer MLP) which predicts the coefficients of the polynomial (Ω_i). These values are then passed to a polynomial layer which computes the appropriate polynomial summation based on the order and basis type (Legendre, R-K or Chebyshev) specified.

sis is used in the output layer [47]. However, we see that the three polynomials perform similarly in this case.

The model was finetuned on two excess property datasets described in Ref [22]. The excess molar volume dataset consisted on 1069 binary mixture data points (28 unique mixtures composed of 25 organic chemicals with varying compositions). The excess molar enthalpy dataset consisted on just 631 data points (34 unique mixtures composed of 35 organic chemicals with varying compositions). All model weights were updated during finetuning, including the pre-trained encoder weights. The task network consisted of 3M learnable parameters. Finetuning was carried out using the AdamW optimizer. The results presented in the main paper are for a polynomial task heads with degree four polynomials.