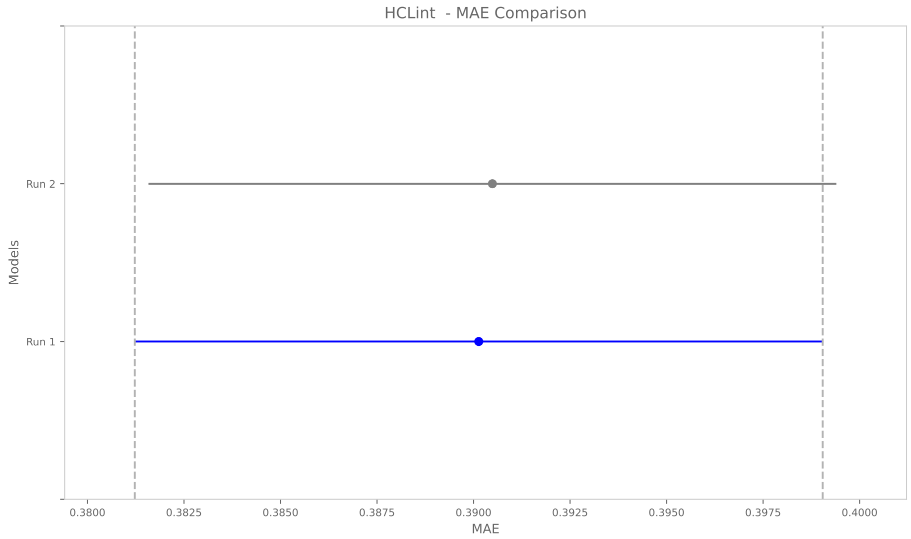
Supplementary Information for MolEncoder: Towards Optimal Masked Language Modeling for Molecules

Fabian P. Krüger[1,2,3]✉, Nicklas Österbacka[1], Mikhail Kabeshov[1], Ola Engkvist[1,4], Igor Tetko[3]

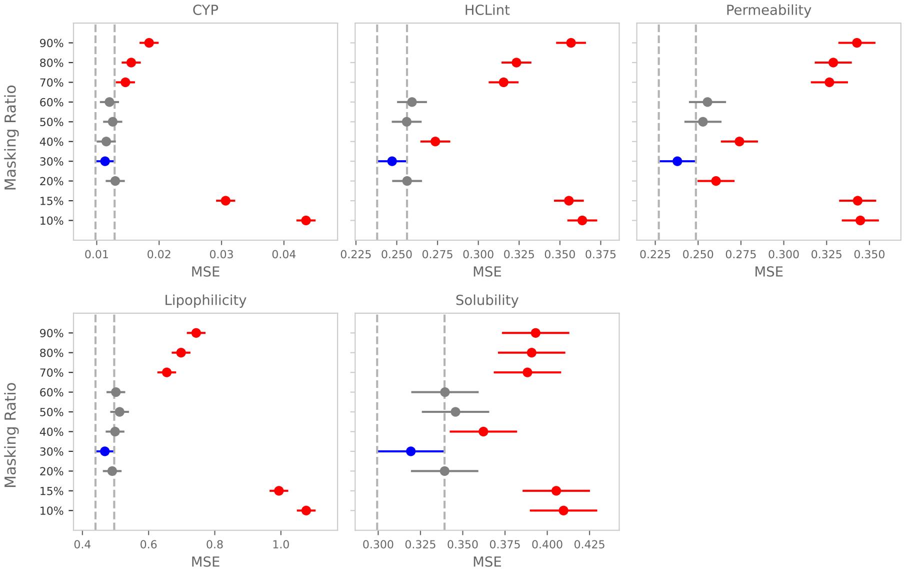
[1]AstraZeneca R&D, Discovery Sciences, Molecular AI, 431 83 Mölndal, Sweden
[2]TUM School of Computation, Information and Technology, Department of Mathematics, Technical University of Munich, 80333 Munich, Germany
[3]Helmholtz Munich – German Research Center for Environmental Health (GmbH), Institute of Structural Biology, Molecular Targets and Therapeutics Center, 85764 Neuherberg, Germany
[4]Department of Computer Science and Engineering, Chalmers University of Technology and University of Gothenburg, Sweden

✉ fabian.krueger@tum.de

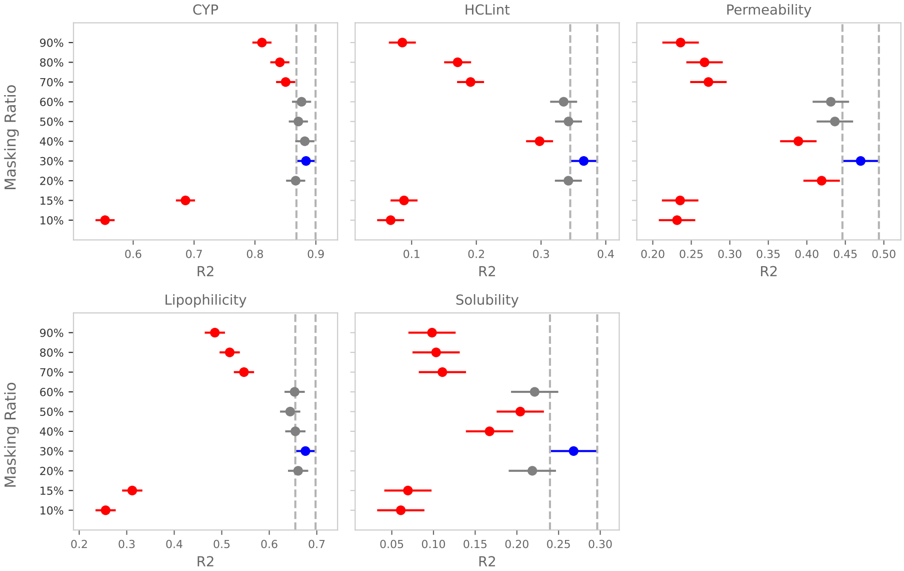
October 2025



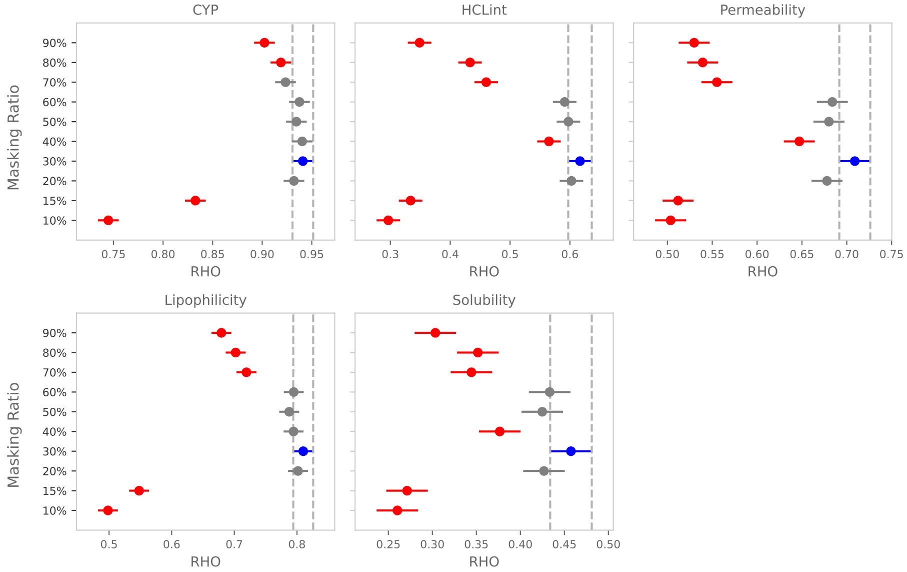
Supplementary Figure S1. Comparison of two runs of the same model in our evaluation pipeline using different random seeds. The null hypothesis that both models have the same performance is not rejected, demonstrating the robustness of the evaluation pipeline to random seed variation (overlapping 95% confidence interval shown in grey).



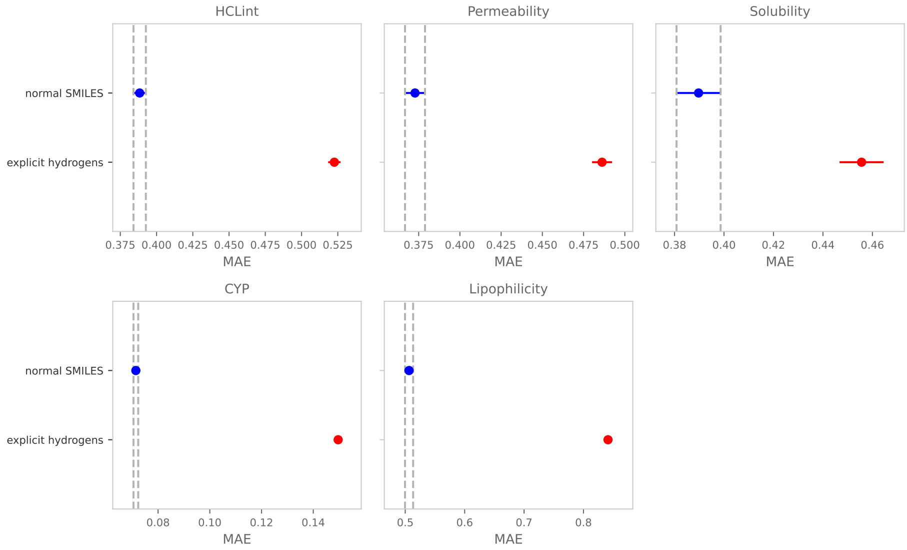
Supplementary Figure S2. Same experiment as Figure 1 in the main document, but with mean squared error as an evaluation metric.



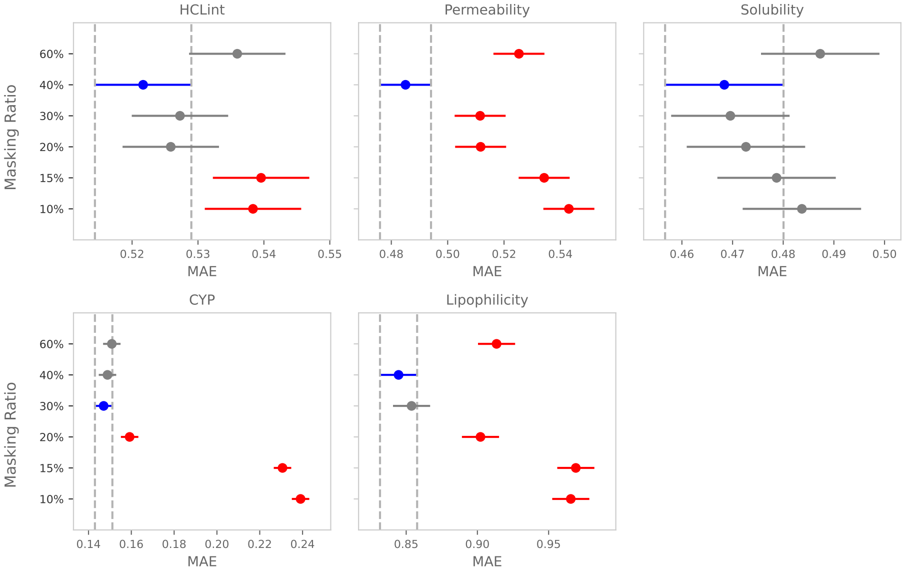
Supplementary Figure S3. Same experiment as Figure 1 in the main document, but with R^2 score as an evaluation metric.



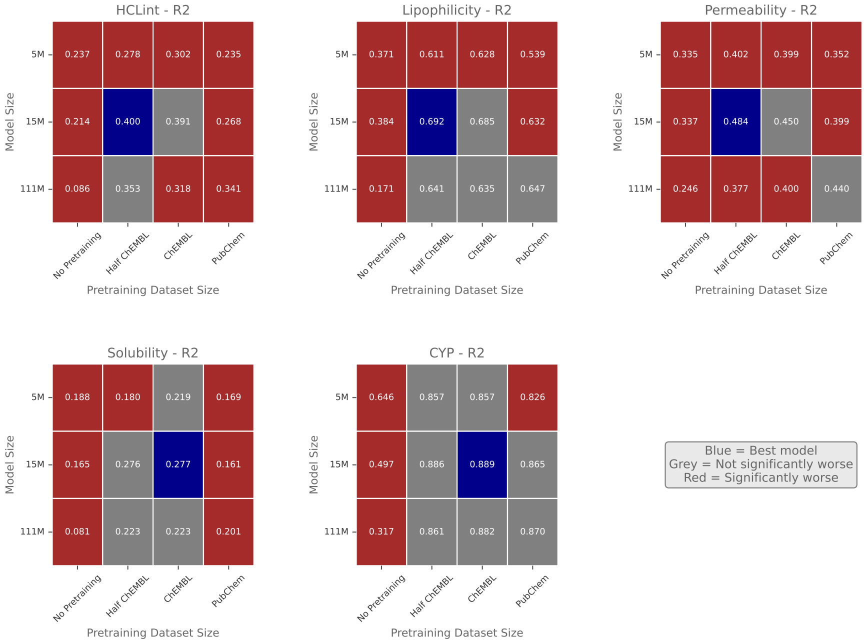
Supplementary Figure S4. Same experiment as Figure 1 in the main document, but with Spearman correlation as an evaluation metric.



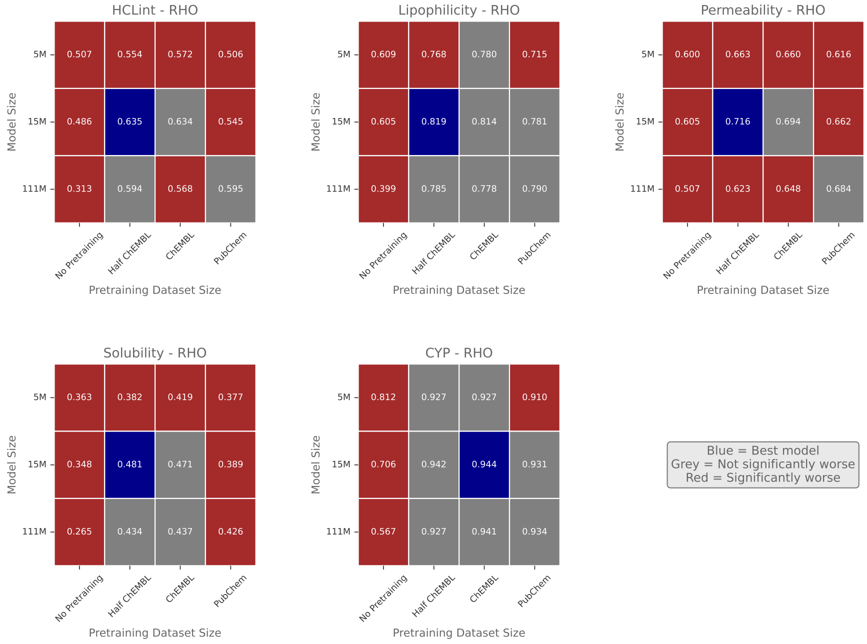
Supplementary Figure S5. Comparison of pretraining and evaluating on SMILES strings with explicit hydrogen atoms against SMILES strings with implicit hydrogen atoms (standard SMILES strings). The models that are compared have 15M parameters and are pretrained on the molecules in ChEMBL.



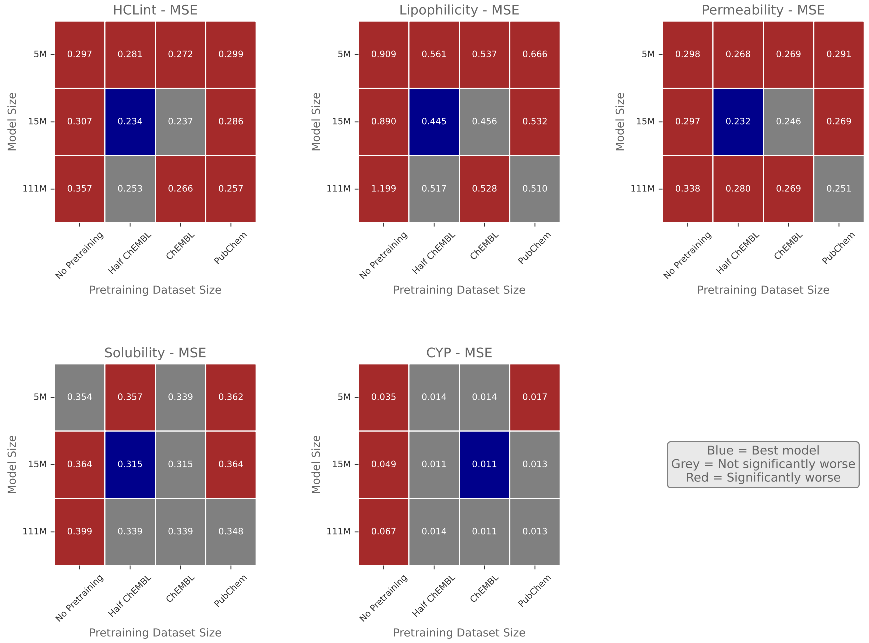
Supplementary Figure S6. Same experiment as Figure 1 in the main document, but with a model pretrained and evaluated on SMILES strings with explicit hydrogen atoms.



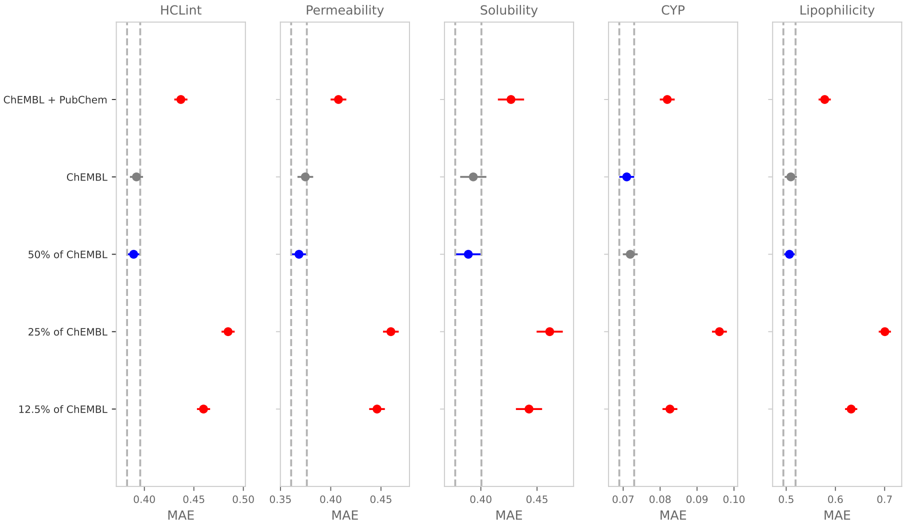
Supplementary Figure S7. Same experiment as Figure 2 in the main document, but with R^2 score as an evaluation metric.



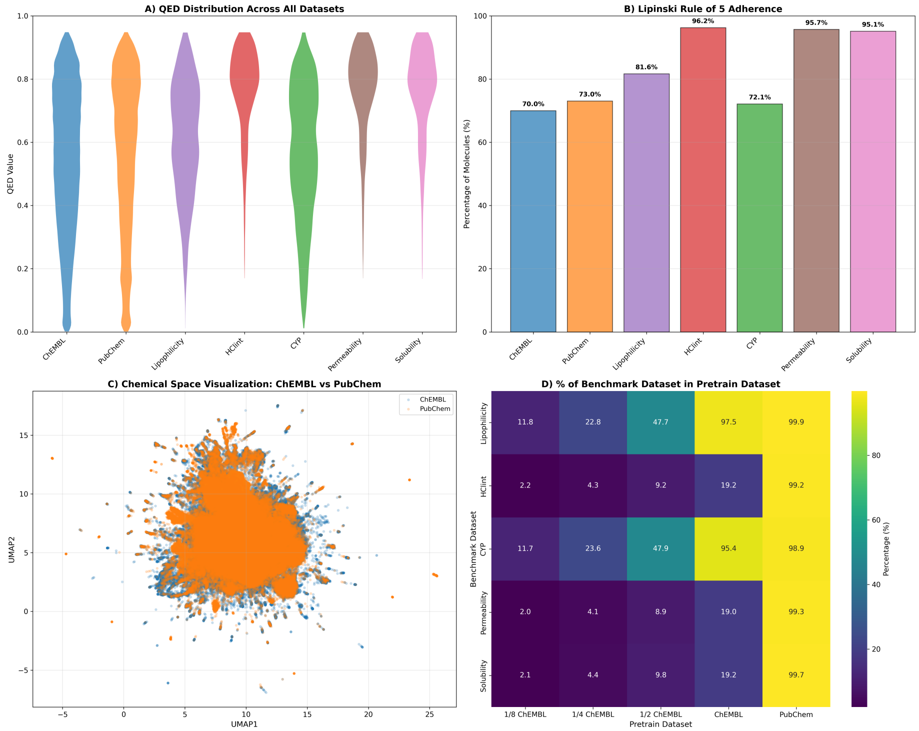
Supplementary Figure S8. Same experiment as Figure 2 in the main document, but with Spearman correlation as an evaluation metric.



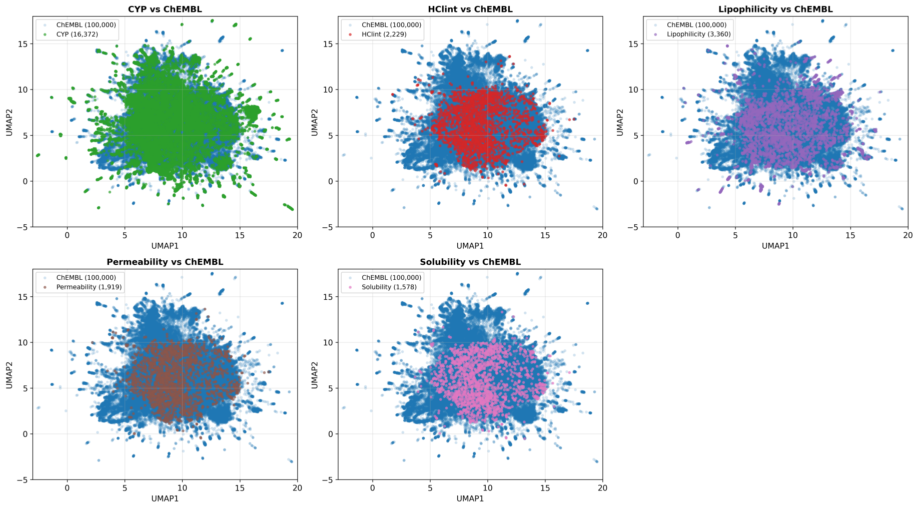
Supplementary Figure S9. Same experiment as Figure 2 in the main document, but with mean squared error as an evaluation metric.



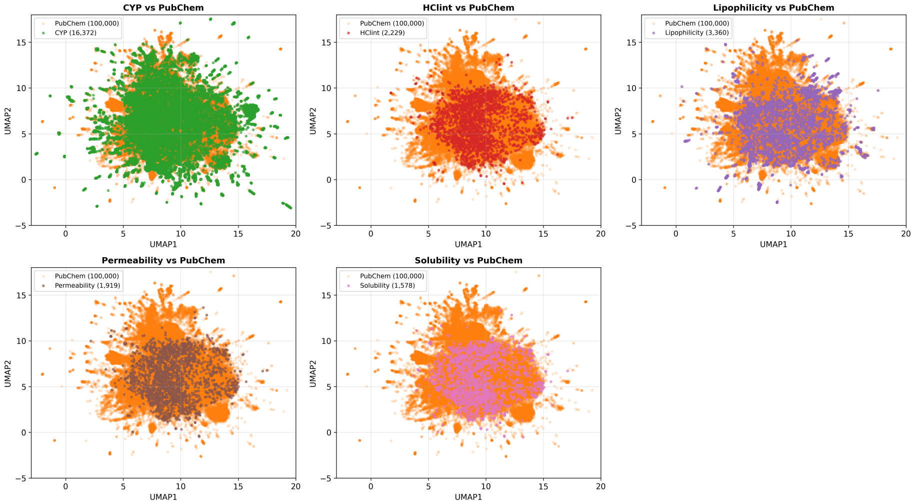
Supplementary Figure S10. Comparison of models pretrained on differently sized datasets using masked language modeling with a masking ratio of 30%. The models used for this comparison had 15M parameters.



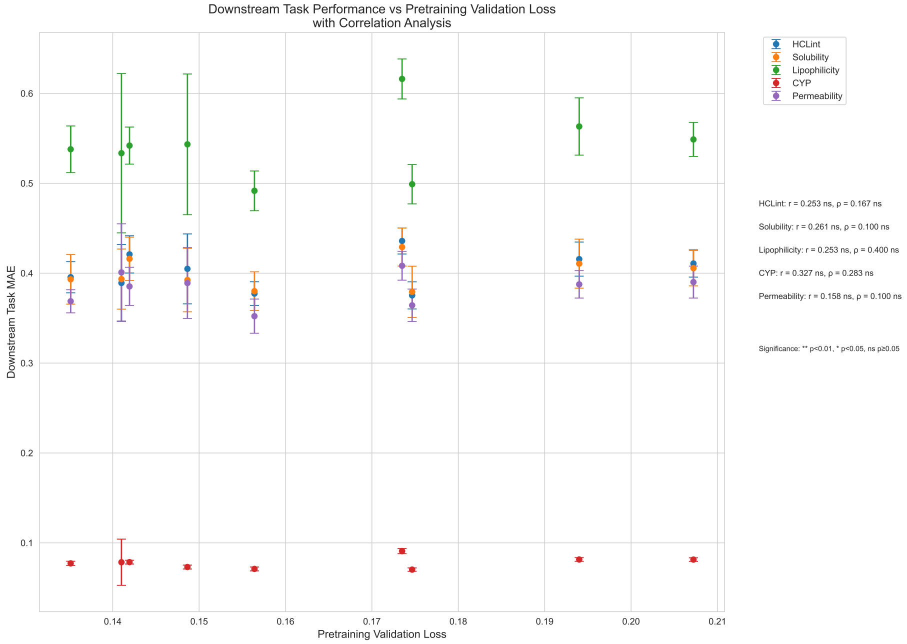
Supplementary Figure S11. Comparison of molecular property and chemical space distributions between ChEMBL and PubChem. (A) Quantitative estimate of drug-likeness (QED) distributions. (B) Fraction of molecules satisfying Lipinski's Rule of Five. (C) Two-dimensional UMAP projection of molecular fingerprints constructed from the first 200 principal components, explaining approximately 55% of the variance. (D) Fraction of SMILES strings from each pretraining dataset that are also present in the downstream evaluation datasets. Both datasets were uniformly subsampled to 100,000 molecules for computational efficiency for the subplots A-C. Subplot D contains the full datasets.



Supplementary Figure S12. Comparison of chemical space between ChEMBL and the downstream evaluation datasets. Each point represents a molecule embedded in a two-dimensional UMAP projection derived from the first 200 principal components of molecular fingerprints (explaining 55% of the variance). The same UMAP embedding used for the ChEMBL–PubChem comparison was applied here to ensure a consistent latent space.



Supplementary Figure S13. Comparison of chemical space between PubChem and the downstream evaluation datasets. Each point represents a molecule embedded in a two-dimensional UMAP projection derived from the first 200 principal components of molecular fingerprints (explaining 55% of the variance). The same UMAP embedding used for the ChEMBL–PubChem comparison was applied here to ensure a consistent latent space.



HCLint: r = 0.253 ns, ρ = 0.167 ns

Solubility: r = 0.261 ns, ρ = 0.100 ns

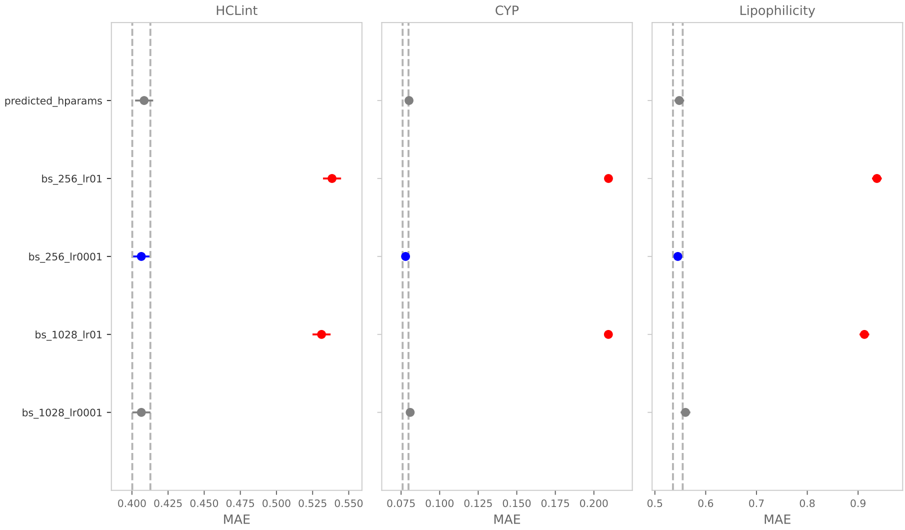
Lipophilicity: r = 0.253 ns, ρ = 0.400 ns

CYP: r = 0.327 ns, ρ = 0.283 ns

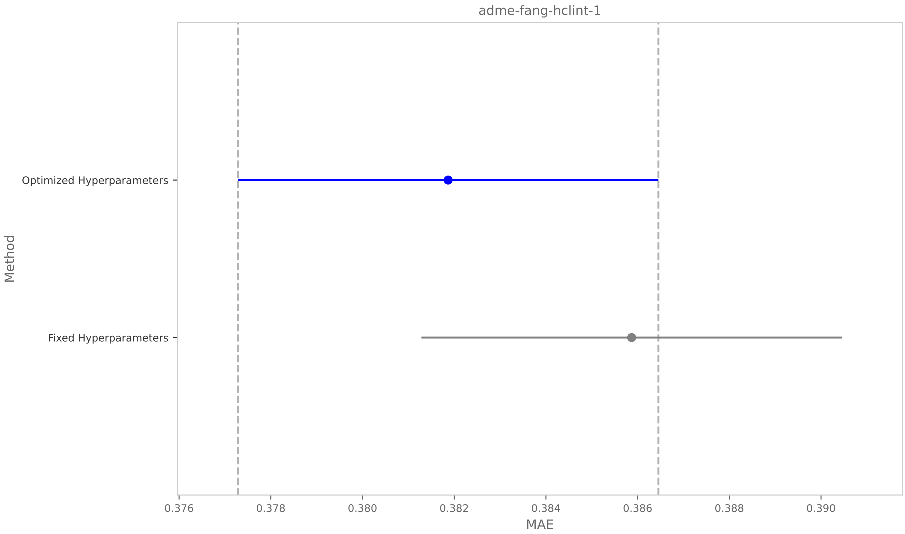
Permeability: r = 0.158 ns, ρ = 0.100 ns

Significance: ** $p<0.01$, * $p<0.05$, ns $p\geq0.05$

Supplementary Figure S14. Downstream performance (measured as MAE) plotted against pretraining performance (cross-entropy loss). Pearson and Spearman correlation coefficients for each evaluation dataset are shown next to the figure. No significant correlation was observed for any of the evaluation datasets.



Supplementary Figure S15. Comparison of models pretrained with different hyperparameters (batch size and learning rate). The predicted hyperparameters correspond to those used in the study and do not differ significantly from the best values identified through manual hyperparameter tuning.



Supplementary Figure S16. Comparison of fine-tuning the pretrained 15M-parameter model using our fixed fine-tuning hyperparameters versus Bayesian hyperparameter optimization for fine-tuning on the specific dataset. The resulting model performances do not differ significantly.

Supplementary Table S1. Mean Absolute Error (MAE) and benchmark-specific ranks for *MolEncoder*. The rank reflects the leaderboard at the 03.07.2025. Archived versions of the leaderboard are available at `https://web.archive.org/web/20250703093745/https://polarishub.io/benchmarks/polaris/adme-fang-perm-1`, `https://web.archive.org/web/20250703094412/https://polarishub.io/benchmarks/polaris/adme-fang-hclint-1`, `https://web.archive.org/web/20250703094653/https://polarishub.io/benchmarks/polaris/adme-fang-solu-1`, `https://web.archive.org/web/20250703094922/https://polarishub.io/benchmarks/novartis/adme-novartis-cyp3a4-reg`, `https://web.archive.org/web/20250703095138/https://polarishub.io/benchmarks/tdcommons/lipophilicity-astrazeneca`.

Benchmark	**MAE**	**Rank (out of N)**
Permeability	0.305	3 / 13
HCLint	0.337	4 / 12
Solubility	0.380	13 / 43
CYP	0.198	2 / 2
Lipophilicity	0.497	4 / 8

Supplementary Table S2. Model configuration details for each model variant.

Model	**Layers**	**Hidden Size**	**Heads**	**Intermediate Size**
5M	8	256	4	384
15M	12	384	6	576
111M	22	768	12	1152

Supplementary Table S3. Pretraining configuration and optimizer settings. Learning rate and batch size formulas are taken from Li et al. 2025.

Parameter	**Value / Description**
Optimizer	Schedule-free AdamW
Learning rate	$\eta(N, D) = 1.79\, N^{-0.713}\, D^{0.307}$ N: number of non-embedding parameters, D: dataset size in tokens
Batch size (in tokens)	$B(D) = 0.58\, D^{0.571}$ D: dataset size in tokens
Weight decay	0.00001
Adam β_1	0.9
Adam β_2	0.999
Warmup steps	1000
Mixed precision	Enabled
Model compilation	`torch.compile` with Inductor backend
Metric for best model	Cross-entropy loss on held-out test set
Dataloader workers	32
Pin memory	True

Supplementary Table S4. Hyperparameter combinations used in the pretraining ablation study. The underlined configuration is the one predicted by the scaling rules of Li et al. 2025.

Learning Rate	**Batch Size (samples)**
$\underline{0.004255}$	$\underline{512}$
0.001	256
0.001	1024
0.1	256
0.1	1024

Supplementary Table S5. Hyperparameter search space for finetuning optimization using TPE.

Hyperparameter	**Search Range**
Learning rate	1e-4 to 1e-3 (log scale)
Weight decay	1e-6 to 1e-2 (log scale)
Warmup steps	0 to 100
Batch size	32, 64, 128, 256, 512

References

Houyi Li, Wenzhen Zheng, Qiufeng Wang, Hanshan Zhang, Zili Wang, Shijie Xuyang, Yuantao Fan, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. Predictable scale: Part i–optimal hyperparameter scaling law in large language model pretraining. arXiv preprint arXiv:2503.04715, 2025.