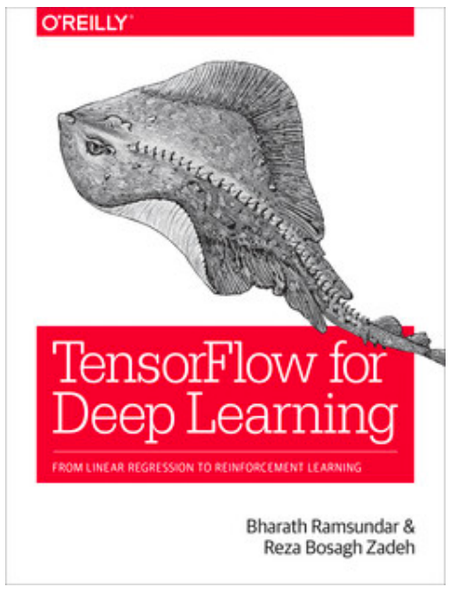
BHARATH RAMSUNDAR

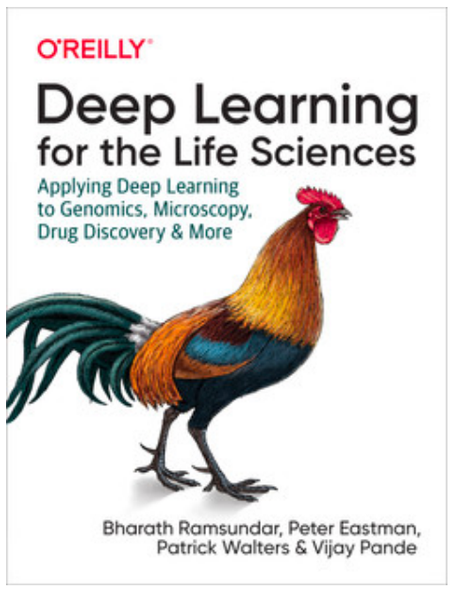
Deep Forest Sciences

TOWARDS BREAKTHROUGH GENERATIVE AI FOR CHEMISTRY

BRIEF INTRO: BHARATH RAMSUNDAR

- Stanford PhD, Pande Group, UCB EECS/Math Alum
- Lead Developer of DeepChem
- Founder/CEO of Deep Forest Sciences





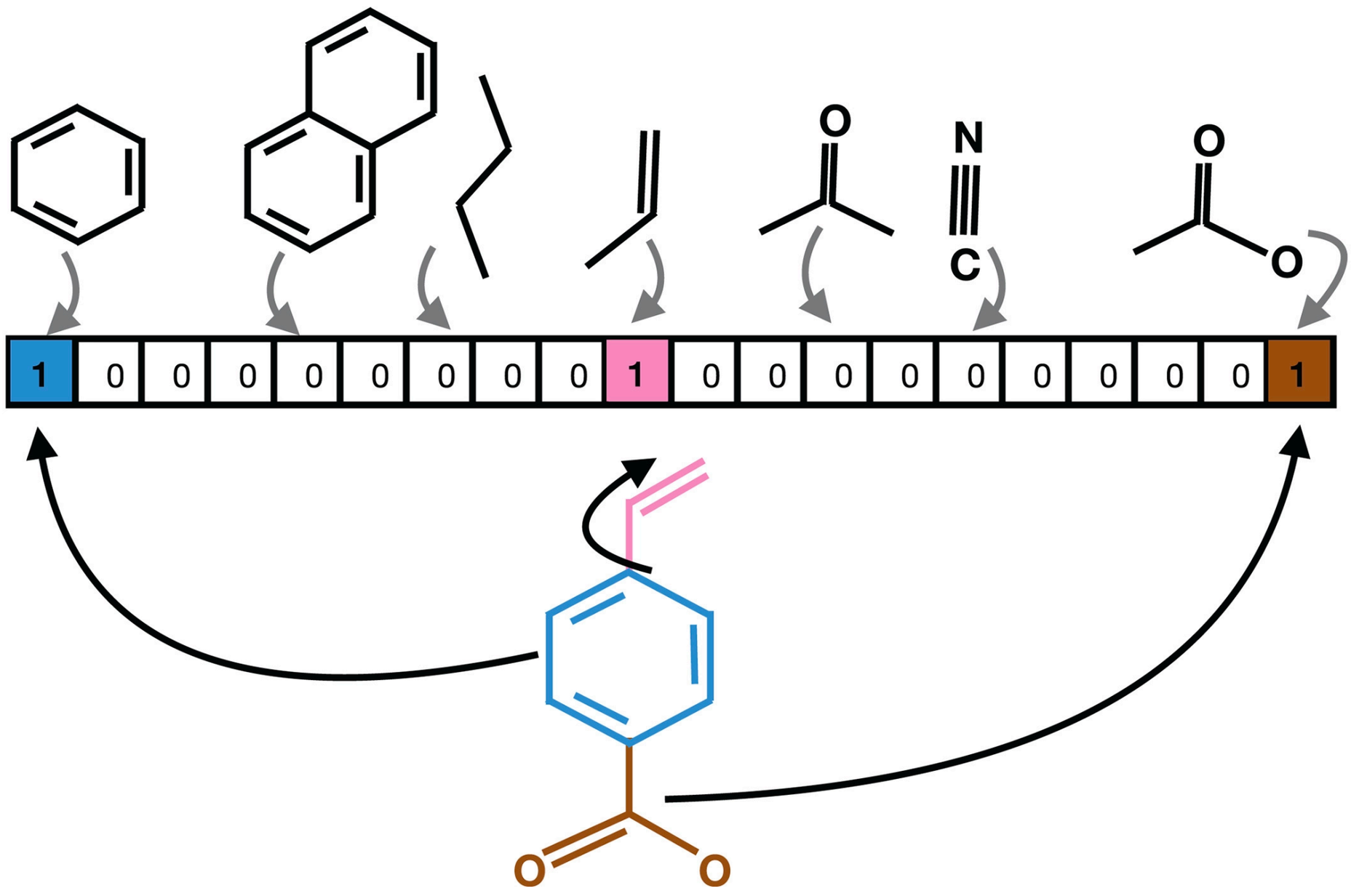
SELECTED PUBLICATIONS AND PATENTS

Research Publications	# Citations
A Guide to Deep Learning in Health Care	5520
MoleculeNet: A Benchmark for Molecular Machine Learning	3803
Low Data Drug Discovery with One-shot Learning	1024
Chemberta: Large Scale Self Supervised Retraining for Molecular Property Prediction	919
Massively Multitask Networks for Drug Discovery	679

Patents	# Citations
Non-volatile key-value store	553
Conditional iteration for a non-volatile device	302

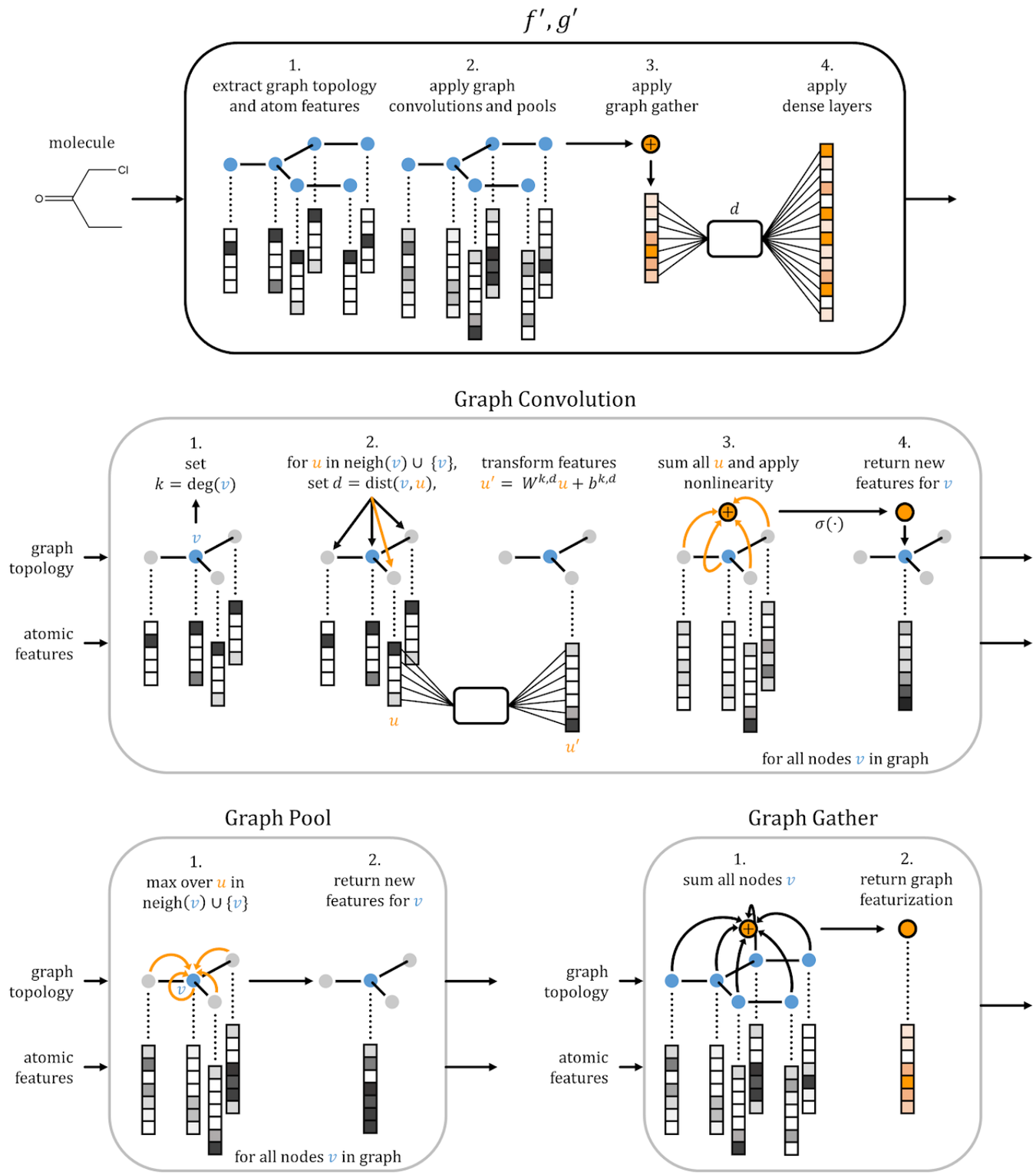
Books	Publisher
Deep Learning for the Life Sciences	O'Reilly Media
TensorFlow for Deep Learning	O'Reilly Media
Differentiable Physics	In Prep.

MOLECULAR MACHINE LEARNING 101



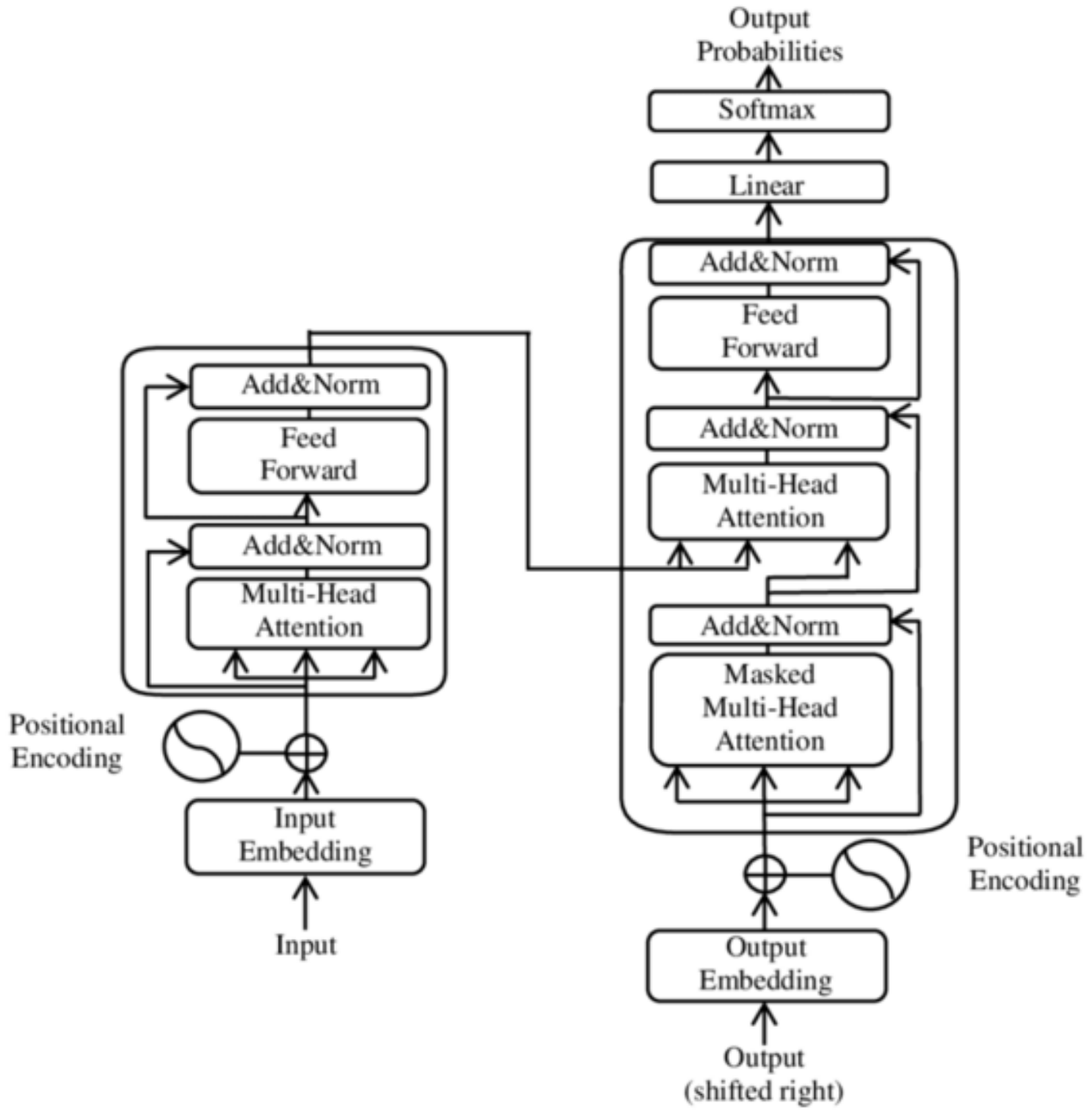
Raghunathan, Shampa, and U. Deva Priyakumar. "Molecular representations for machine learning applications in chemistry." International Journal of Quantum Chemistry 122.7 (2022): e26870.

GRAPH CONVOLUTIONS LEARN FUNCTIONS ON MOLECULES



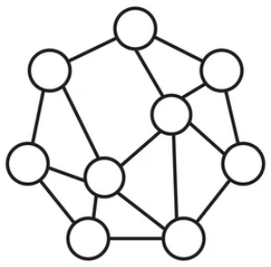
THE AGE OF GENERATIVE AI?

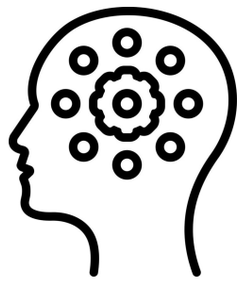
What ChatGPT thinks AI hallucination means!



HOW CAN WE MAKE GENERATIVE AI WORK FOR CHEMISTRY?

CHEMICAL FOUNDATION MODELS

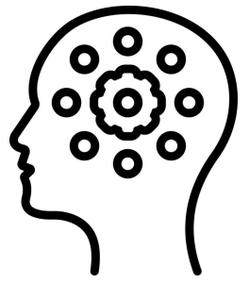




GENERATIVE AI REASONING

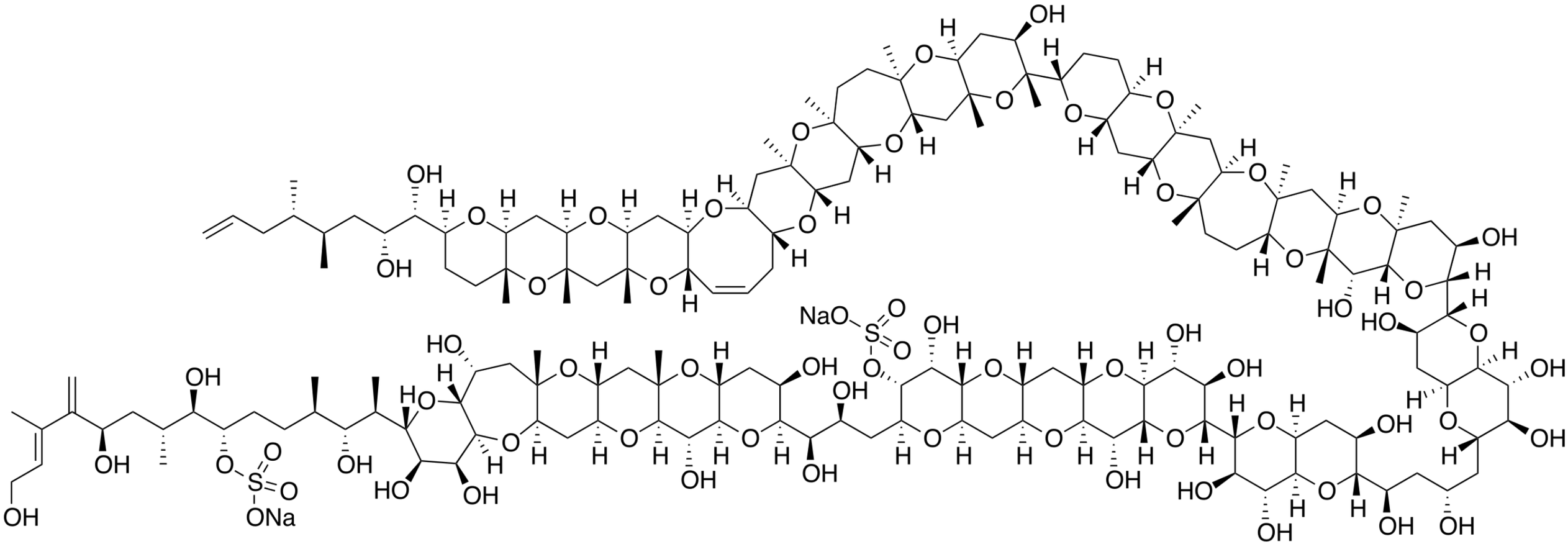
Breakthrough Generative AI for Chemistry

IMPROVED SCIENTIFIC MACHINE LEARNING



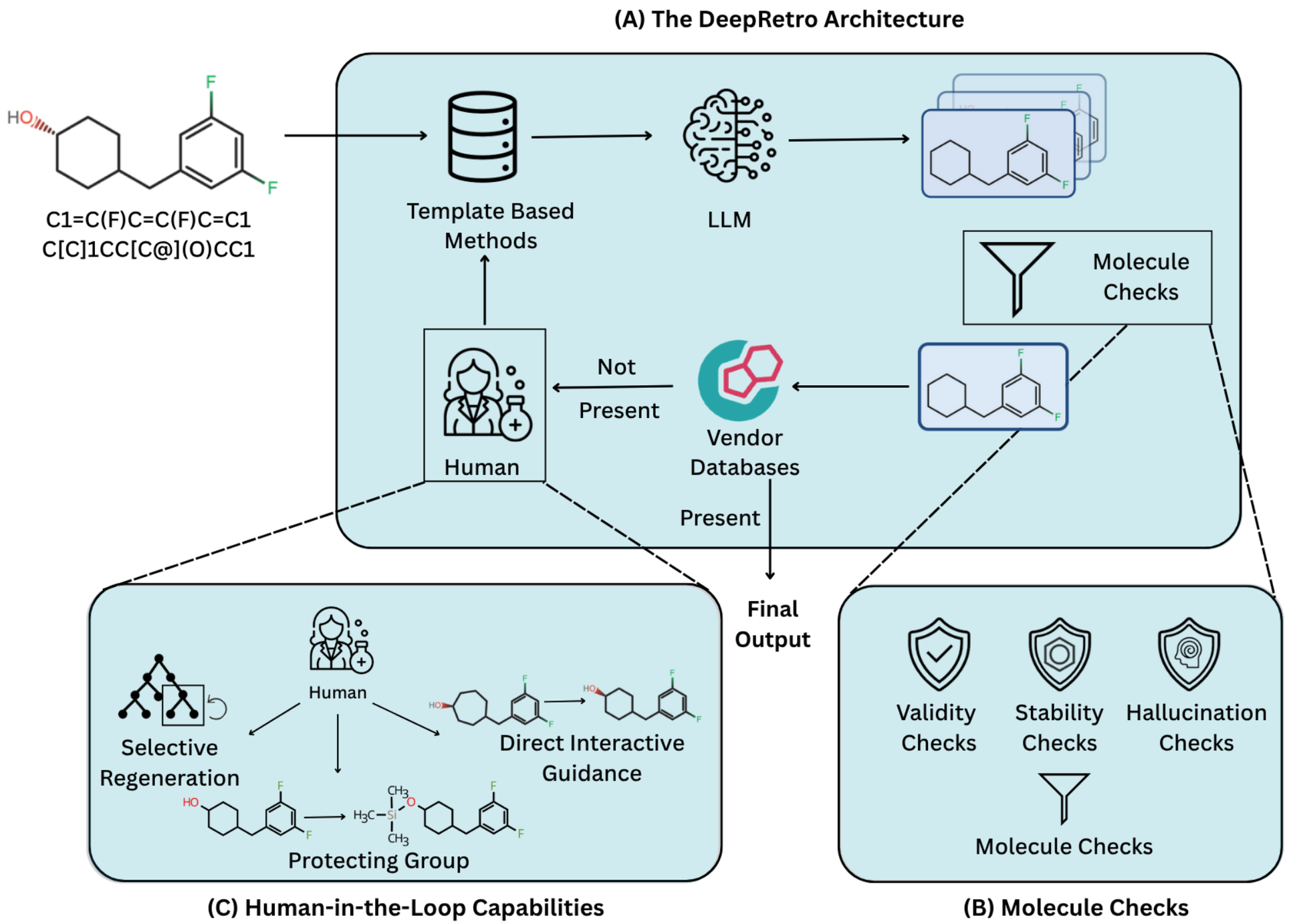
DEEPRETRO

CHEMICAL SYNTHESIS CAN BE CHALLENGING



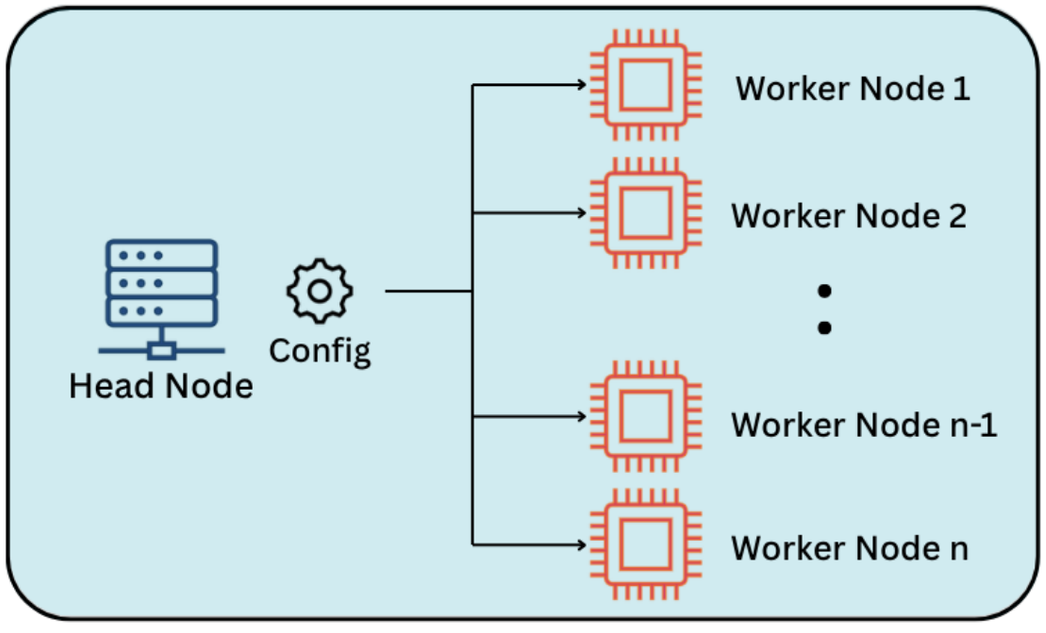
Maitotoxin

DEEPRETRO OFFERS A NOVEL GENERATIVE AI – HUMAN LOOP FOR SYNTHESIS

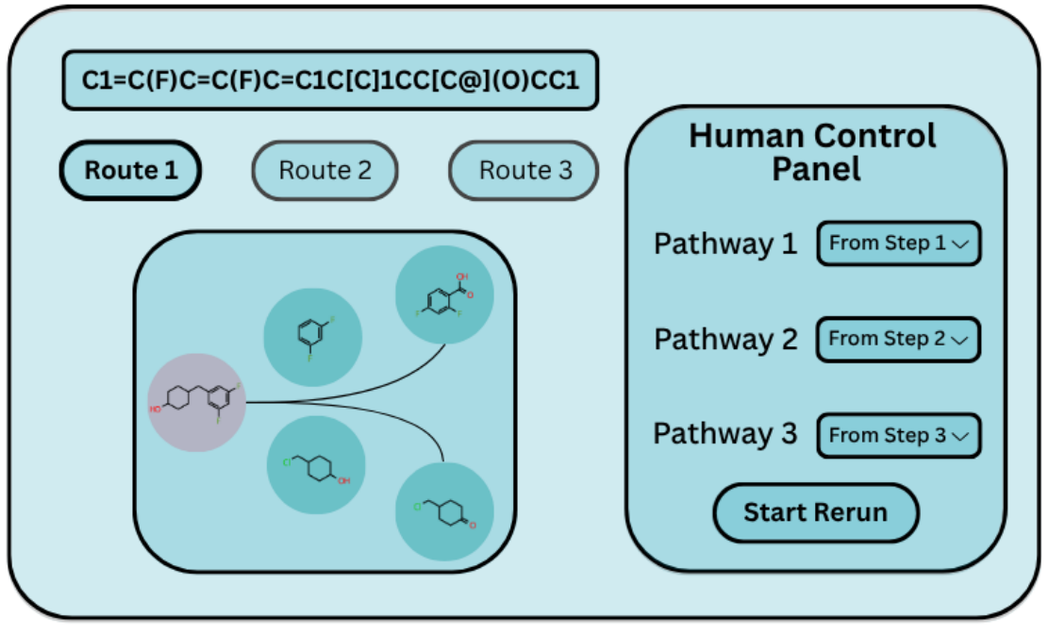


*Sathyanarayana, S. V., Hiremath, S. D., Shah, R., Panda, R., Jana, R., Singh, R., ... & **Ramsundar, B.** (2025). Deepretro: Retrosynthetic pathway discovery using iterative llm reasoning. arXiv preprint arXiv:2507.07060.*

DEEPRETRO RUNS ON SCALED CLOUD INFRASTRUCTURE WITH HUMAN UI

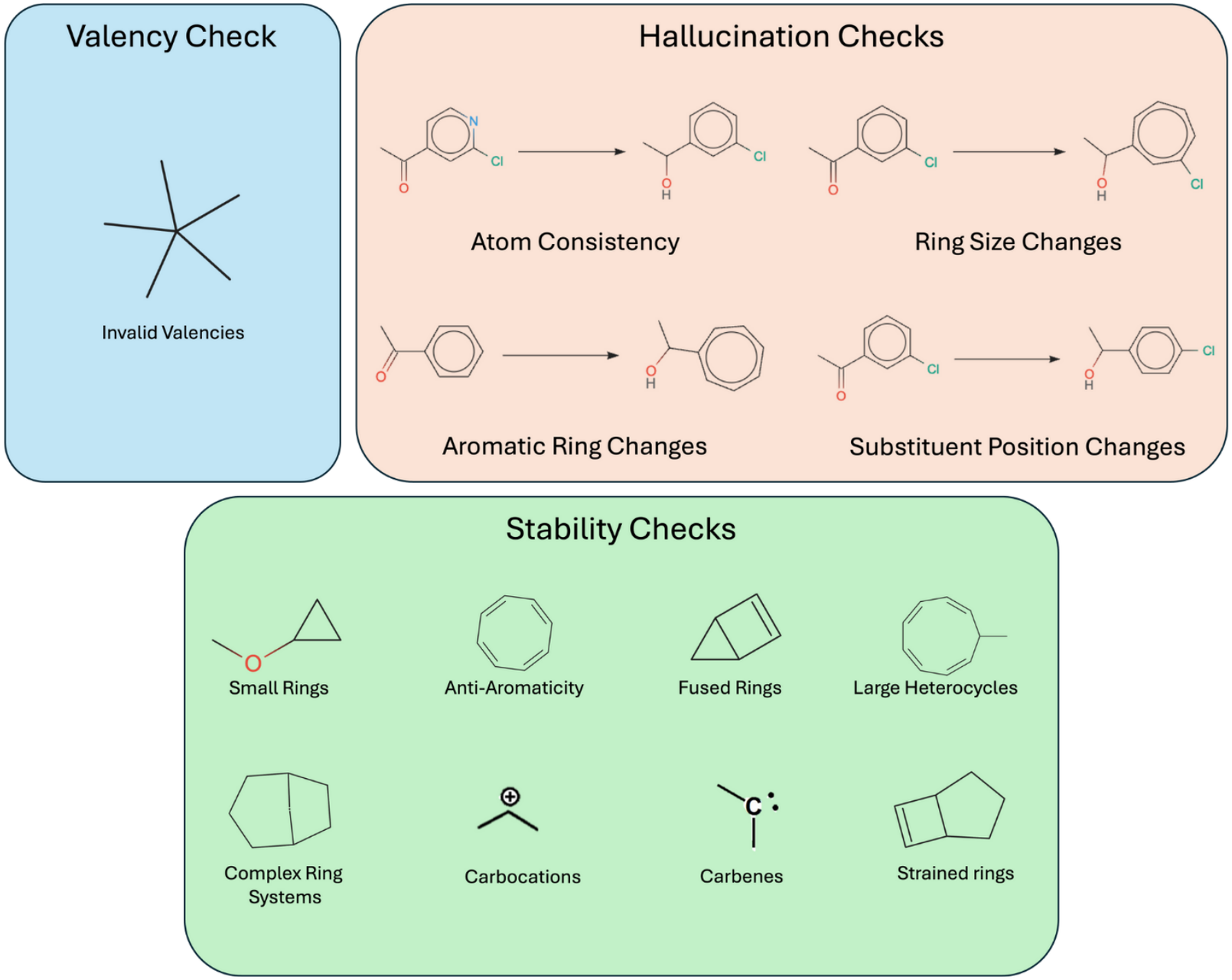


(D) Infrastructure Overview

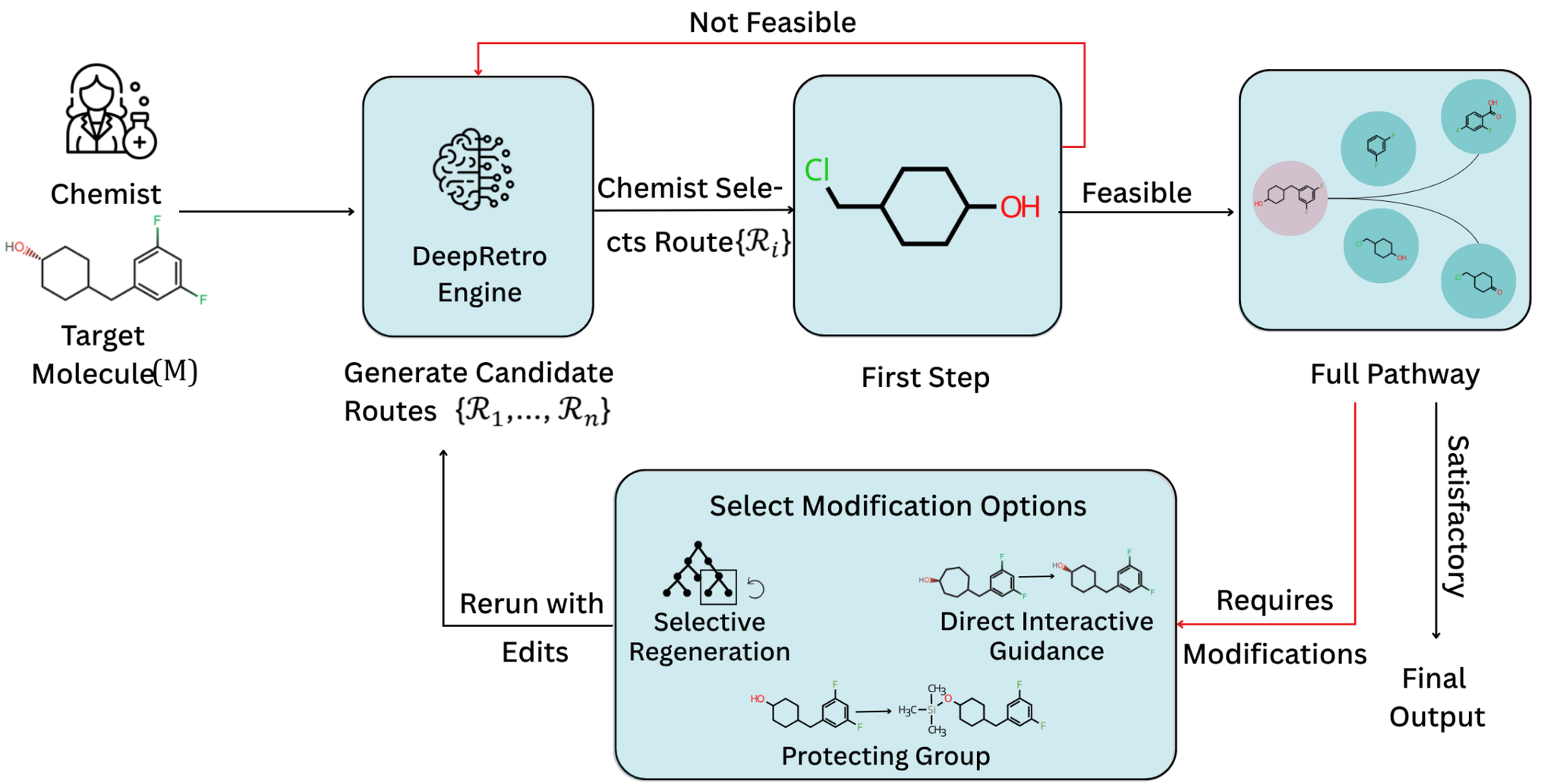


(E) User Interface

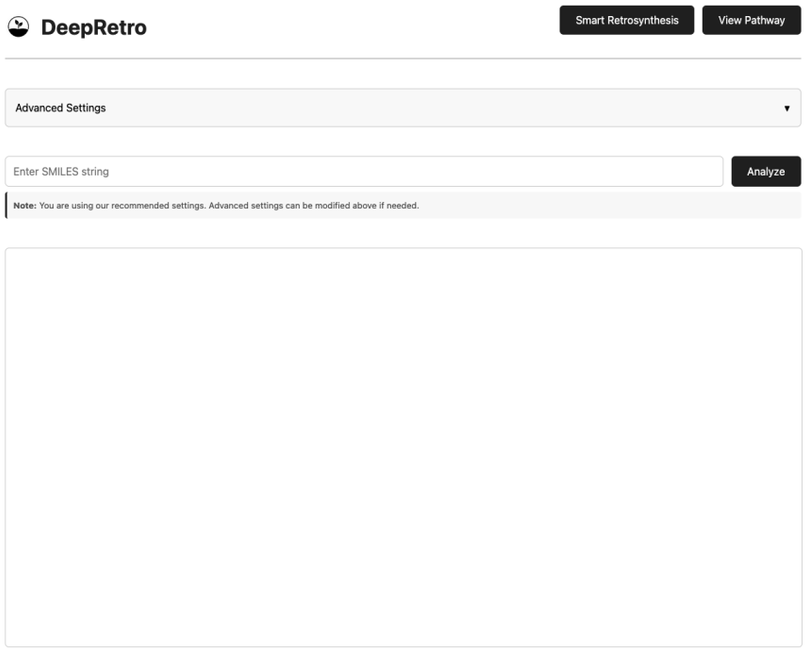
AUTOMATED VALIDATION CHECKS ATTEMPT TO CONTROL FOR HALLUCINATIONS

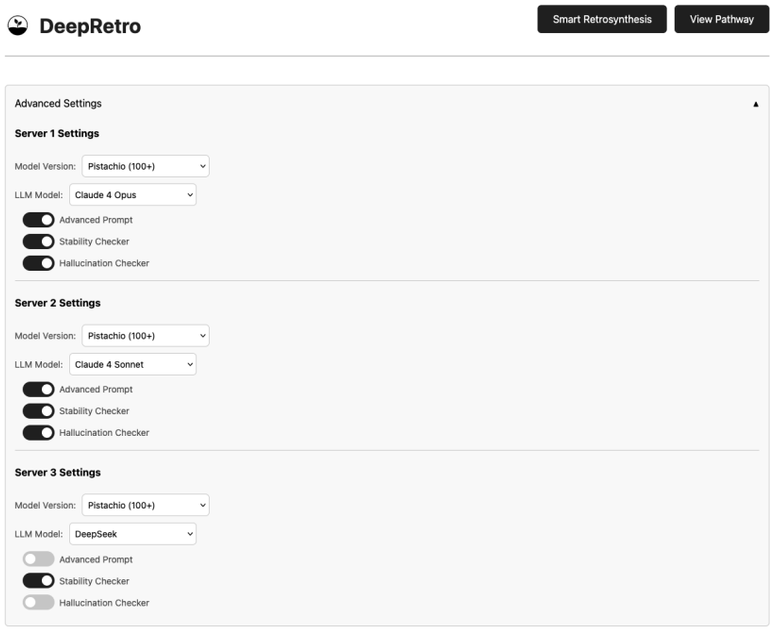


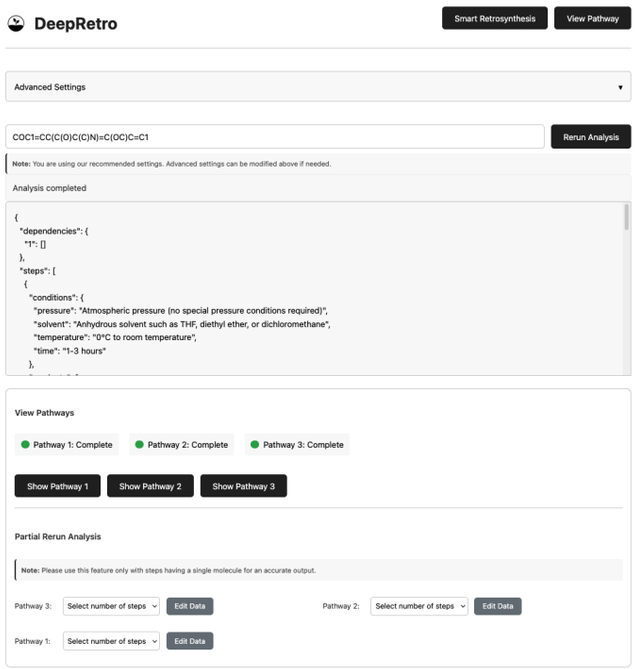
HUMAN CHEMISTS HAVE SEVERAL INTERVENTION POINTS...

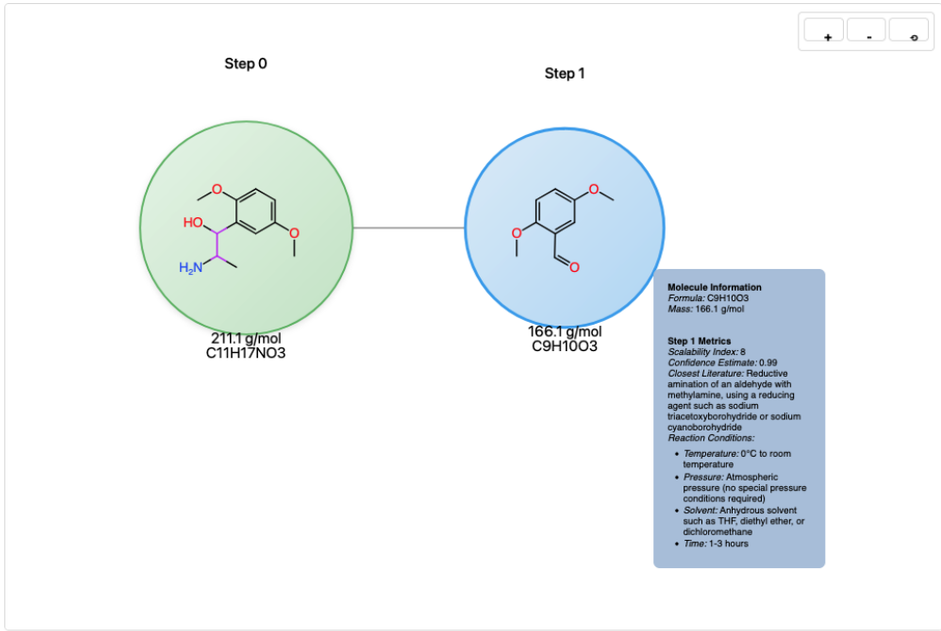


THROUGH A USER-FRIENDLY GUI









DEEPRETRO ACHIEVES STATE OF ART SINGLE STEP (AUTOMATIC) RETROSYNTHESIS

Table 1: This table showcases the Single-Step Retrosynthesis Prediction Accuracy (Top-1) on a 250 subset of USPTO-50k. The numbers reported are out of 250 tested molecules. DeepRetro's performance depends on the choice of underlying LLM and training dataset for the template-based algorithm. With Claude 4 Opus and Pistachio, DeepRetro outperforms strong baselines like ASKCOS. DeepRetro was run in automatic mode with no human intervention. The "*" represents currently deprecated models

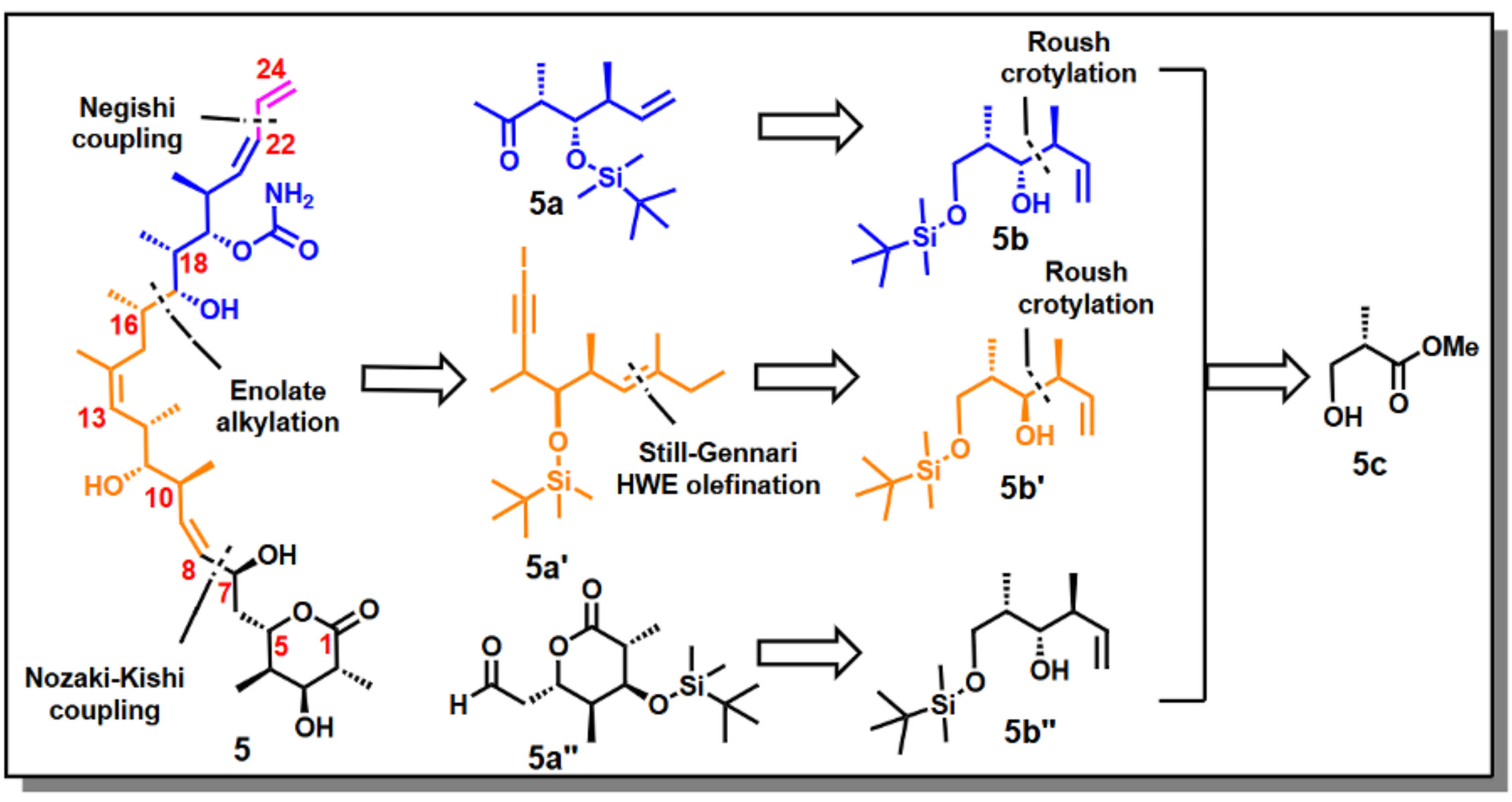
Model	LLM	Dataset	All Correct Accuracy (/250)	Any Correct Accuracy (/250)	MaxFrag Accuracy (/250)
ASKCOS	-	Reaxys	105	118	115
Aizynthfinder	-	Pistachio	73	83	79
DeepRetro	Claude 3 Opus	Pistachio	93	109	102
DeepRetro	Claude 3.5 Sonnet	Pistachio	90	102	*
DeepRetro	Claude 3.7 Sonnet	Pistachio	96	111	105
DeepRetro	Claude 4 Opus	Pistachio	**111**	**137**	**126**
DeepRetro	Claude 4 Sonnet	Pistachio	<u>107</u>	<u>129</u>	<u>120</u>
DeepRetro	DeepSeek R1	Pistachio	95	110	*
DeepRetro	GPT-5	Pistachio	101	125	118
Aizynthfinder	-	USPTO	63	70	67
DeepRetro	Claude 3 Opus	USPTO	80	90	86
DeepRetro	Claude 3.5 Sonnet	USPTO	82	89	*
DeepRetro	Claude 3.7 Sonnet	USPTO	85	95	-
DeepRetro	Claude 4 Opus	USPTO	**95**	**107**	102
DeepRetro	Claude 4 Sonnet	USPTO	<u>93</u>	<u>103</u>	99
DeepRetro	DeepSeek R1	USPTO	83	92	*

DEEPRETRO HAS STATE-OF-ART RESULTS ON A MORE REALISTIC DATASET (AUTOMATIC)

Table 3: This table showcases the number of solved molecules of Different Retrosynthesis Models on Drug Hunter 172 Dataset ([58]). The numbers reported are out of 172. DeepRetro was run in automatic mode with no human intervention.

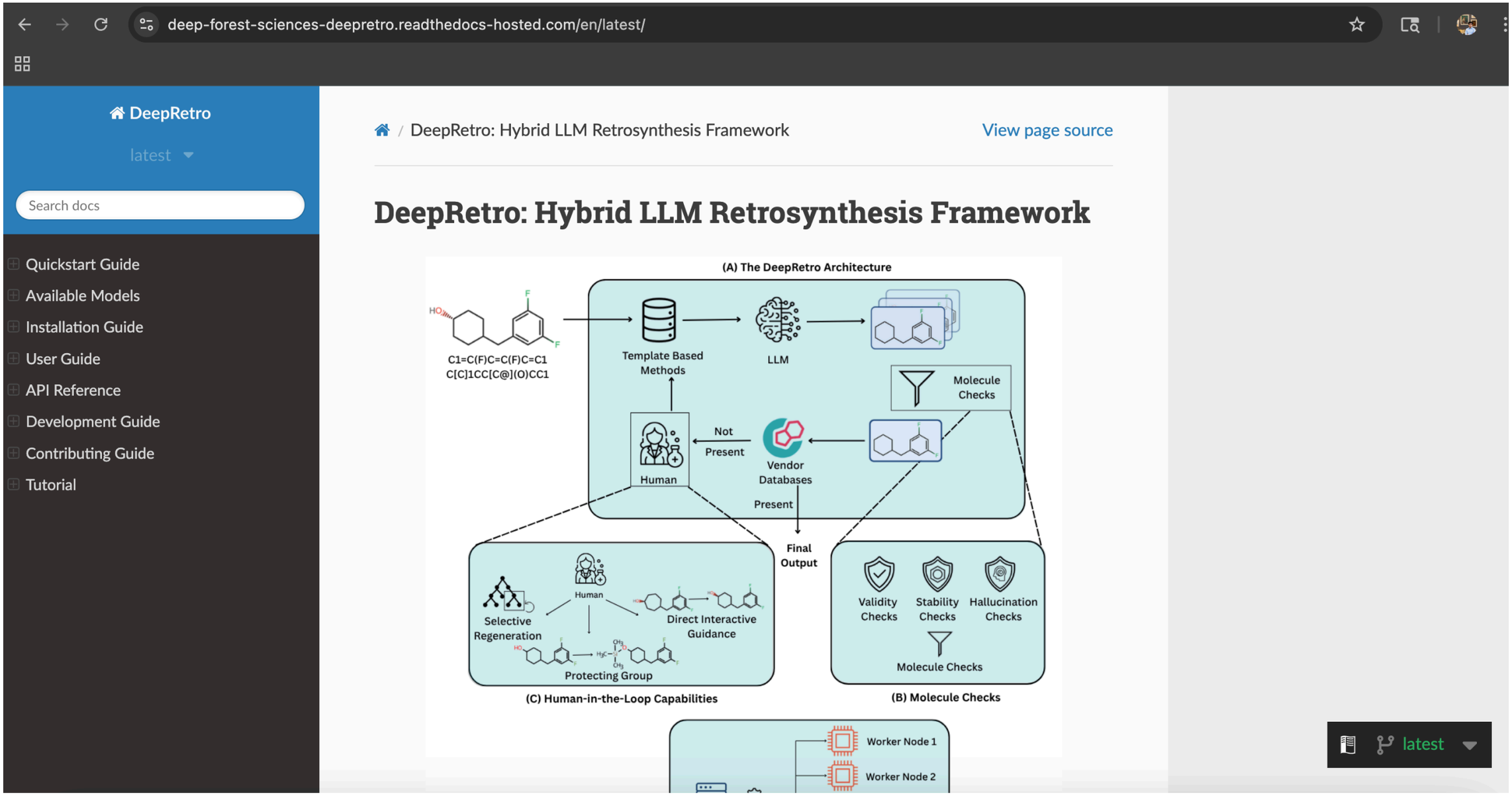
Model	LLM	Dataset	Number of solved molecules (/172)
DeepRetro	Claude 3 Opus	Pistachio	164
DeepRetro	Claude 3.7 Sonnet	Pistachio	160
DeepRetro	Claude 4 Opus	Pistachio	**168**
DeepRetro	Claude 4 Sonnet	Pistachio	<u>165</u>
DeepRetro	GPT-5	Pistachio	162
DeepRetro	Claude 4 Opus	USPTO	161
Retro* [59]	NA	USPTO	99
PDVN [60]	NA	USPTO	113

DEEPRETRO CREATES PATHWAYS TO COMPLEX MOLECULES (HUMAN-IN-THE-LOOP)

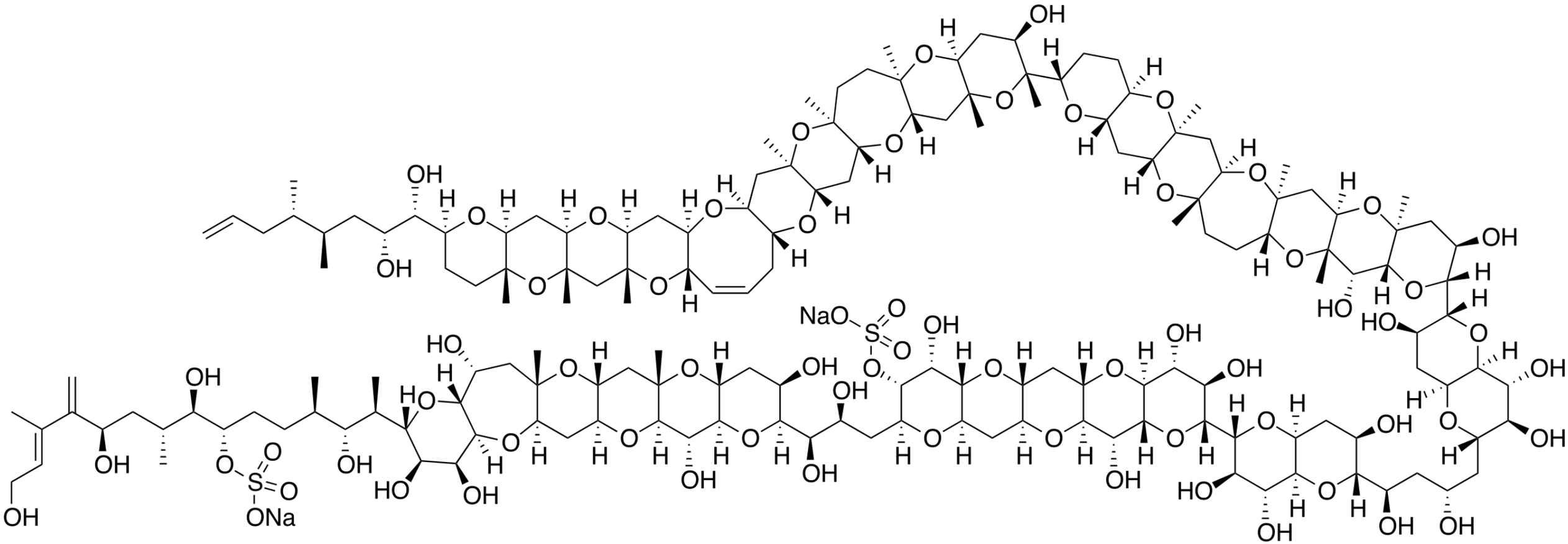


Chithrananda, Seyone, Gabriel Grand, and Bharath Ramsundar. "Chemberta: Large-scale self-supervised pretraining for molecular property prediction." arXiv preprint arXiv:2010.09885 (2020).

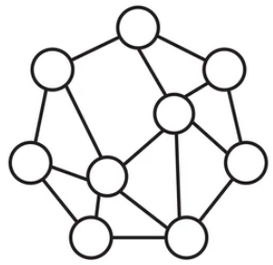
DEEPRETRO IS OPEN-SOURCE AND DOCUMENTED



A PATH TO SOLVE MAITOTOXIN?

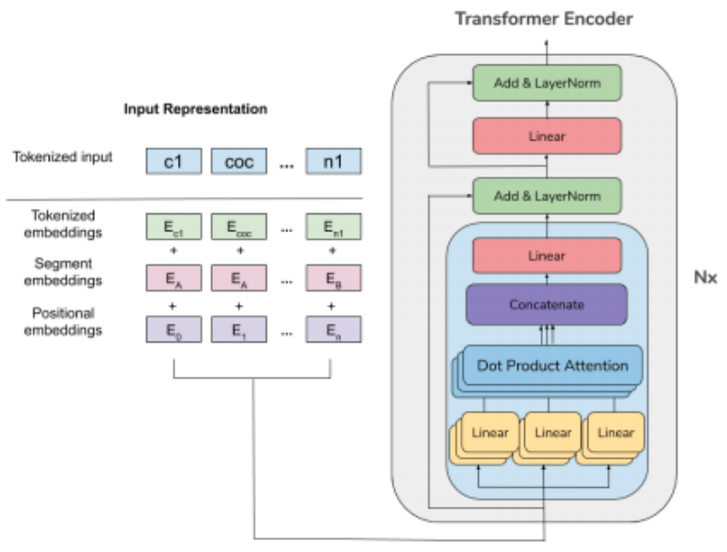


Lawrence Livermore National Laboratory



CHEMBERTA-3

CHEMBERTA (2020): SELF-SUPERVISION IMPROVES WITH MORE STRUCTURES SEEN (TO 10M)



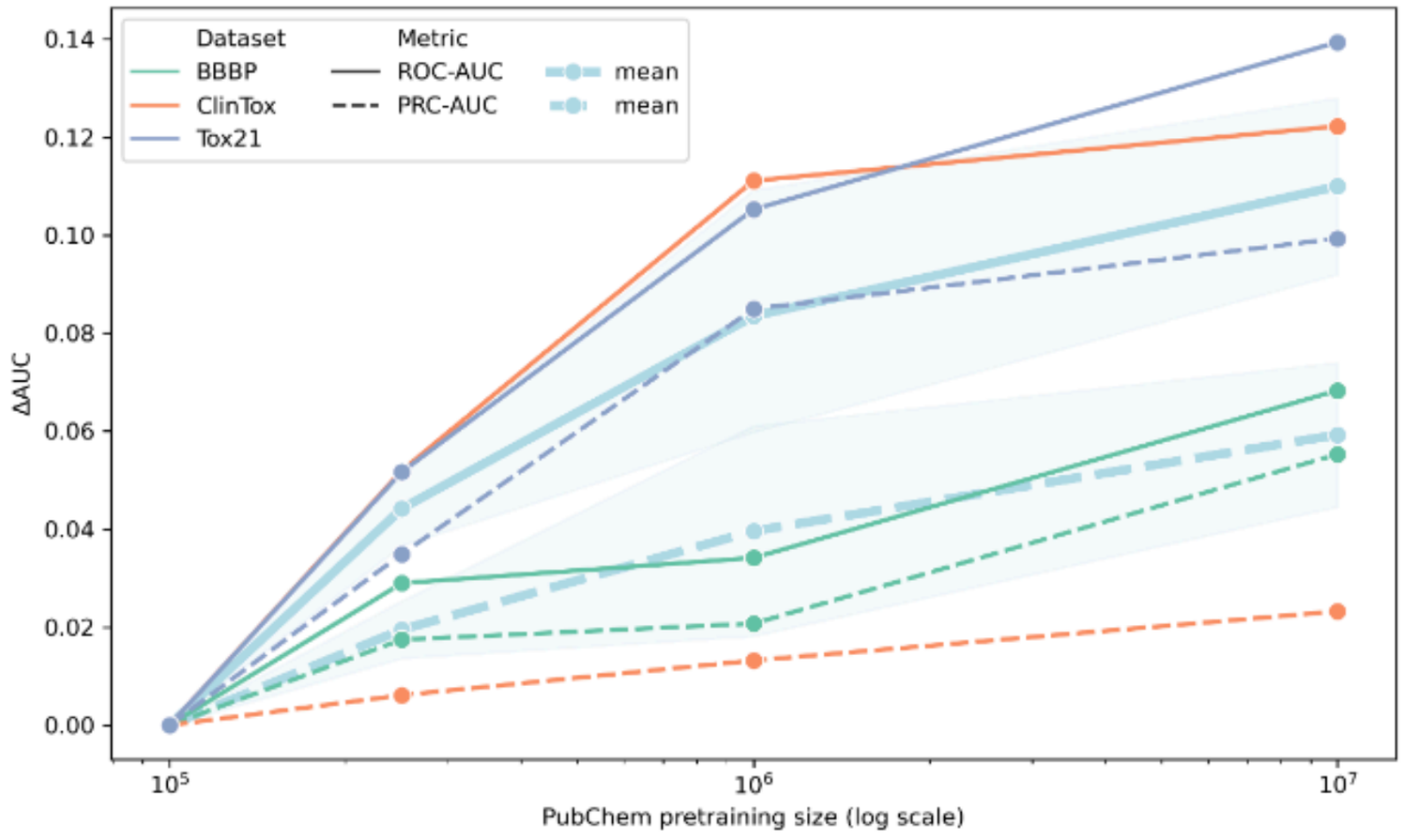
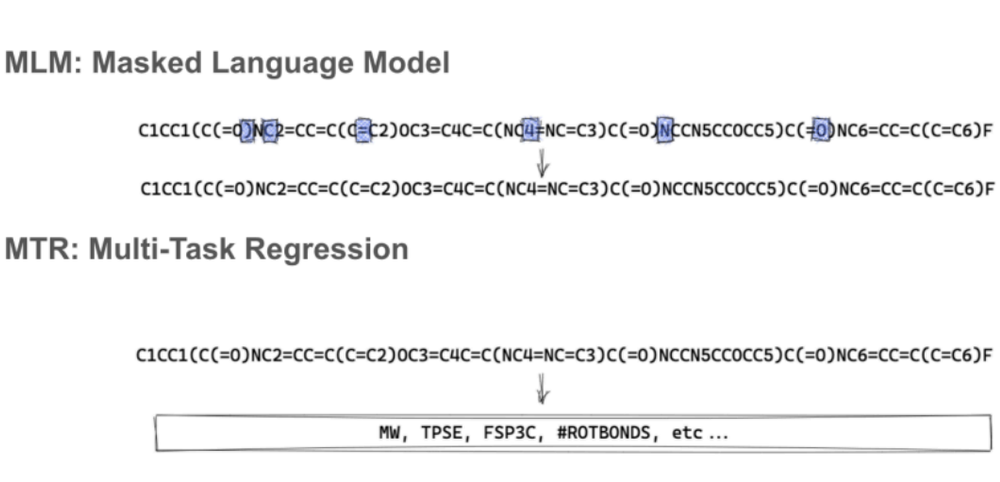
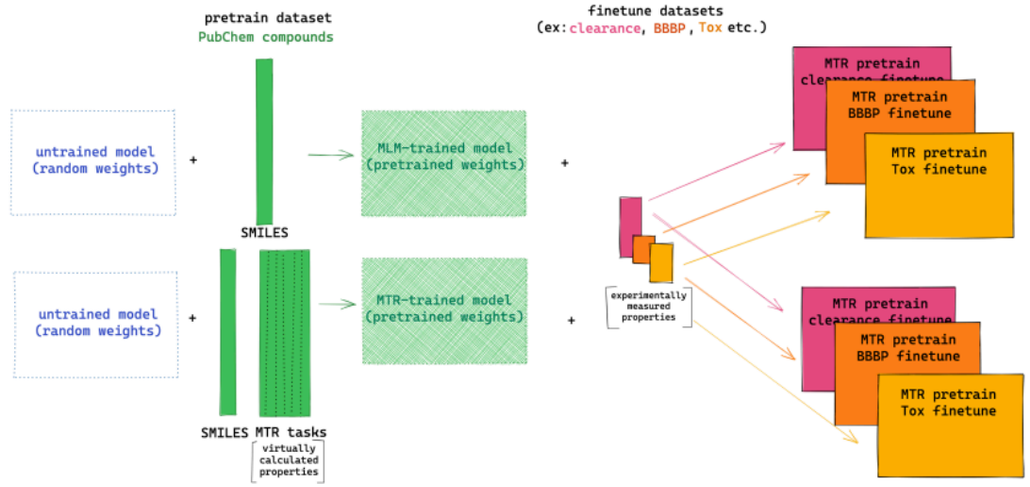


Figure 1: Scaling the pretraining size (100K, 250K, 1M, 10M) produces consistent improvements in downstream task performance on BBBP, ClinTox, and Tox21. (HIV was omitted from this analysis due to resource constraints.) Mean ΔAUC across all three tasks with a 68% confidence interval is shown in light blue.

Chithrananda, S., Grand, G., and ***Ramsundar, B.*** *"ChemBERTa: large-scale self-supervised pretraining for molecular property prediction." arXiv preprint arXiv:2010.09885 (2020).*

CHEMBERTA-2 (2022): MTR PRETRAINING AND SCALING TO 80M





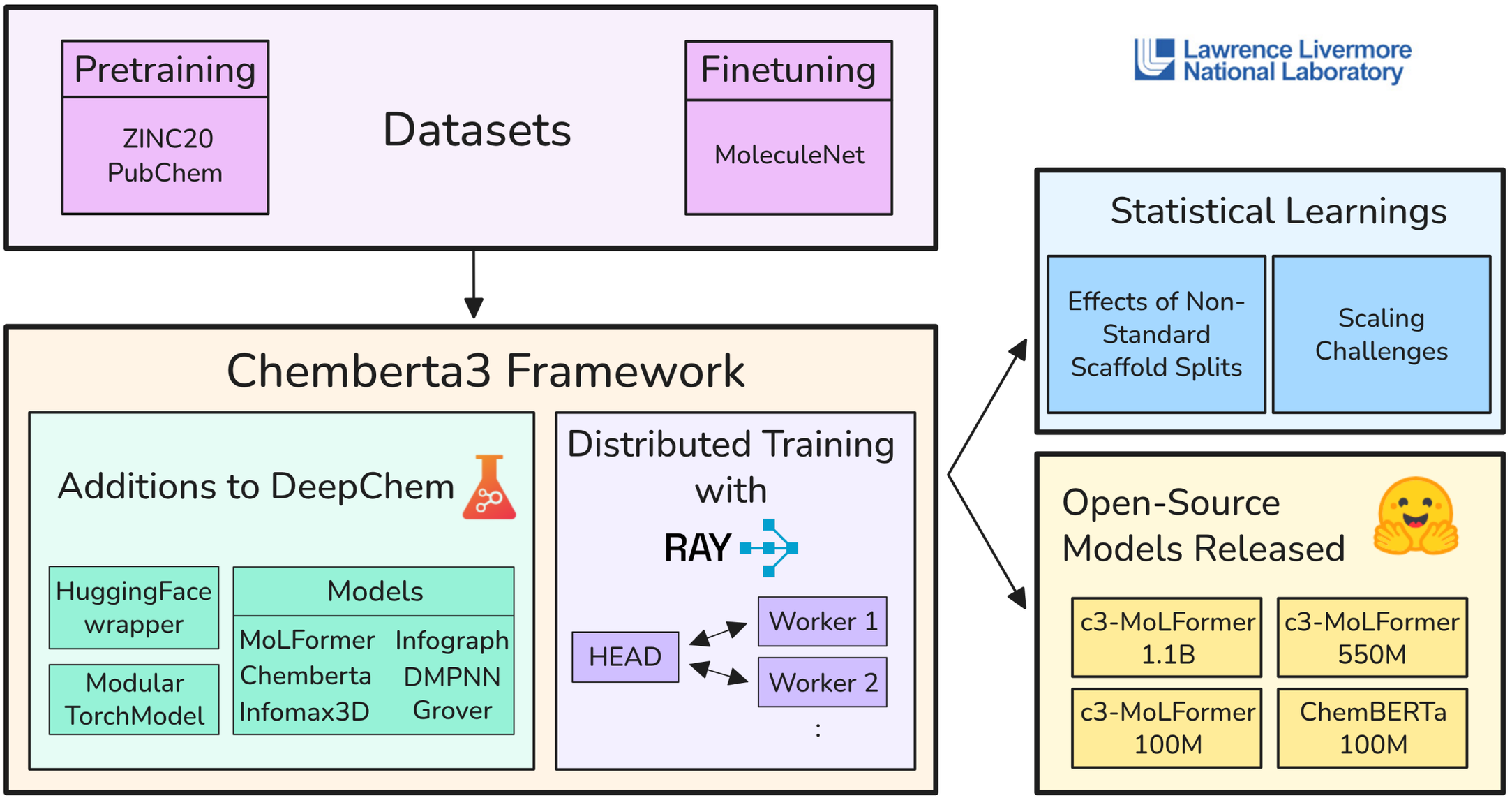
(a) MLM vs. MTR (b) Training pipeline

Ahmad, W., Simon, E., Chithrananda, S., Grand, G. and ***Ramsundar, B.****, 2022. Chemberta-2: Towards chemical foundation models. arXiv preprint arXiv:2209.01712.*

	BACE *RMSE*	**Clearance** *RMSE*	**Delaney** *RMSE*	**Lipo** *RMSE*	**BACE** *ROC*	**BBBP** *ROC*	**ClinTox** *ROC*	**SR-p53** *ROC*
D-MPNN	2.253	49.754	1.105	1.212	0.812	0.697	**0.906**	0.719
RF	**1.3178**	52.0770	1.7406	0.9621	**0.8507**	0.7194	0.7829	0.724
GCN	1.6450	51.2271	0.8851	0.7806	0.818	0.676	0.907	0.688
ChemBERTa-1						0.643	0.733	0.728
ChemBERTa-2								
MLM-5M	1.451	54.601	0.946	0.986	0.793	0.701	0.341	0.762
MLM-10M	1.611	53.859	0.961	1.009	0.729	0.696	0.349	0.748
MLM-77M	1.509	52.754	1.025	0.987	0.735	0.698	0.239	0.749
MTR-5M	1.477	50.154	0.874	0.758	0.734	**0.742**	0.552	**0.834**
MTR-10M	1.417	48.934	**0.858**	**0.744**	0.783	0.733	0.601	0.827
MTR-77M	1.363	**48.515**	0.889	0.798	0.799	0.728	0.563	0.817

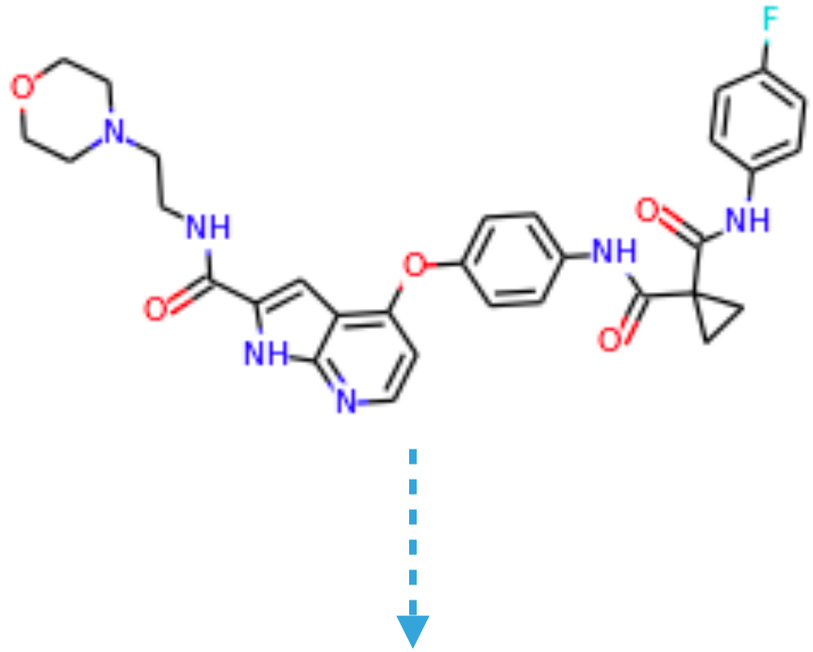
Table 1: Comparison of ChemBERTa-2 pretrained on different tasks (MLM and MTR) and on different dataset sizes (5M, 10M, and 77M), vs. existing architectures on selected MoleculeNet tasks. We report ROC-AUC (↑) for classification and RMSE (↓) for regression tasks. D-MPNNs were trained with the `chemprop` [20] library. We could not benchmark easily against Grover [11] due to differences in benchmarking procedures.

CHEMBERTA3: INFRASTRUCTURE TO TRAIN UP TO 1.1B AND BENCHMARK RIGOROUSLY



*Singh, R., Barsainyan, A. A., Irfan, R., Amorin, C. J., He, S., Davis, T., ... & **Ramsundar, B**. (2025). ChemBERTa-3: An Open Source Training Framework for Chemical Foundation Models.*

DEEPCHEM'S FRAMEWORK FOR PRETRAINING MODELS

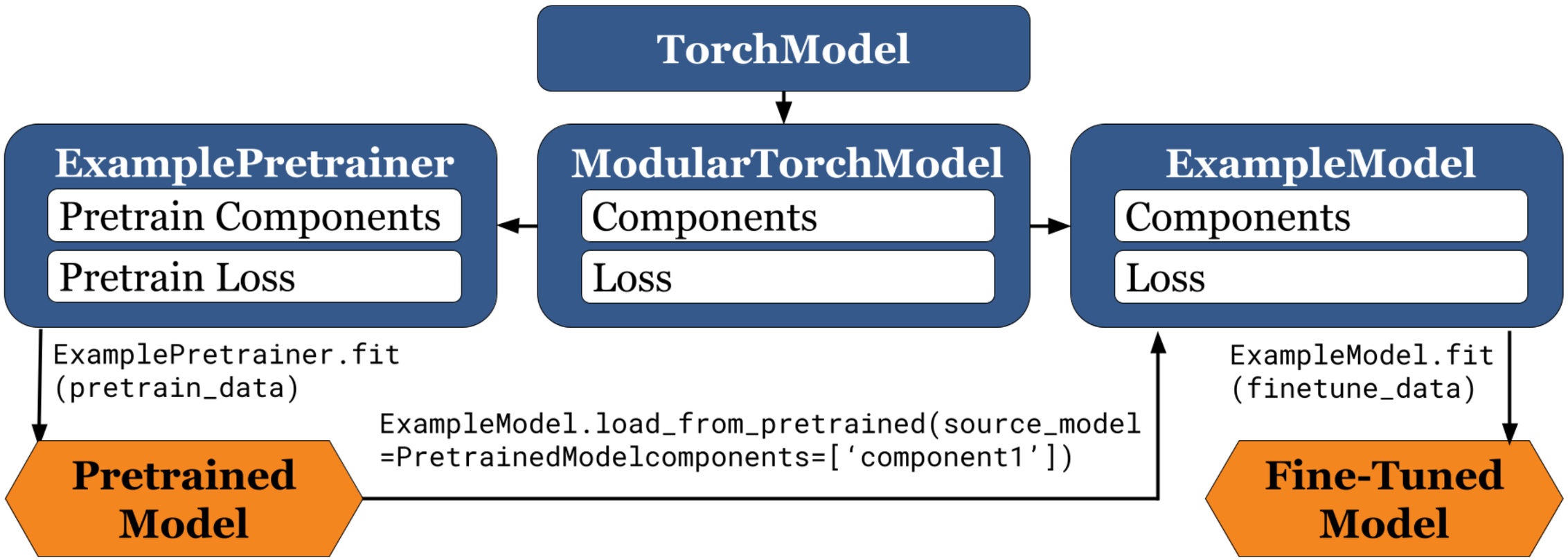


C1CC1(C(=O)NC2=CC=C(C=C2)OC3=C4C=C(NC4=NC=C3)C(=O)NCCN5CCOCC5)C(=O)NC6=CC=C(C=C6)F

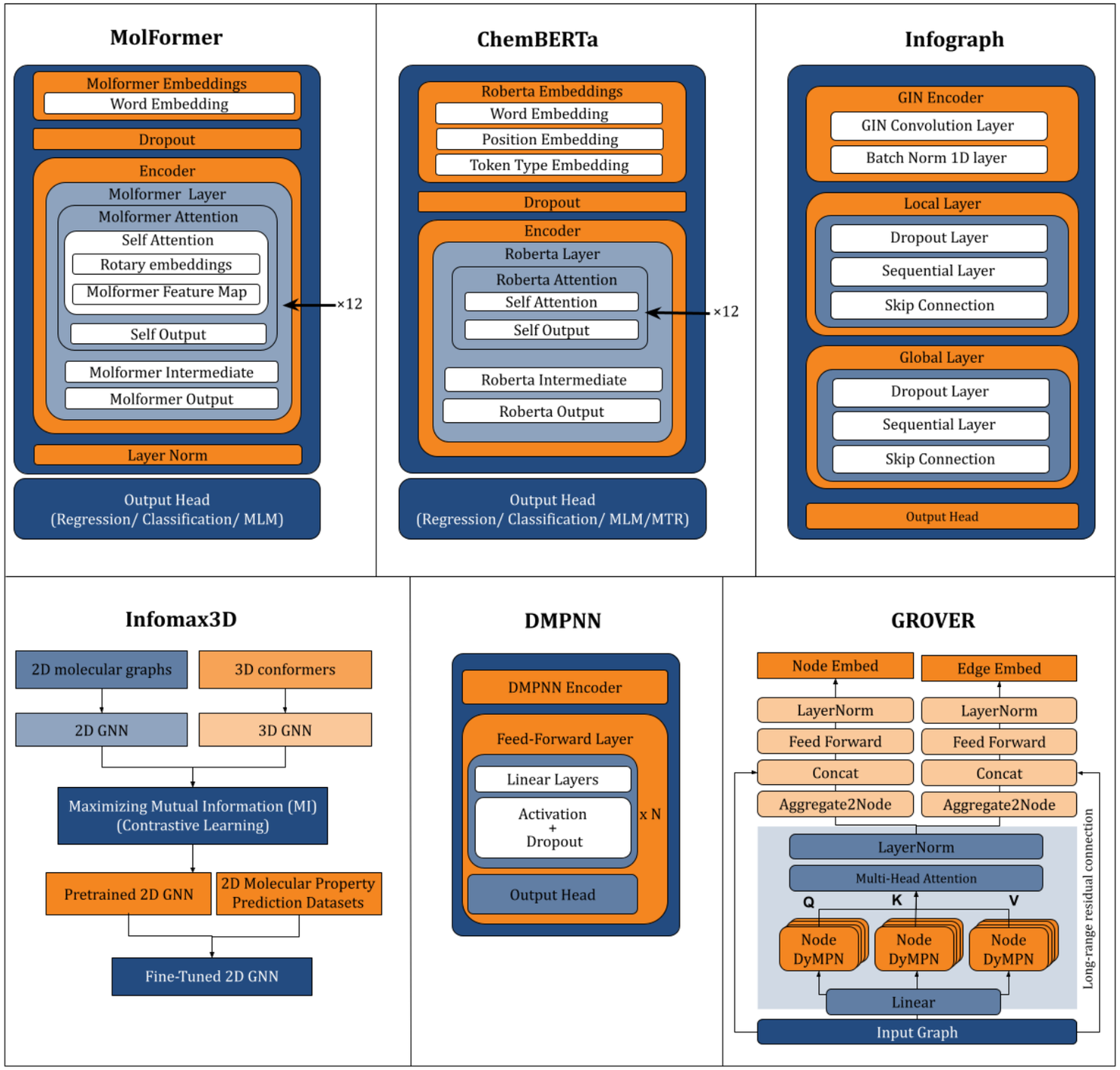
Fill in Missing Atoms and Bonds

Predict Basic Chemical Properties

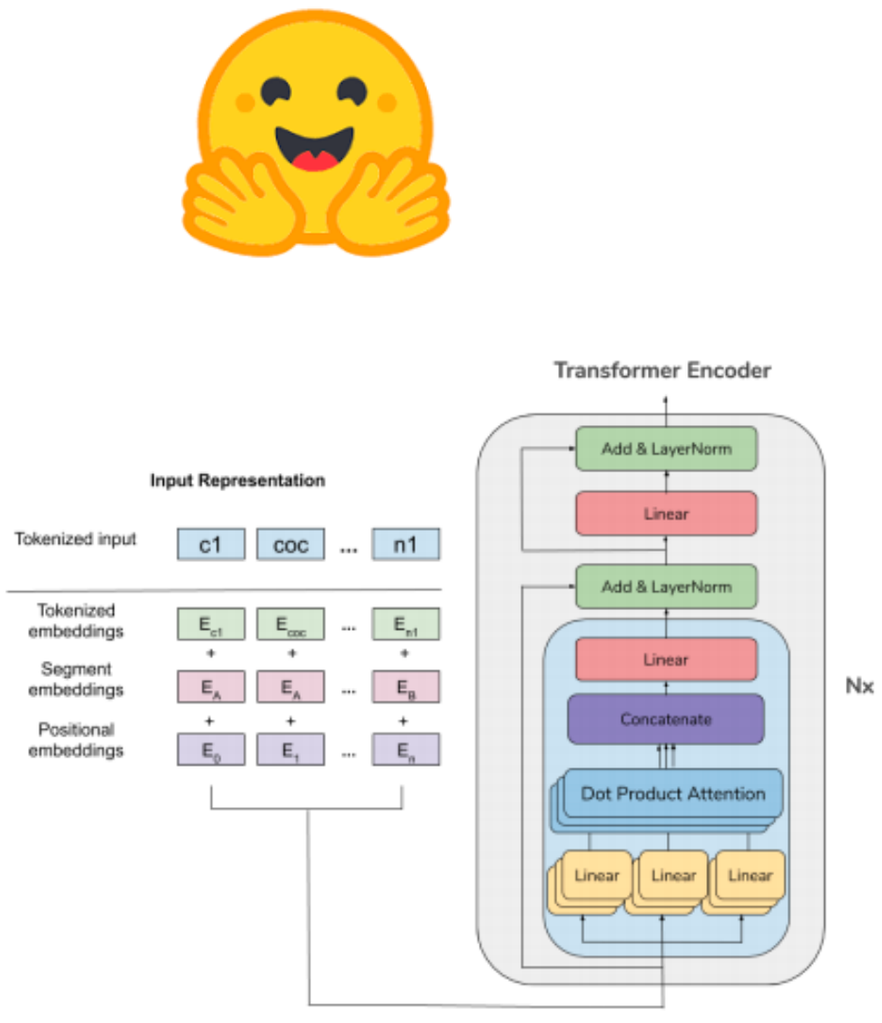
DEEPCHEM'S FRAMEWORK FOR PRETRAINING MODELS

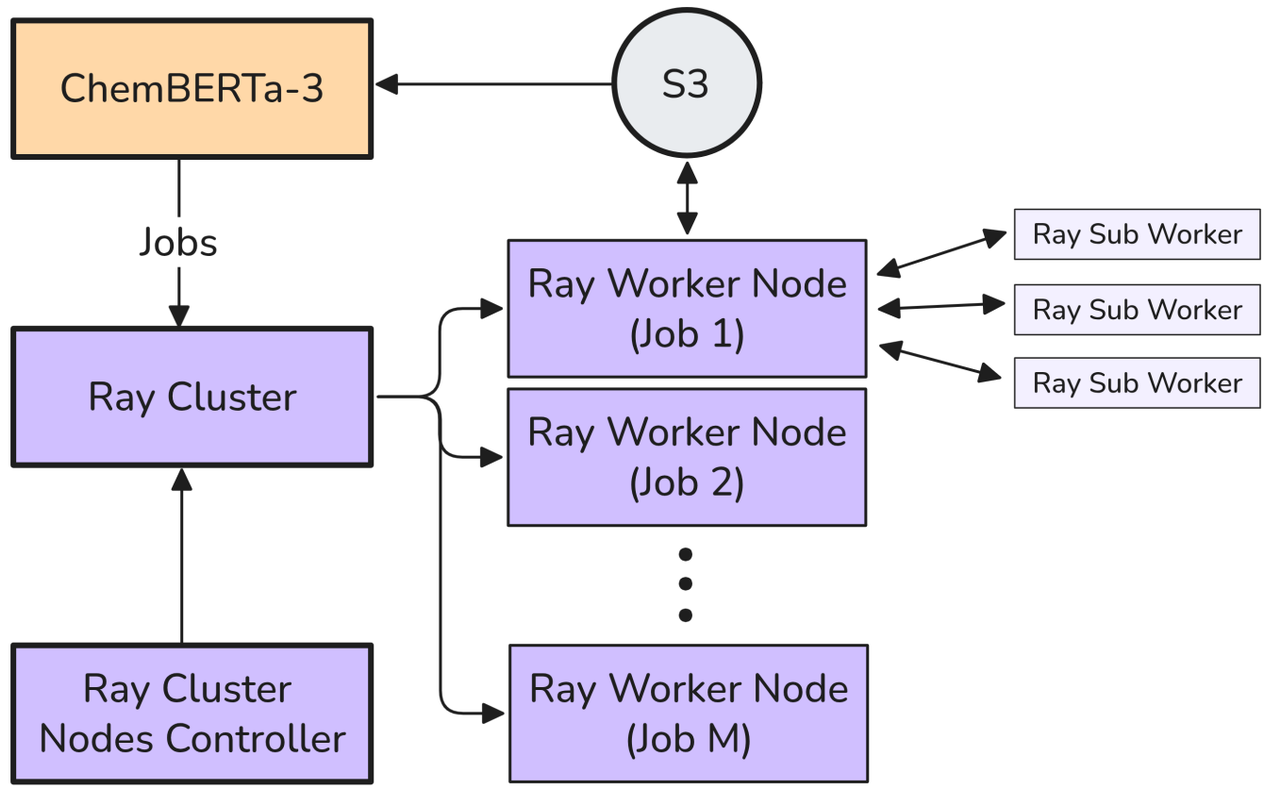


SEVERAL NEW MODELS AND PRETRAINING METHODS ADDED TO DEEPCHEMM

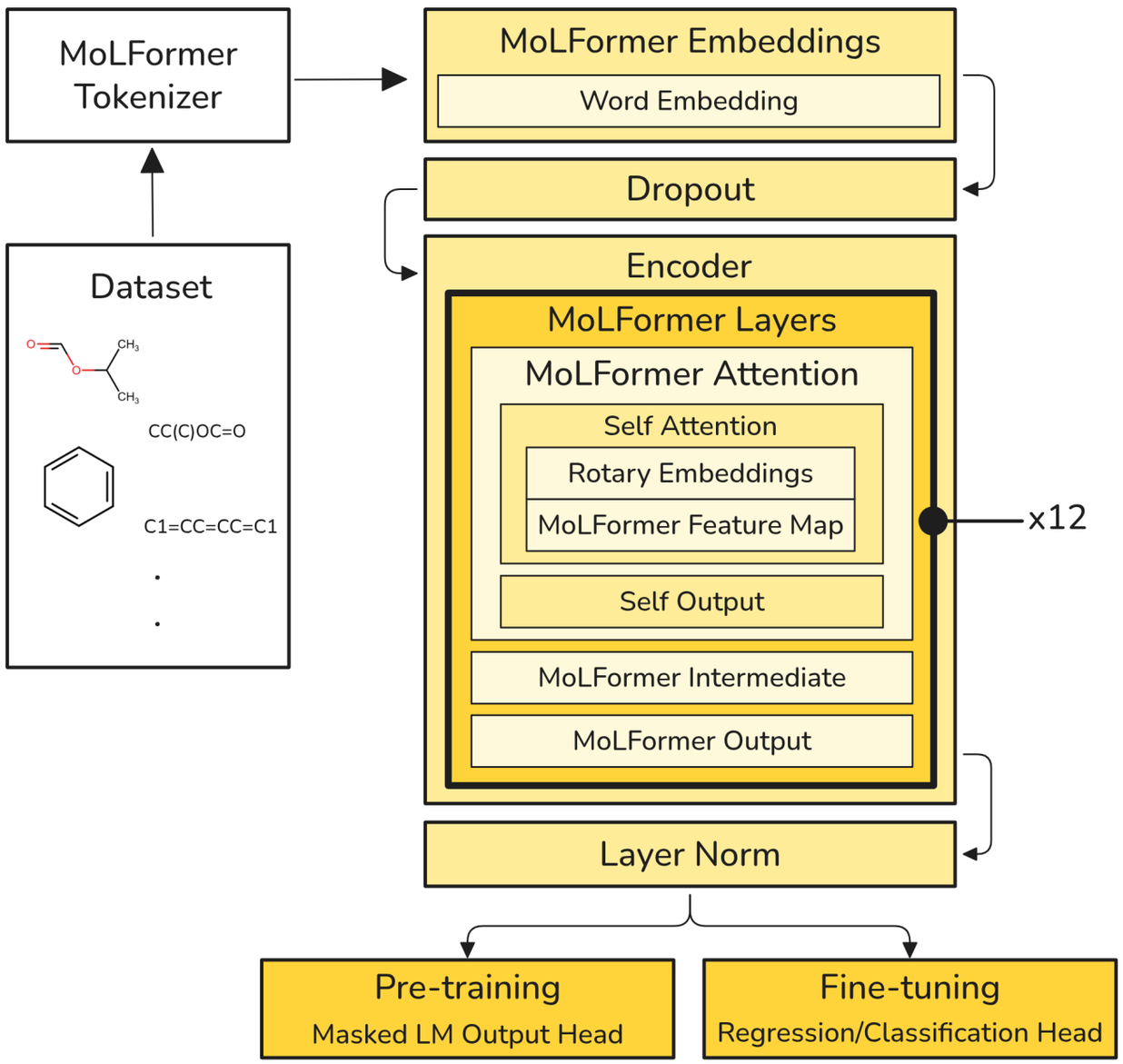


NEW INFRASTRUCTURE TO INTEGRATE DEEPCHEM, HUGGINGFACE, AND RAY





HIGHLIGHT: MOLFORMER ARCHITECTURE



Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4(12), 1256-1264.

MOLFORMER SCAFFOLD SPLITS ARE EASIER THAN DEEPCHEM SCAFFOLD SPLITS!

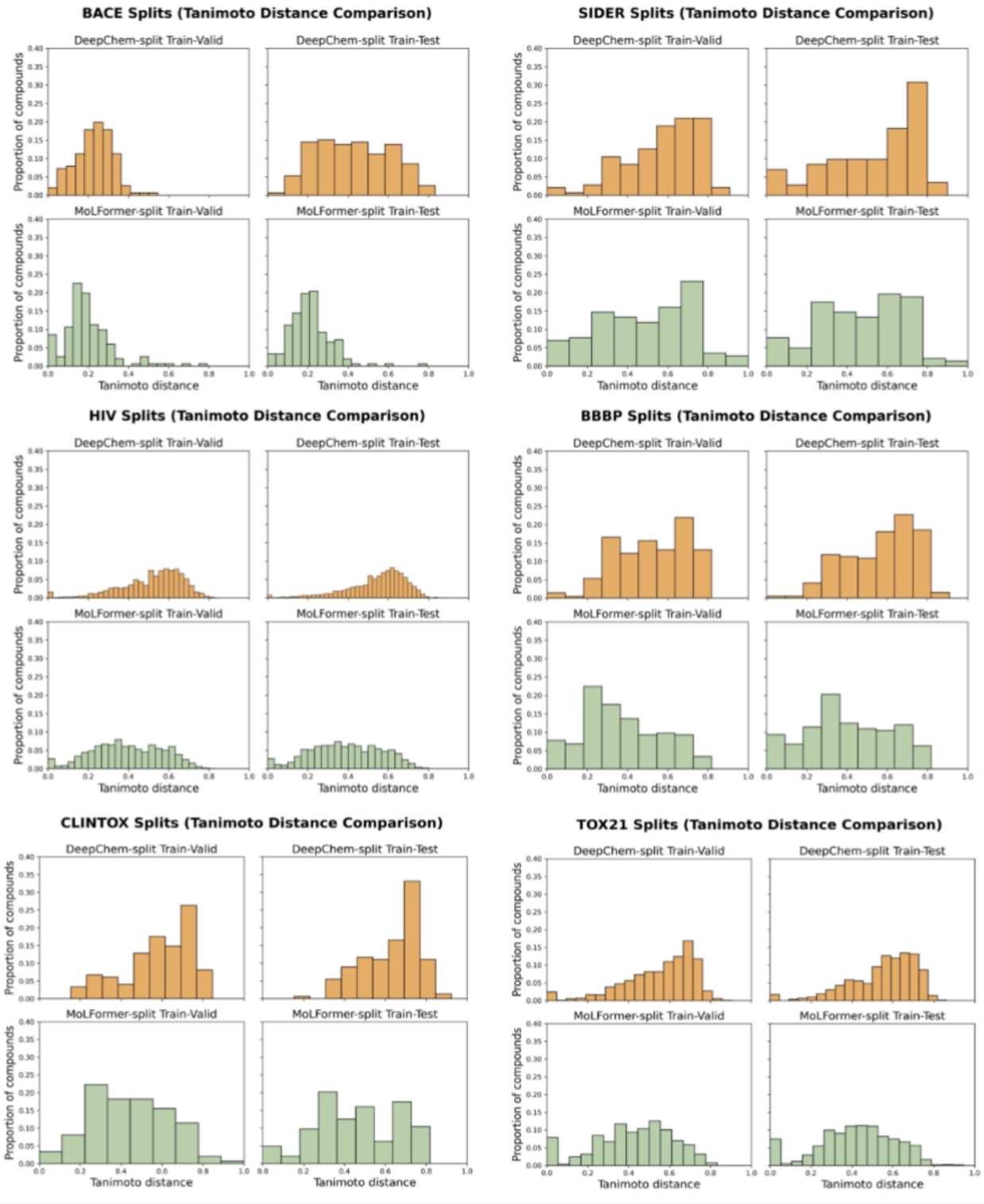


Fig. 7 Histograms of Minimum Tanimoto Distance (MTD) distributions comparing validation and test sets across multiple MoleculeNet classification datasets: BACE, SIDER, HIV, BBBP, CLINTOX and TOX21.

1.1B MODELS GENERALLY OUTPERFORM SMALLER MODELS

Table 3 The tables compare different baseline models (RF, GCN, DMPNN, Infograph, Infomax3D, and Grover) to the transformer architecture models, ChemBERTa and MoLFormer, on various **classification datasets**, in block 1 and **regression datasets**, in block 2 and repost ROC-AUC scores (Higher is better) and RMSE (Lower is better) respectively. We used the **deepchem scaffold splitter** to split the datasets provided by MoleculeNet. Here, c3-MoLFormer indicates that the MoLFormer model is trained using Chemberta3 infrastructure and MoLFormer-LHPC is trained using the HPC clusters. (green=top ranked, yellow=second rank)

Classification Datasets (Higher is better)

Dataset Tasks	BACE ↑ 1	BBBP ↑ 1	TOX21 ↑ 12	HIV ↑ 1	SIDER ↑ 27	CLINTOX ↑ 2
Random Forest	0.866 ± 0.004	0.694 ± 0.013	0.674 ± 0.007	0.794 ± 0.007	0.630 ± 0.002	0.689 ± 0.011
GCN	0.778 ± 0.008	0.642 ± 0.011	0.710 ± 0.005	0.759 ± 0.007	0.613 ± 0.010	0.870 ± 0.020
DMPNN	0.626 ± 0.004	0.661 ± 0.001	0.706 ± 0.001	0.752 ± 0.007	0.524 ± 0.029	0.642 ± 0.005
Infograph-250K	0.739 ± 0.019	0.639 ± 0.054	0.684 ± 0.010	0.755 ± 0.007	0.627 ± 0.010	0.845 ± 0.004
Infomax3D-250K	0.658 ± 0.008	0.624 ± 0.020	0.645 ± 0.006	0.704 ± 0.056	0.588 ± 0.010	0.860 ± 0.023
Grover-250K	0.825 ± 0.006	0.674 ± 0.006	0.692 ± 0.003	0.759 ± 0.002	0.619 ± 0.010	0.642 ± 0.020
ChemBERTa-MLM-100M	0.781 ± 0.019	0.700 ± 0.027	0.718 ± 0.011	0.740 ± 0.013	0.611 ± 0.002	0.979 ± 0.022
c3-MoLFormer-1.1B	0.819 ± 0.018	0.735 ± 0.019	0.723 ± 0.012	0.762 ± 0.005	0.618 ± 0.005	0.839 ± 0.013
MoLFormer-LHPC	0.887 ± 0.004	0.908 ± 0.013	0.791 ± 0.014	0.750 ± 0.003	0.622 ± 0.007	0.993 ± 0.004

Regression Datasets (Lower is better)

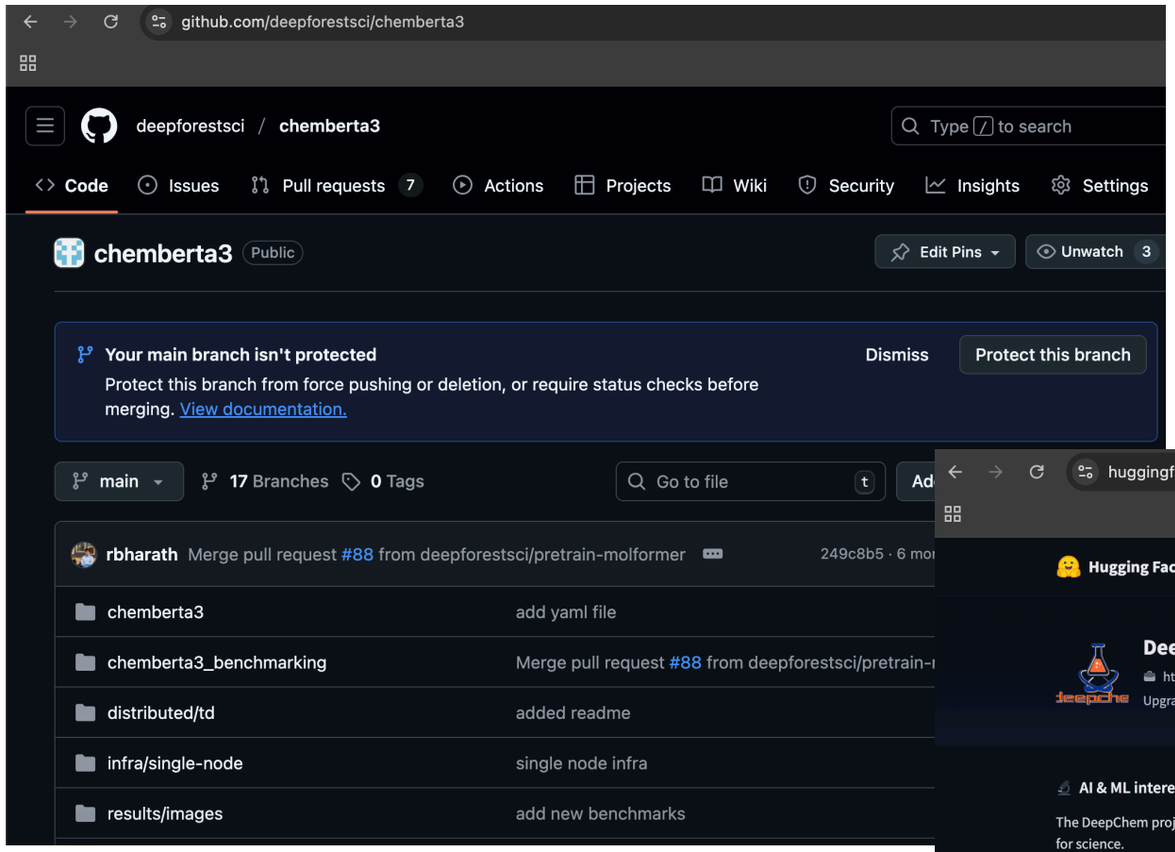
Dataset	ESOL↓	FREESOLV ↓	LIPO ↓	BACE ↓	CLEARANCE ↓
Random Forest	1.697 ± 0.005	1.138 ± 0.017	0.963 ± 0.003	1.249 ± 0.011	51.683 ± 0.402
GCN	1.002 ± 0.034	0.624 ± 0.031	0.879 ± 0.071	1.259 ± 0.028	54.599 ± 1.984
DMPNN	1.068 ± 0.033	0.596 ± 0.033	0.690 ± 0.015	1.146 ± 0.100	50.974 ± 0.542
Infograph-250K	1.410 ± 0.196	0.988 ± 0.063	0.898 ± 0.012	1.440 ± 0.137	92.646 ± 22.630
Infomax3D-250K	1.467 ± 0.013	0.623 ± 0.024	0.787 ± 0.022	1.440 ± 0.174	58.270 ± 0.642
Grover-250K	1.845 ± 0.037	1.038 ± 0.008	0.816 ± 0.027	1.563 ± 0.058	64.452 ± 0.287
ChemBERTa-MLM-100M	0.920 ± 0.011	0.536 ± 0.016	0.758 ± 0.013	1.011 ± 0.038	51.582 ± 3.079
c3-MoLFormer-1.1B	0.829 ± 0.019	0.572 ± 0.023	0.728 ± 0.016	1.094 ± 0.126	52.058 ± 2.767
MoLFormer-LHPC	0.848 ± 0.031	0.683 ± 0.040	0.895 ± 0.080	1.201 ± 0.100	45.74 ± 2.637

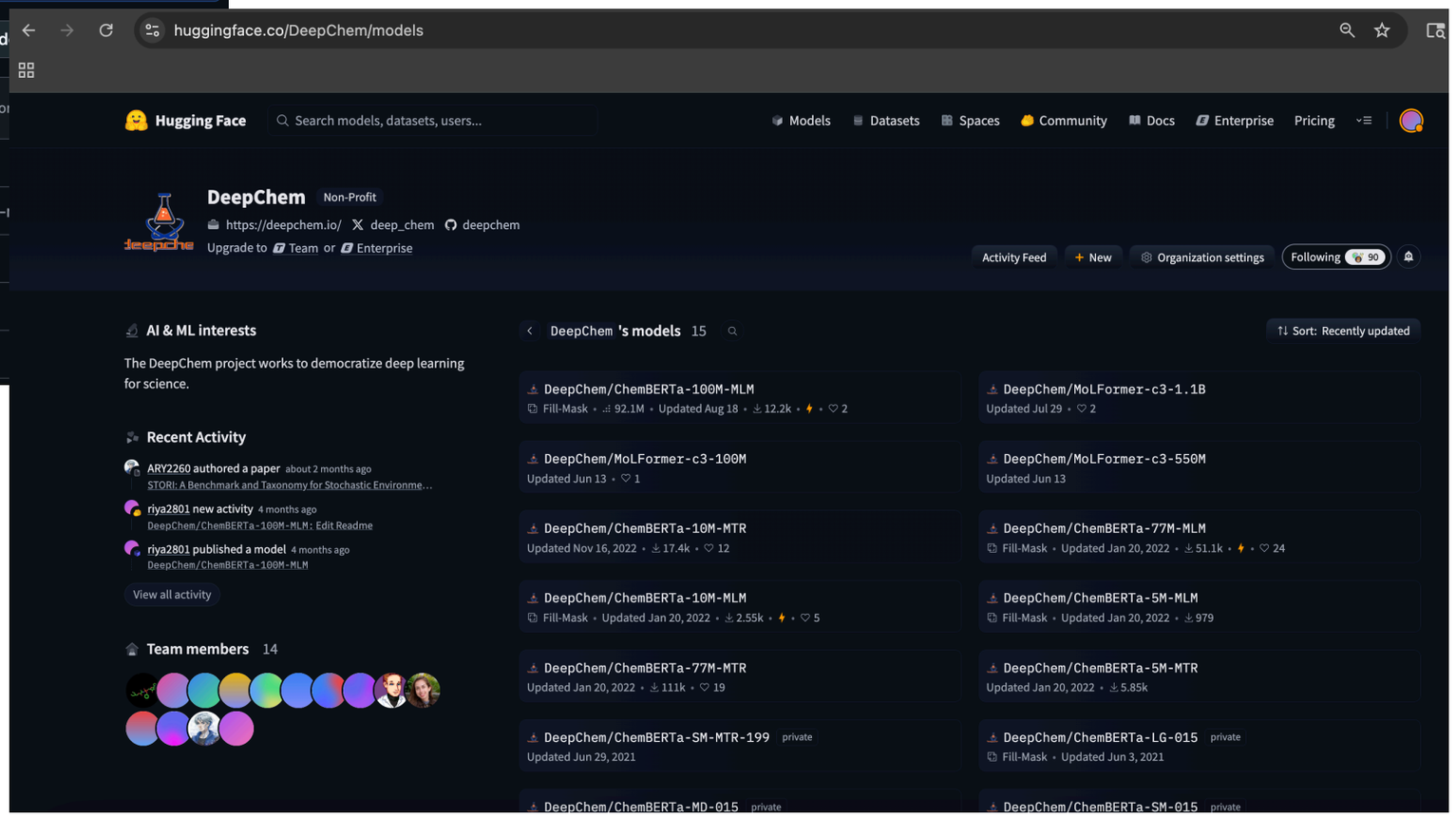
BUT, STARTING TO HIT DIMINISHING RETURNS WITH SCALE?

Table 4 This table compares the ChemBERTa and MoLFormer models pretrained on ZINC and PubChem datasets of varying sizes on various **classification datasets** and reports ROC AUC scores (Higher is better). We used **deepchem scaffold splits** and pretrained ChemBERTa models on the ZINC 10M and 100M dataset.

Dataset Tasks	BACE ↑ 1	BBBP↑ 1	TOX21↑ 12	HIV↑ 1	SIDER↑ 27	CLINTOX↑ 2
ChemBERTa-MLM-10M	0.773 ± 0.010	0.715 ± 0.006	0.713 ± 0.014	0.725 ± 0.017	0.616 ± 0.010	0.983 ± 0.010
ChemBERTa-MLM-100M	0.781 ± 0.019	0.700 ± 0.027	0.718 ± 0.011	0.747 ± 0.009	0.629 ± 0.023	0.979 ± 0.022
c3-MoLFormer-10M	0.776 ± 0.031	0.715 ± 0.021	0.718 ± 0.003	0.711 ± 0.014	0.618 ± 0.005	0.847 ± 0.024
c3-MoLFormer-100M	0.809 ± 0.019	0.730 ± 0.016	0.729 ± 0.005	0.747 ± 0.017	0.631 ± 0.009	0.854 ± 0.036
c3-MoLFormer-550M	0.812 ± 0.017	0.742 ± 0.020	0.726 ± 0.002	0.659 ± 0.140	0.594 ± 0.007	0.856 ± 0.020
c3-MoLFormer-1.1B	0.819 ± 0.018	0.735 ± 0.019	0.723 ± 0.012	0.762 ± 0.005	0.618 ± 0.005	0.839 ± 0.013
MoLFormer-LHPC	0.887 ± 0.004	0.908 ± 0.013	0.791 ± 0.014	0.750 ± 0.003	0.622 ± 0.007	0.993 ± 0.004

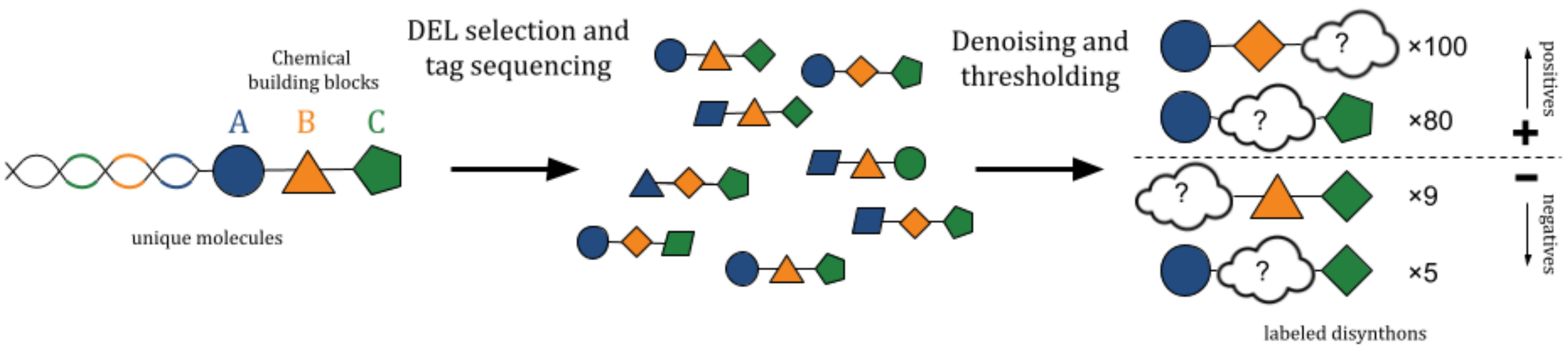
CHEMBERTA-3 CODE AND MODELS ARE FULLY OPEN SOURCED

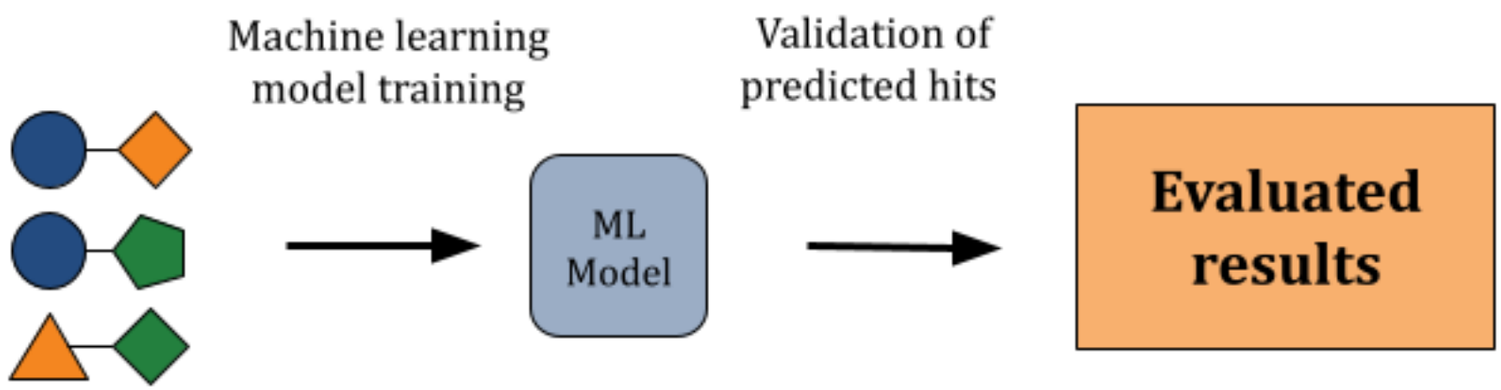




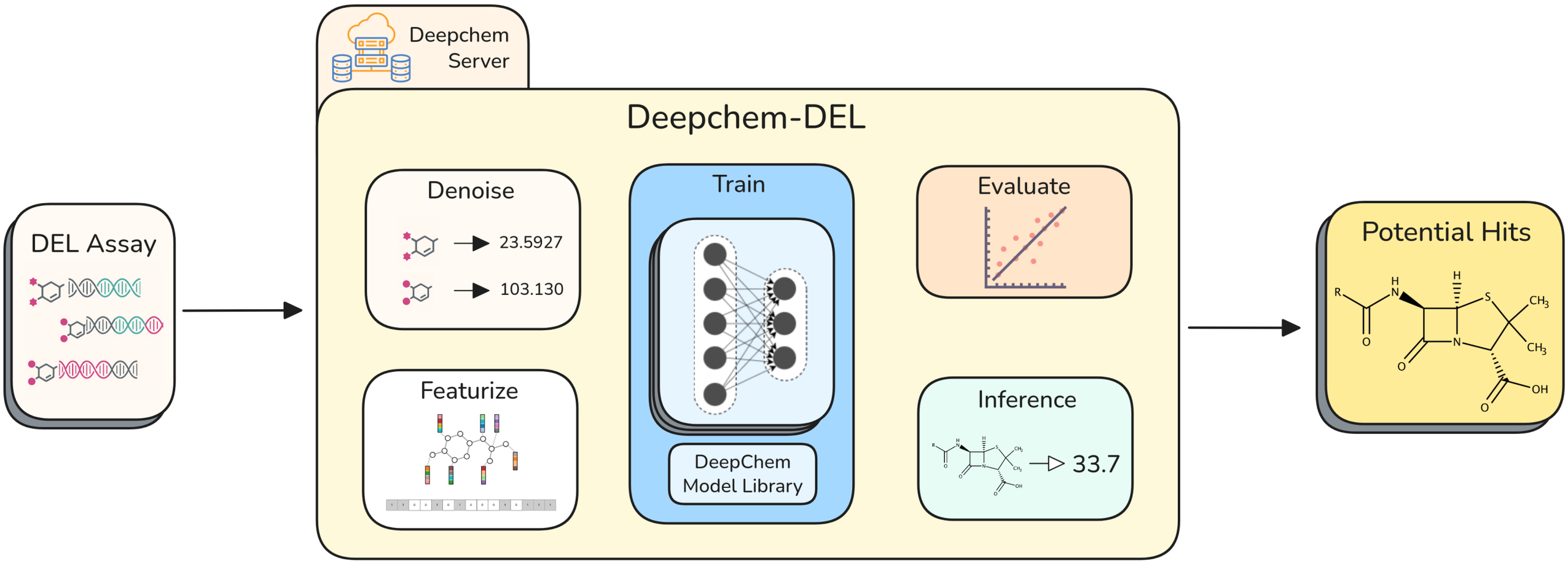
DEEPCHEM-DEL

MACHINE LEARNING FOR DNA ENCODED LIBRARIES (DEL) IS POWERFUL



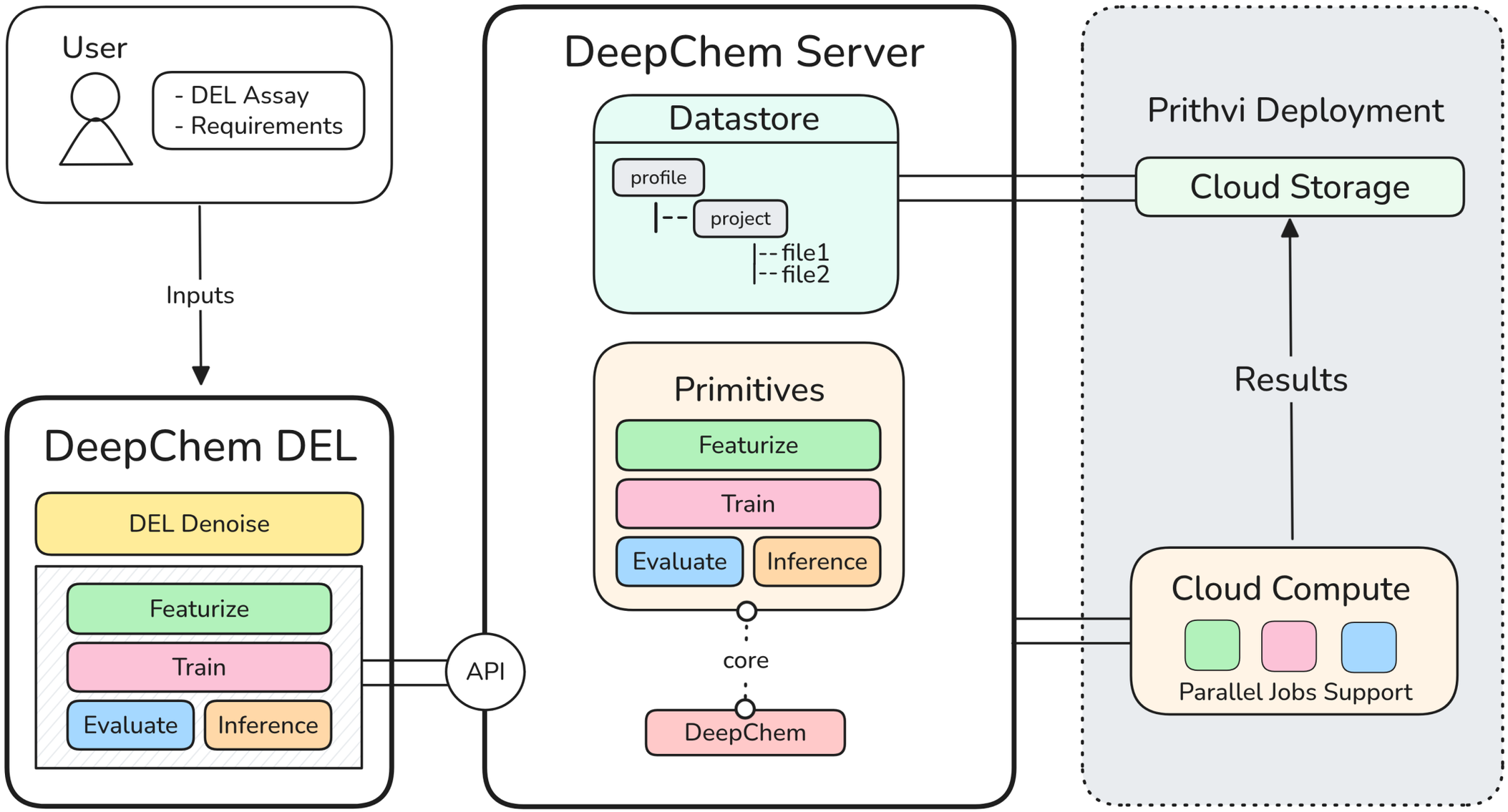


DEEPCHEM-DEL: A FRAMEWORK FOR BUILDING DEL MODELS

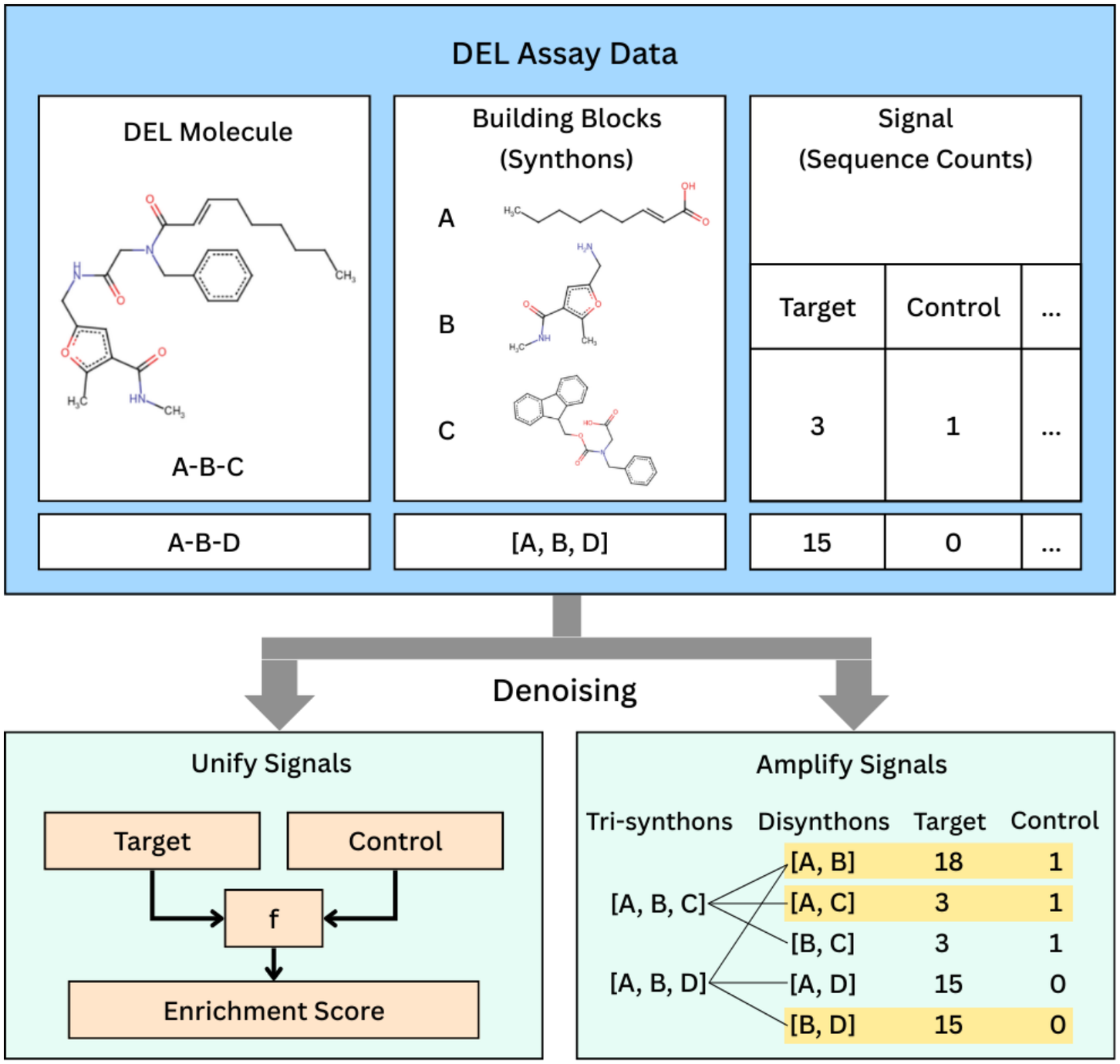


*Barsainyan, A. A., Singh, R., Mengade, A. P., Irfan, R., & **Ramsundar, B. (2025)**. DeepChem-DEL: An Open Source Framework for Reproducible DEL Modeling and Benchmarking.*

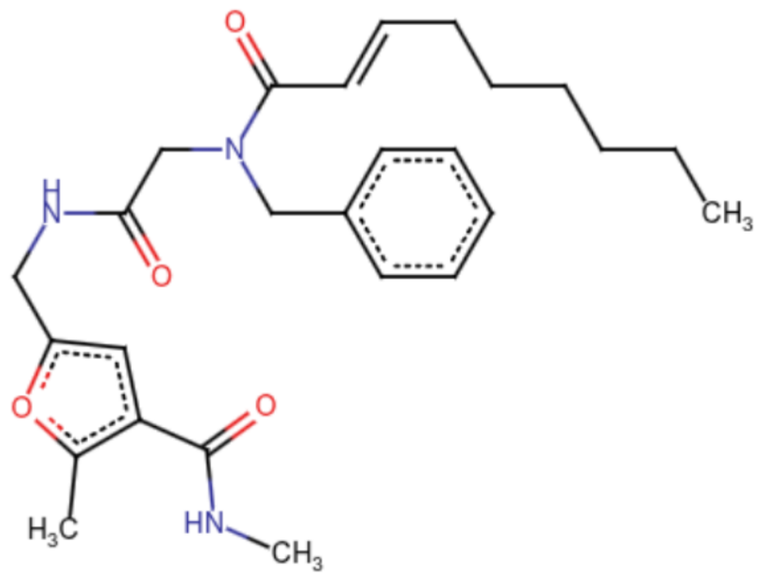
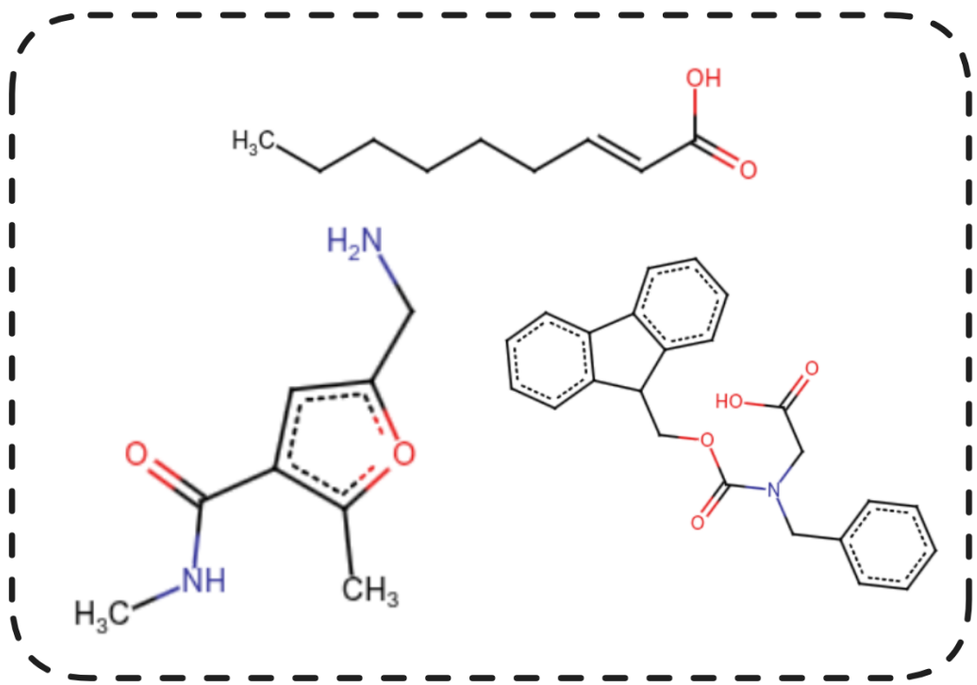
DEEPCHEM-SERVER FACILITATES LARGE SCALE CHEMOINFORMATICS



DEEPCHEM-DEL SUPPORTS MULTIPLE DENOISING STRATEGIES



CONNECTED SYNTHONS VERSUS DISJOINTED SYNTHONS

Connected Synthons Graph (DEL Molecule)	Disjointed Synthons Graph
	
CCCCCC/C=C/C(=O)N(CC(=O)NCc1cc(C(=O)NC)c(C)o1)Cc1ccccc1	CNC(=O)c1cc(CN)oc1C. CCCCCC\C=C\C(O)=O. OC(=O)CN(Cc1ccccc1)C(=O)OCC1c2ccccc2-c2ccccc12

DEEPCHEM-DEL CAN SIGNIFICANTLY IMPROVE PREDICTIVE POWER

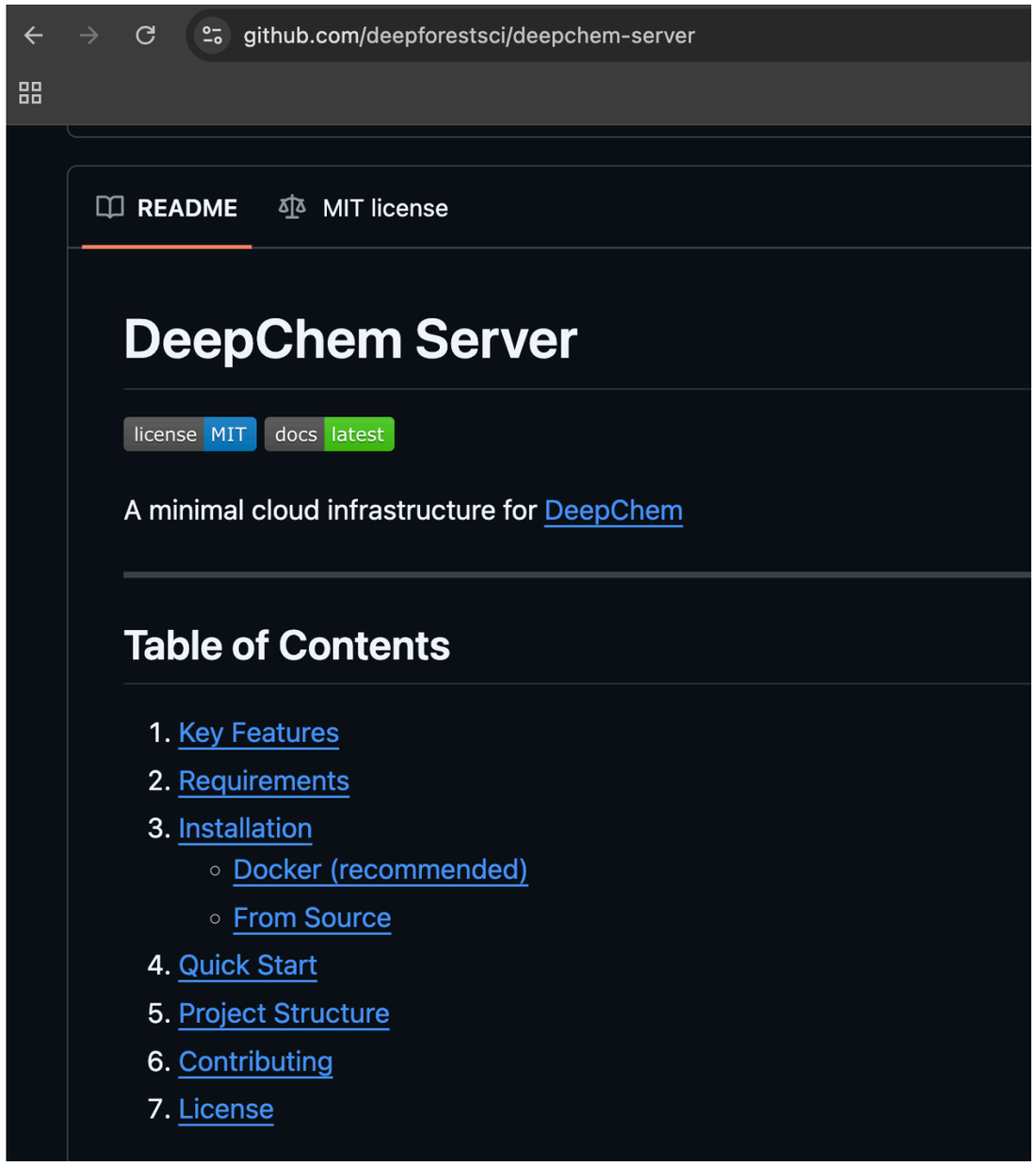
Table 1: Performance of unified regression models for amplified (di) and non-amplified (tri) on the DDR1 kinase target under random split evaluation. Models were trained jointly across targets to predict enrichment values. Results are reported as test RMSE (lower is better) and negative Spearman correlation for held-out in-library and held-out extended library sets.

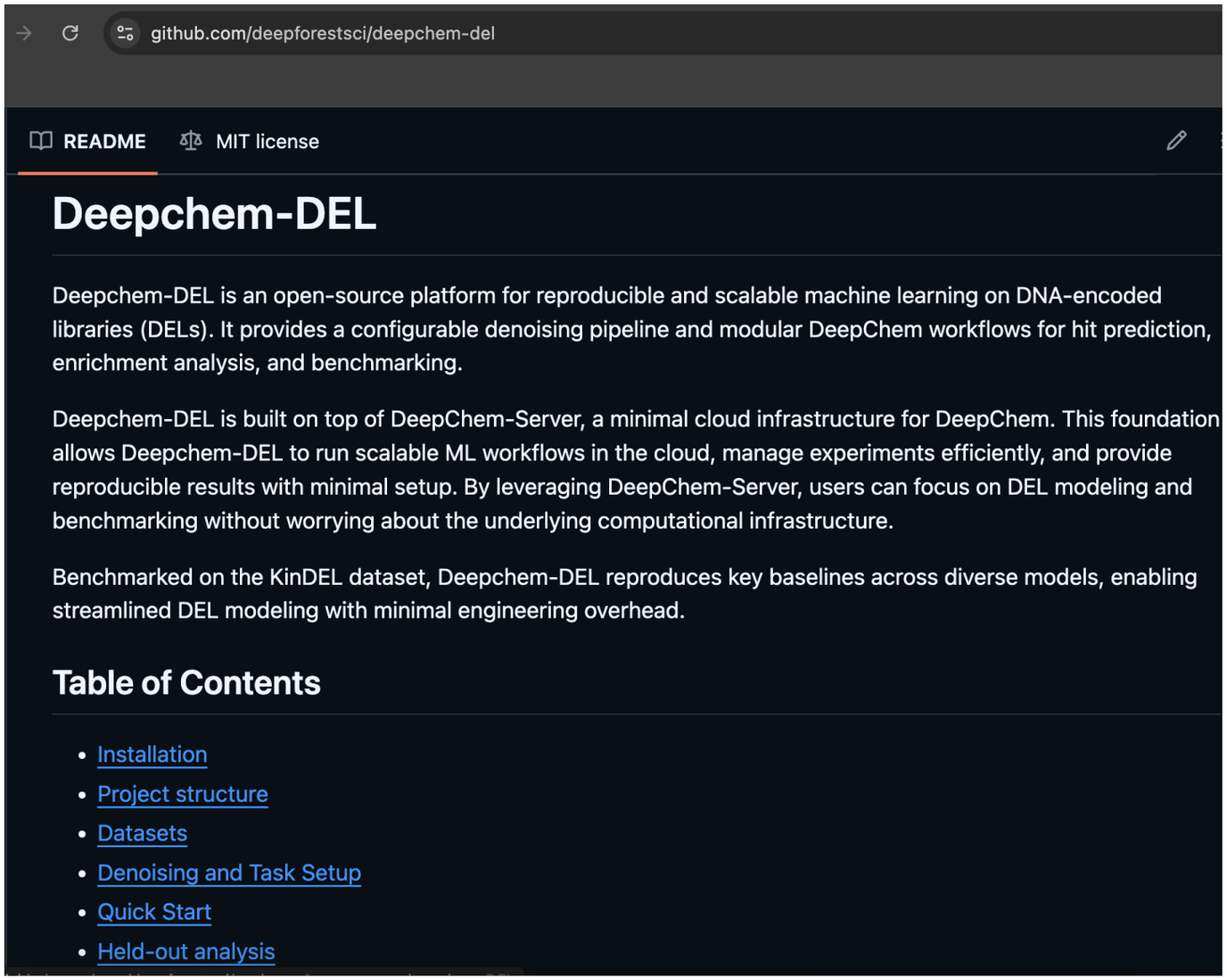
Input Type	Models	Test RMSE↓	Heldout In-Library		Heldout Extended	
			on-DNA↑	off-DNA↑	on-DNA↑	off-DNA↑
Di	RF	1.677 ± 0.001	-0.207 ± 0.011	-0.139 ± 0.007	-0.021 ± 0.021	-0.104 ± 0.012
	GCN	1.874 ± 0.234	0.361 ± 0.170	0.134 ± 0.041	0.289 ± 0.230	0.086 ± 0.026
Tri	RF	0.671 ± 0.037	0.714 ± 0.014	0.399 ± 0.017	0.665 ± 0.006	0.362 ± 0.017
	GCN	0.462 ± 0.212	0.465 ± 0.093	0.233 ± 0.055	0.489 ± 0.058	0.197 ± 0.057
	RF (KinDEL)	0.685 ± 0.011	0.578 ± 0.034	0.267 ± 0.022	0.608 ± 0.021	-
	GIN (KinDEL)	0.454 ± 0.012	0.572 ± 0.044	0.283 ± 0.028	0.579 ± 0.037	-

Table 3: The table summarizes the performance of the Unified Trisynthon Regression task on the DDR1 dataset, comparing results obtained under multiple DEL assay training set sizes, architecture choices, and hyperparameter selections.

Model-data configs	Heldout in-Library		Heldout Extended	
	on-DNA↑	off-DNA↑	on-DNA↑	off-DNA↑
GCN-128hf-1M-100e	0.465 ± 0.093	0.233 ± 0.055	0.489 ± 0.058	0.197 ± 0.057
DMPNN-1M-100e	**0.694**	**0.296**	**0.711**	**0.255**
GCN-256hf-1M-30e	0.481	0.217	0.537	0.198
GCN-256hf-1M-100e	0.552	0.258	0.544	0.216
GCN-128hf-5M-10e	0.137	-0.08	0.027	-0.119
GCN-256hf-5M-10e	0.167	0.049	-0.059	0.046
GCN-128hf-5M-30e	0.078	-0.048	0.069	-0.098
GCN-256hf-5M-30e	0.193	0.067	-0.007	0.067
GCN-256hf-5M-50e	0.373	0.188	0.402	0.197
GCN-128hf-11M-10e	0.073	-0.078	0.406	-0.073

DEEPCHEM-DEL IS OPEN-SOURCED



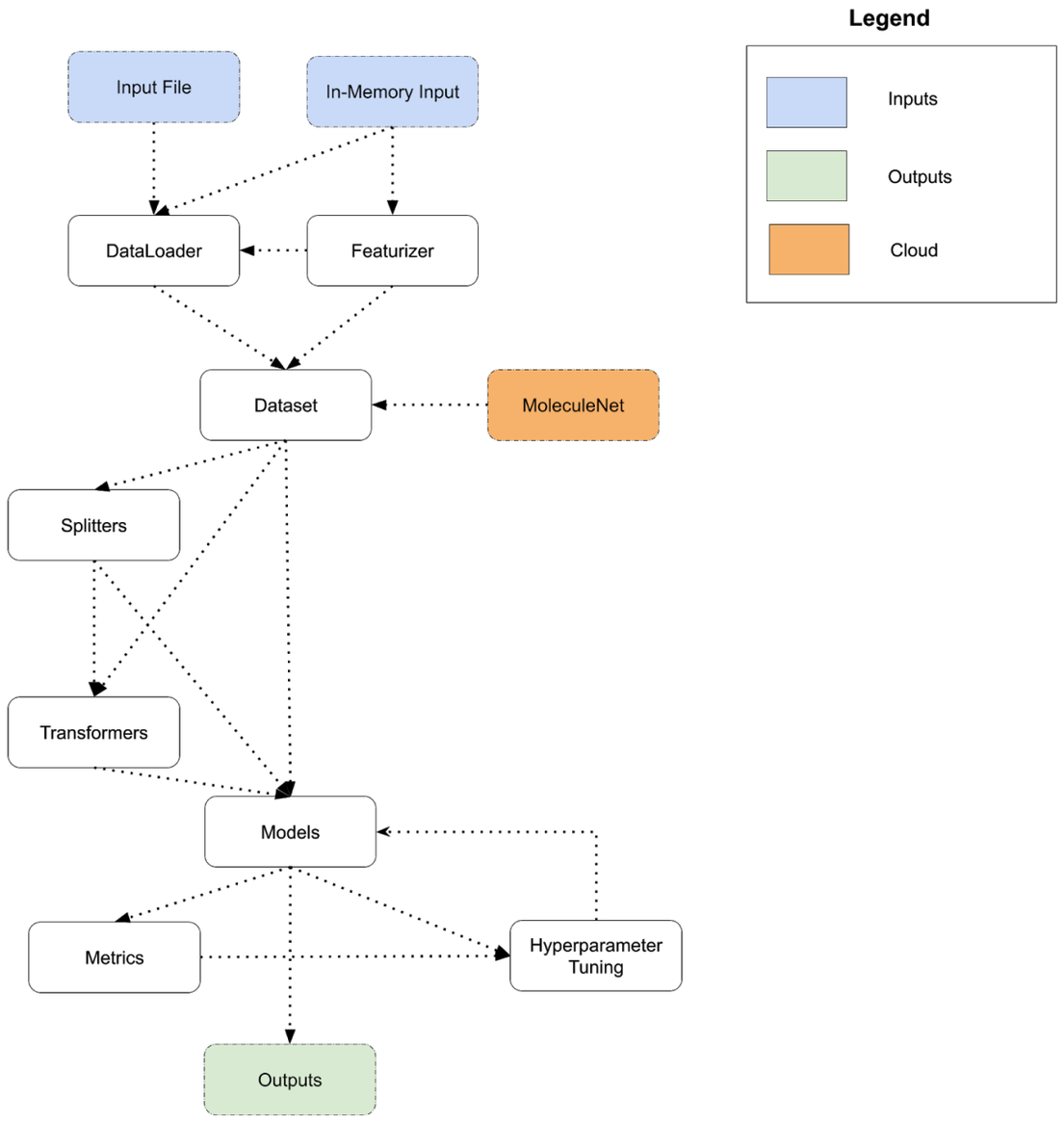


Chithrananda, Seyone, Gabriel Grand, and Bharath Ramsundar. "Chemberta: Large-scale self-supervised pretraining for molecular property prediction." arXiv preprint arXiv:2010.09885 (2020).

DEEPCHEM

SCIENTIFIC MACHINE LEARNING WITH DEEPCHEM

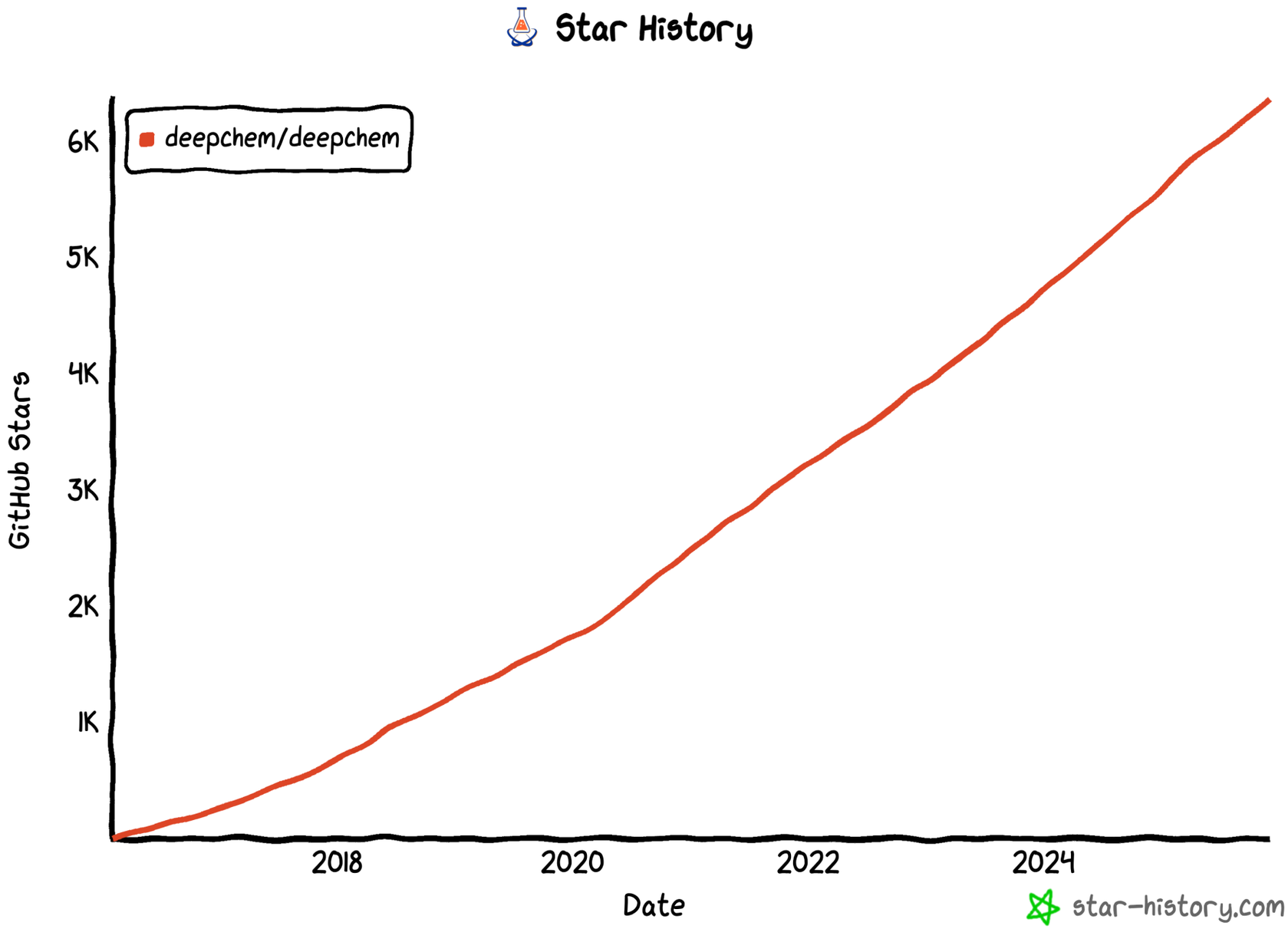
- DeepChem is a framework to apply AI to drug discovery and other scientific problems.
- DeepChem offers composable re-usable pipelines for scientific workflows
- DeepChem increasingly supports non-molecular applications for general scientific computing.



DEEPCHEM IS BROADLY USED

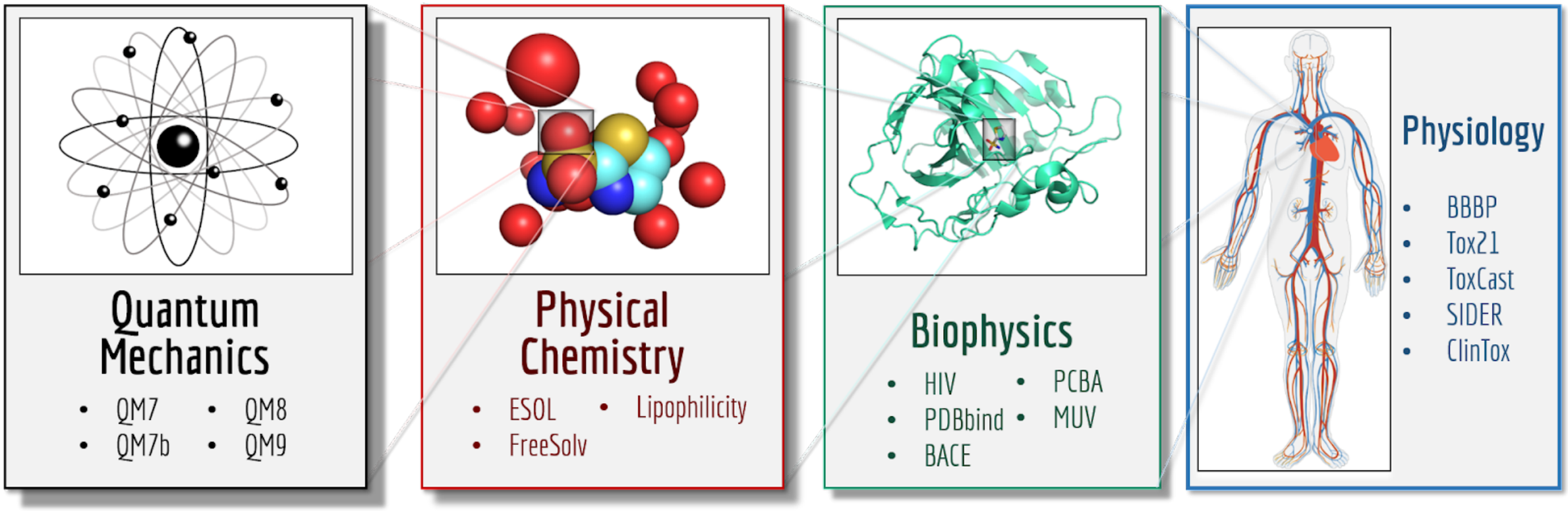
RIKEN
aws
Skoltech
Skolkovo Institute of Science and Technology
Carnegie Mellon University
WISCONSIN
UNIVERSITY OF WISCONSIN-MADISON
GHDDI
novo nordisk®
MERCK
IBM Watson
Expedia®
gsk
GlaxoSmithKline
Australian National University
SRI International
BROAD INSTITUTE
nurix
BAYER
Pfizer
Firmenich
Mizzou
University of Missouri
JKU
JOHANNES KEPLER UNIVERSITÄT LINZ
UNIVERSITAS AMOIENSIS
JIANGSU UNIVERSITY OF TECHNOLOGY
ReverieLabs
UCSF
University of California San Francisco
janssen
Weights & Biases
UNIVERSITÉ DE GENÈVE
SCHRÖDINGER.
Valo
OWKIN
GSI TECHNOLOGY
Open Analytics
UNIVERSITY OF VIENNA

DEEPCHEM HAS AN ACTIVE OPEN SOURCE ECOSYSTEM

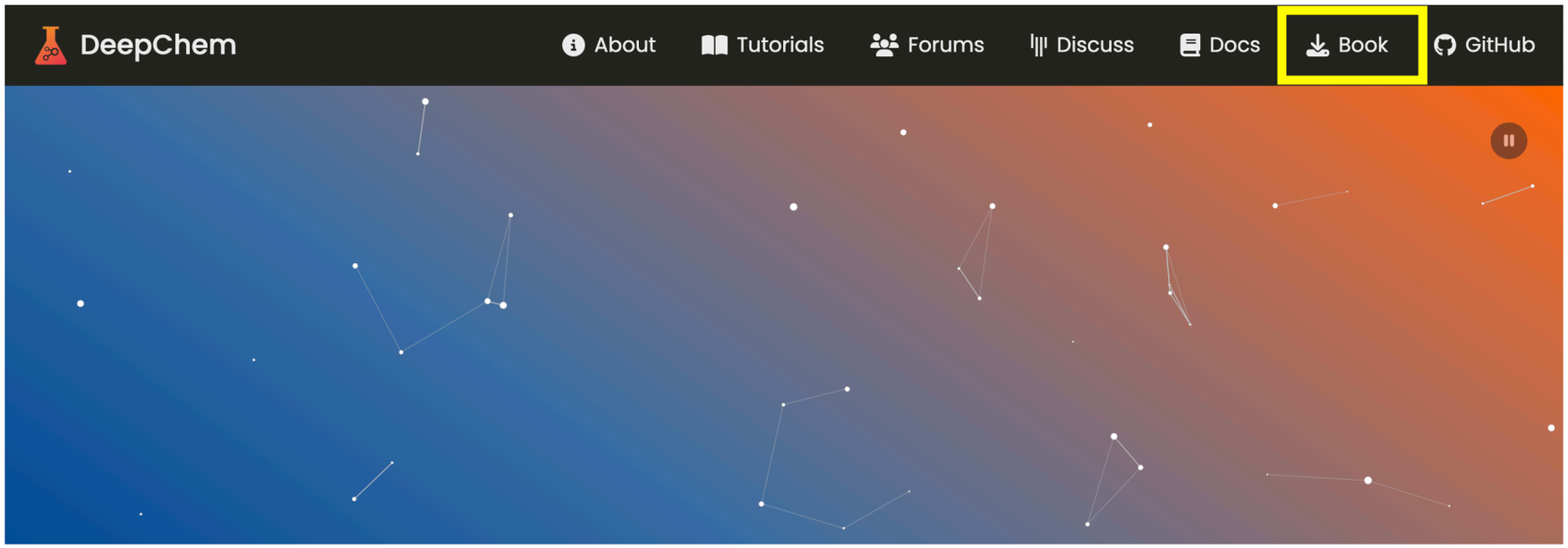


DeepChem	
Github Stars	6.4K
Downstream Packages	633
Contributors	188
Releases	20

MOLECULENET PROVIDES INTEGRATED DATASETS AND BENCHMARKS



JOIN THE COMMUNITY!

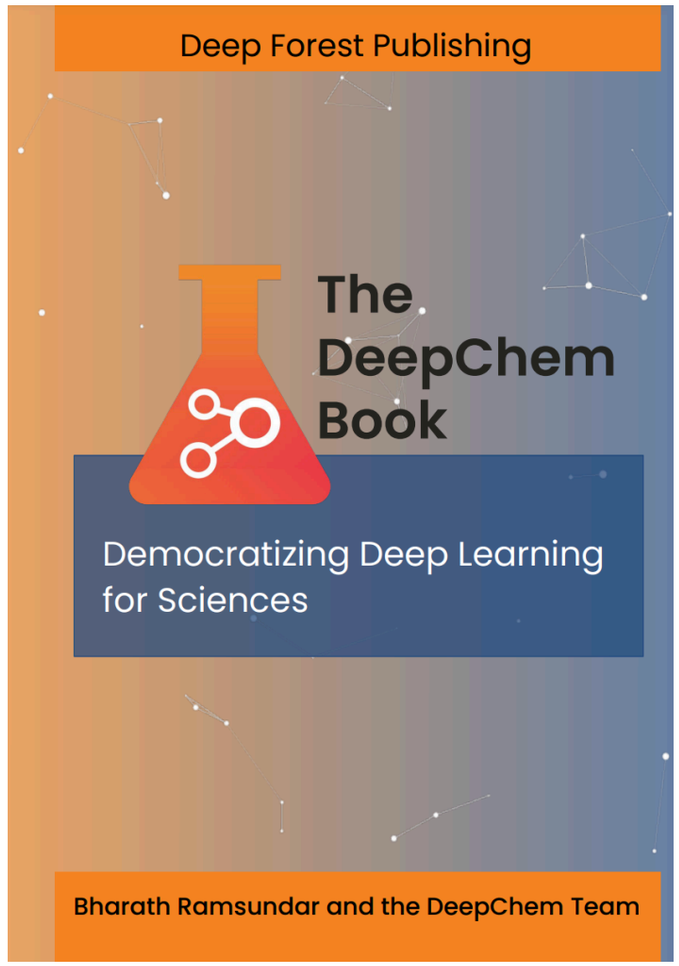


Democratising Deep Learning for
Material Science

DeepChem

"Chemistry itself knows altogether too well that - given the real fear that the scarcity of global resources and energy might threaten the unity of mankind - chemistry is in a position to make a contribution towards securing a true peace on earth."

~ Kenichi Fukui



Join the Discord!

DEEPCHEM-DEL

DEEPRETRO

CHEMBERTA-3

Our thanks to collaborators at Lawrence Livermore National Laboratory

CONTACT US

bharath@deepforestsci.com

@rbhar90