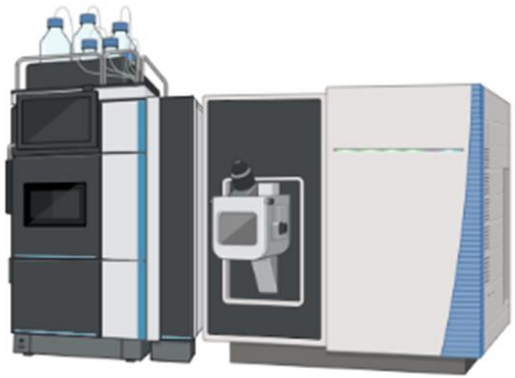
A Task-Specific Transfer Learning Approach to Enhancing Small Molecule Retention Time Prediction with Limited Data

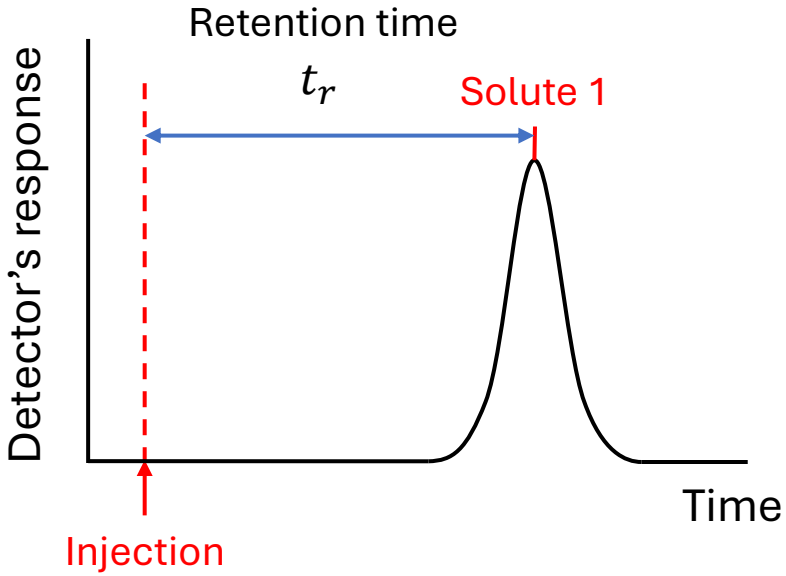
Yuhui Hong, Haixu Tang
Luddy School of Informatics, Computing, and Engineering
Indiana University, Bloomington, IN 47408, USA

The authors declare no competing financial interest.

Background | LC-MS retention time prediction

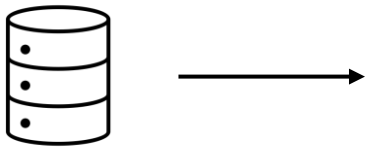


Liquid Chromatography-Mass Spectrometer



Background | LC-MS retention time prediction

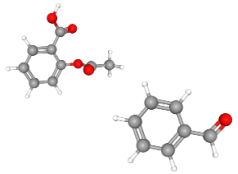
METLIN-SMRT
80,038 small molecules [1]



Small molecules datasets

Neural Networks [1, 2]

under the same standard condition

Molecule	Predicted retention time
	t'_{r1}
	t'_{r2}
.....	.....

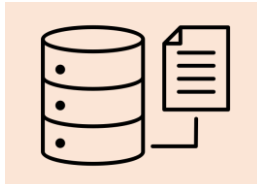
Instrumentation	
LC System	Agilent 1100/1200 series
Mass Spectrometer	Quadrupole-time of flight (Q-TOF) G6538A (Agilent Technologies)
Column	Zorbax Extend-C18 reverse-phase (2.1 × 50 mm, 1.8 µm, Agilent Technologies)

Chromatographic Conditions	
Flow rate	100 µL/min
Mobile phase A	Water + 0.1% formic acid
Mobile phase B	Acetonitrile + 0.1% formic acid
Dead volume	40 µL
Dwell volume	900 µL

[1] Domingo-Almenara X, et al. Nat. Commun. 10, 5811 (2019)
[2] Xue J, et al. Bioinformatics 40.3, btae084 (2024)

Challenge | limited data for different conditions

Retention time dataset (RT) for standard condition
(e.g. METLIN-SMRT [1])



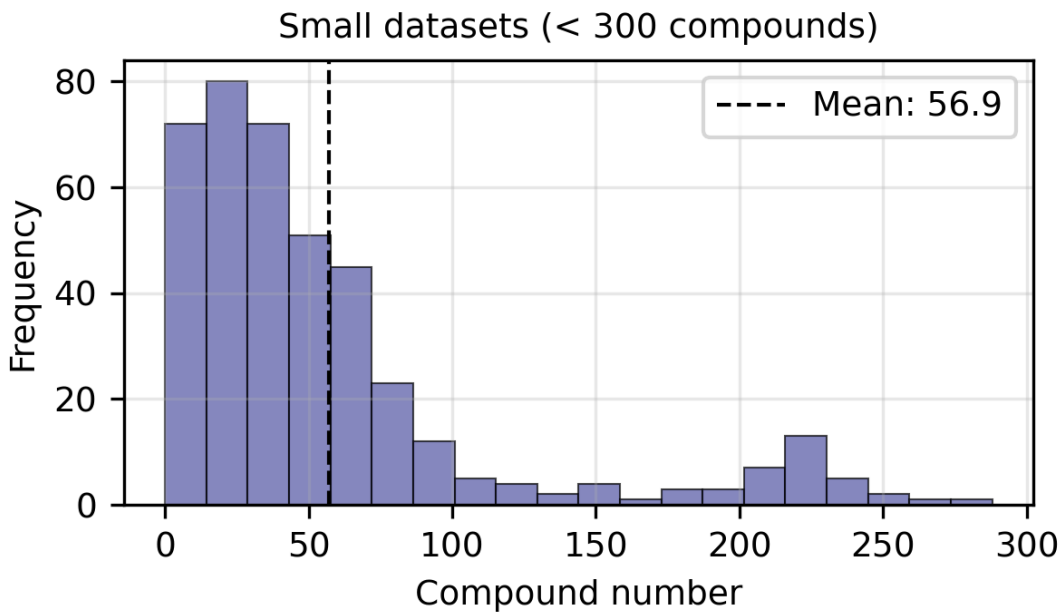
Retention time (RT) dataset for specialized conditions
(e.g. RepoRT [3])

Neural Networks

Insufficient data for training

.....

Only 10 datasets ≥ 300 compounds

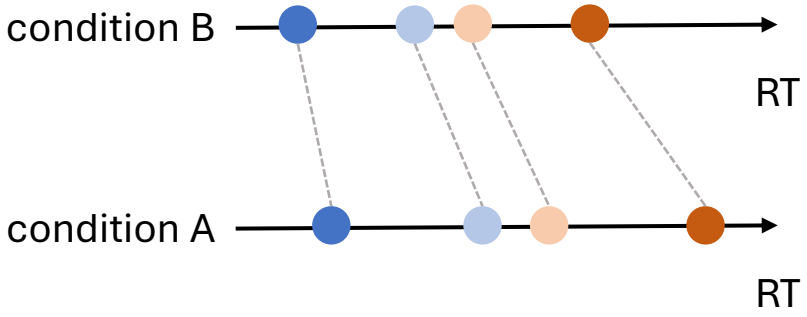


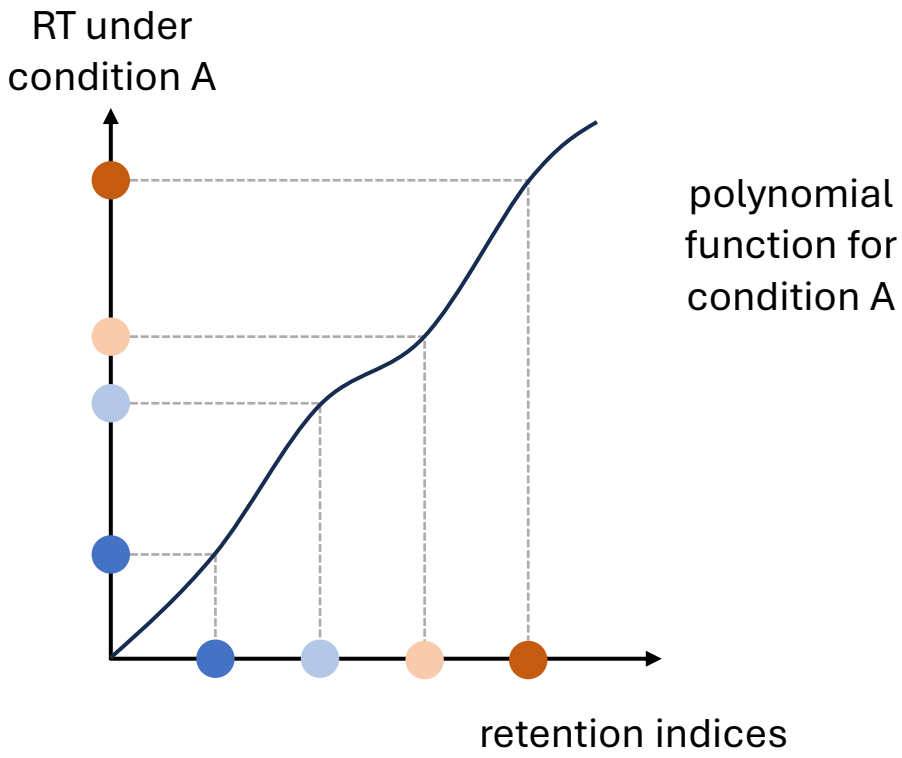
Compounds number (isomeric records) distribution of RT datasets in the RepoRT database [1] before preprocessing

[1] Domingo-Almenara X, et al. Nat. Commun. 10, 5811 (2019)
[3] Kretschmer F, et al. Nat. Methods 21.2, 153-155 (2024)

Challenge | limited data for different conditions

Potential solution 1:
Converting retention times to retention indices [4]

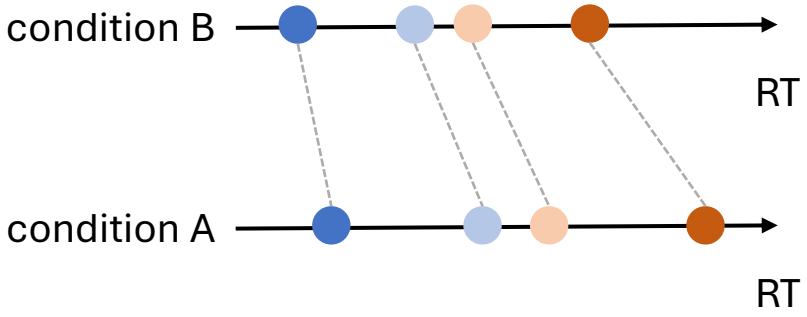


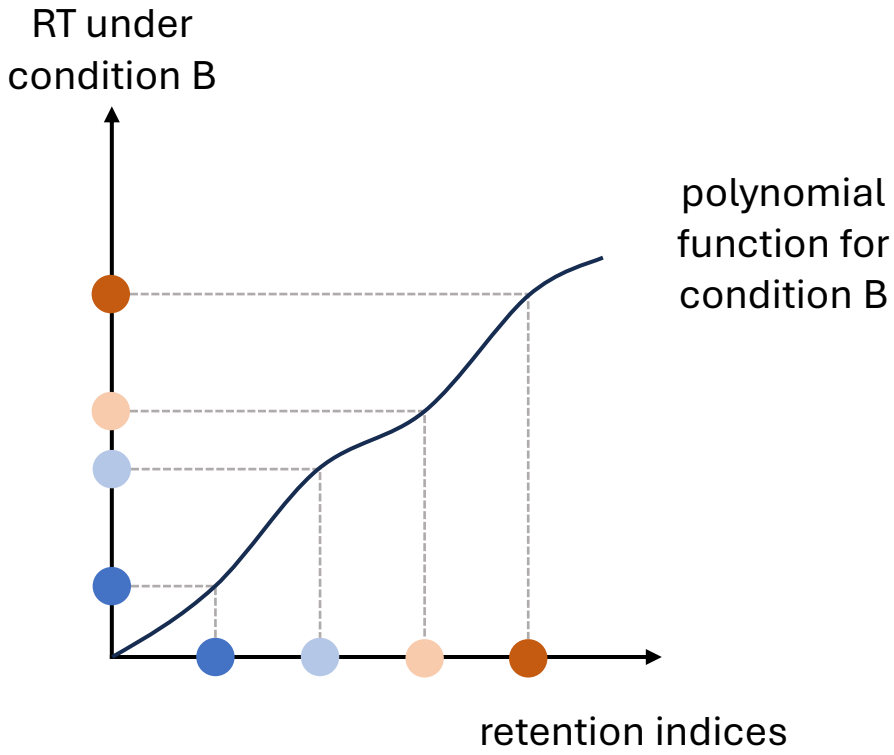


[4] Kretschmer F, et al. arXiv preprint (2024)

Challenge | limited data for different conditions

Potential solution 1:
Converting retention times to retention indices [4]

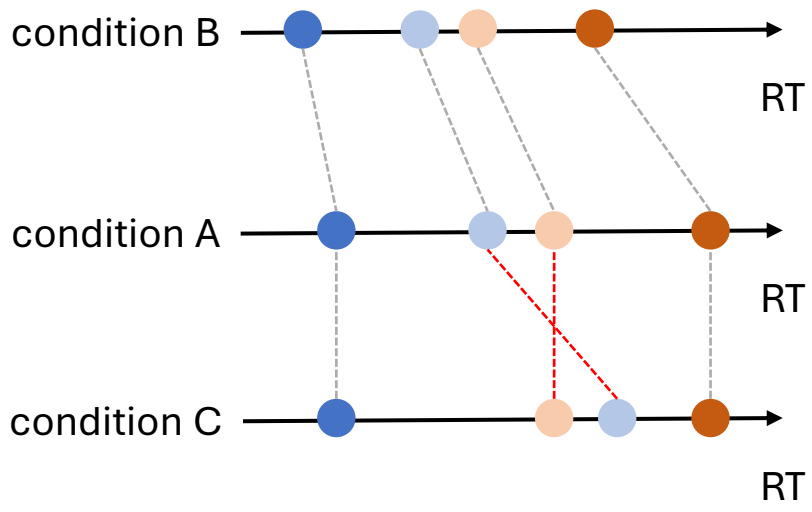


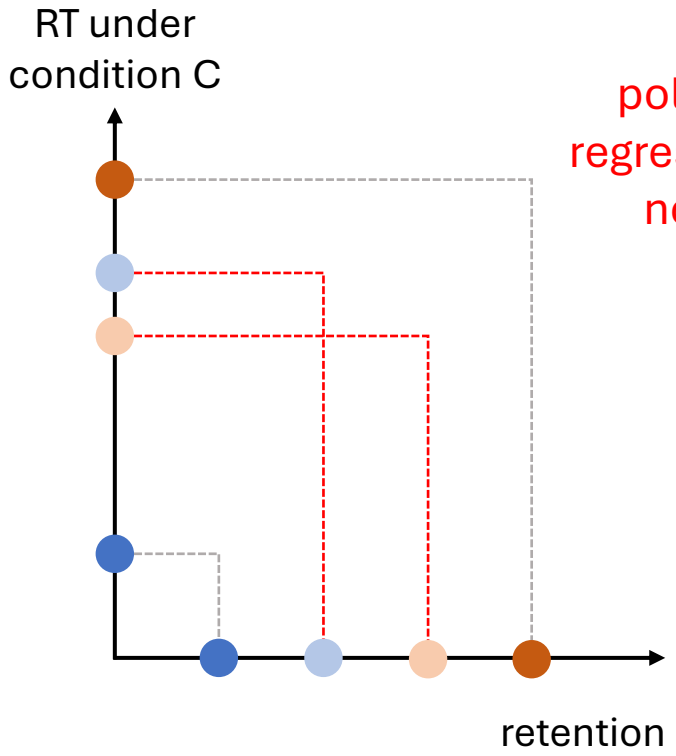


[4] Kretschmer F, et al. arXiv preprint (2024)

Challenge | limited data for different conditions

Potential solution 1: ✗
Converting retention times to retention indices





polynomial regression does not work

Challenge | limited data for different conditions

Potential solution 2: Transfer learning [5, 6]

Upstream dataset,
e.g. METLIN-SMRT [5],
LC-MSMS datasets [6], etc.

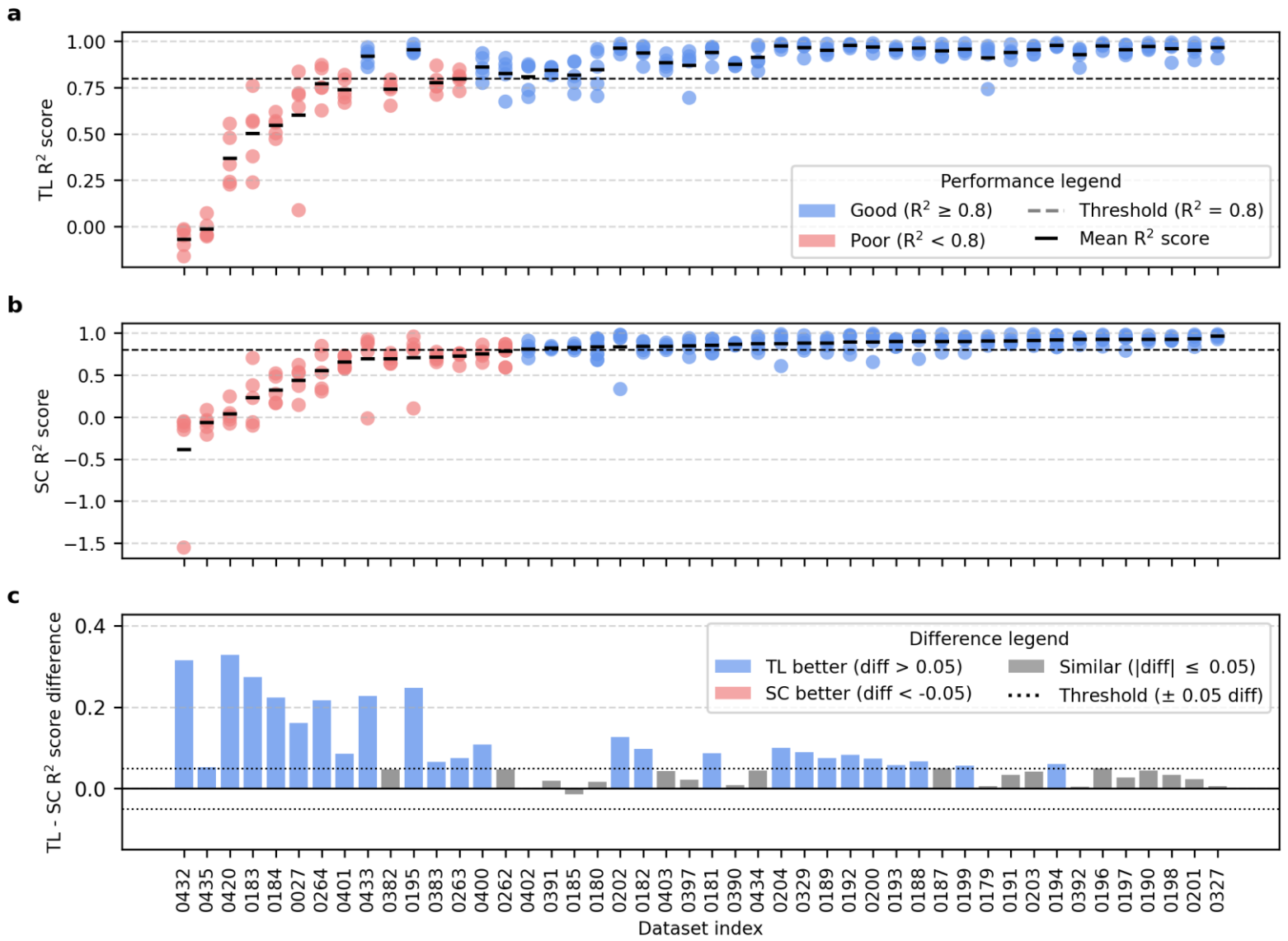
Downstream
dataset

[5] Kwon Y, et al. Anal. Chem. 95.47, 17273-17283 (2023)
[6] Hong Y, et al. Bioinformatics 39.6, btad354 (2023)

Challenge | limited data for different conditions

Potential solution 2: Transfer learning [5]

Performance of Graph Isomorphism Network (GIN) on 45 datasets from the RepoRT database using METLIN-SMRT pre-training [3]



[5] Kwon Y, et al. Anal. Chem. 95.47, 17273-17283 (2023)

Challenge | limited data for different conditions

Potential solution 2: Transfer learning ✗

- Large domain gap between datasets
- Limited downstream data

Upstream dataset, e.g. METLIN-SMRT

Downstream dataset

Our solution:

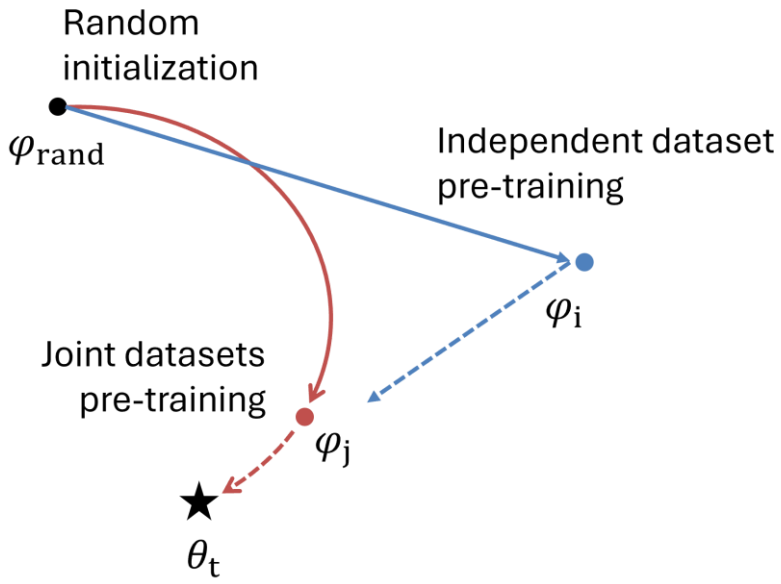
Task-Specific Transfer Learning (TSTL)

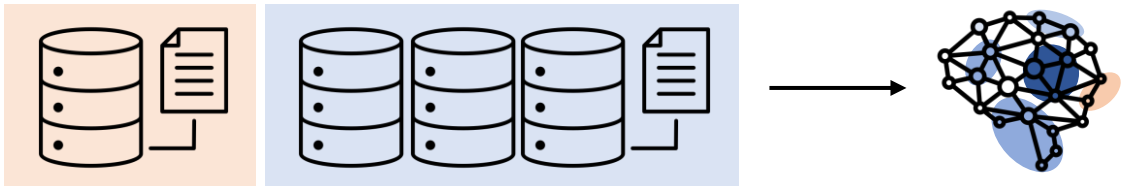
Step 1: Joint datasets pre-training

Step 2: Integration of multiple fine-tuned models

Methodology | task-specific transfer learning

Step 1: Joint datasets pre-training





Upstream dataset
+ Downstream dataset

Efficient fine-tuning

Downstream
dataset

Methodology | task-specific transfer learning

Optimizer methods

$\mathcal{L}(\emptyset)$

pre-trained model

best initialization for
specific downstream task

Methodology | task-specific transfer learning

Optimizer methods

$$\mathcal{L}(\emptyset) = l_t(\theta_t)$$

pre-trained model

best initialization for specific downstream task

fine-tuned model

Methodology | task-specific transfer learning

Optimizer methods

$$\mathcal{L}(\emptyset) = l_t(\theta_t) = l_t(\emptyset - \alpha \nabla_{\emptyset} l_t(\emptyset))$$

inner loop: update θ_t initialized as $\emptyset$ on downstream task

learning rate: α, β

Methodology | task-specific transfer learning

Optimizer methods

$$\emptyset \leftarrow \emptyset - \beta \nabla_{\emptyset} l_{\text{up}}(\emptyset)$$

outer-loop: update randomly initialized $\emptyset$ according to upstream task

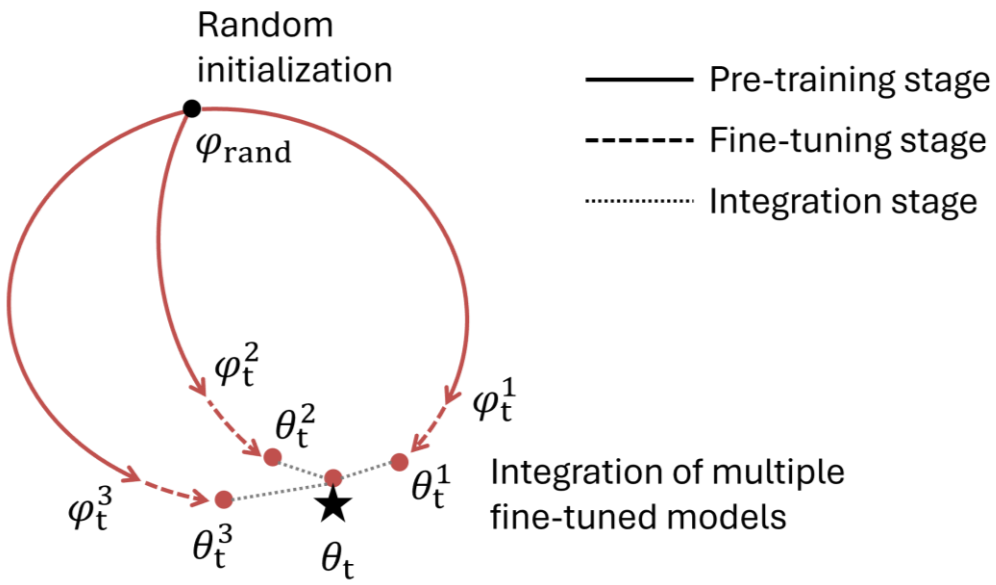
$$\mathcal{L}(\emptyset) = l_t(\theta_t) = l_t(\emptyset - \alpha \nabla_{\emptyset} l_t(\emptyset))$$

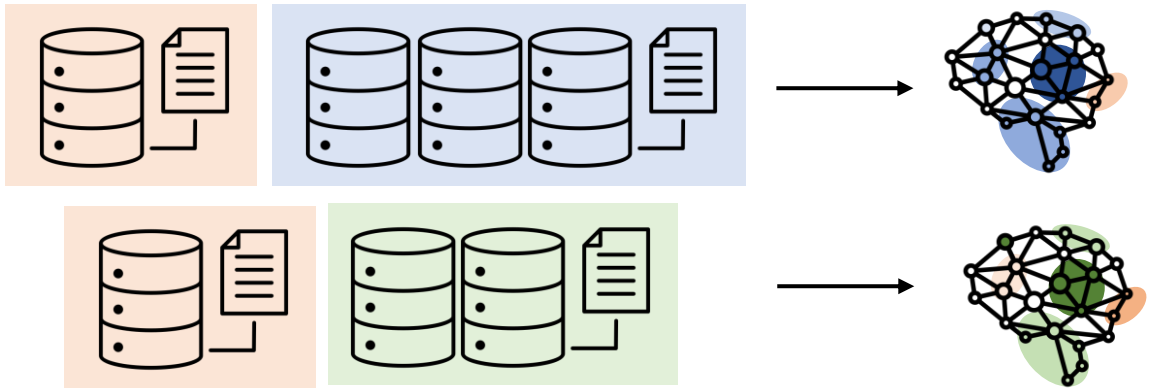
inner loop: update θ_t initialized as $\emptyset$ on downstream task

learning rate: α, β

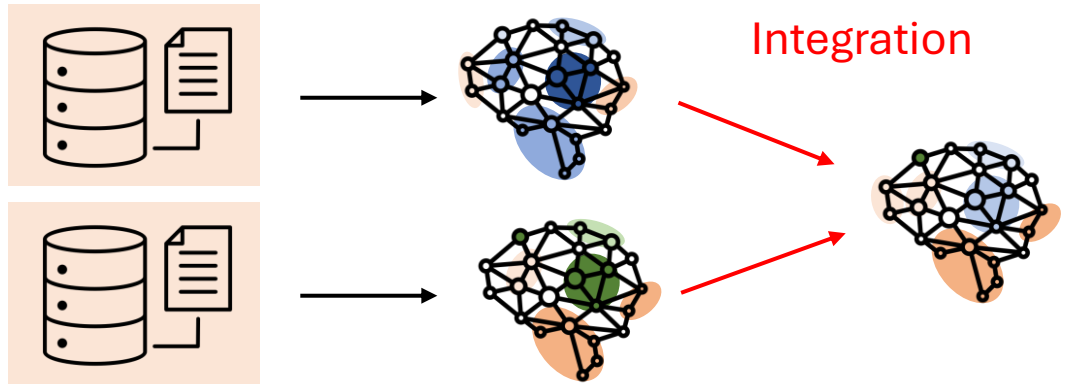
Methodology | task-specific transfer learning

Step 2: Integration of multiple fine-tuned models



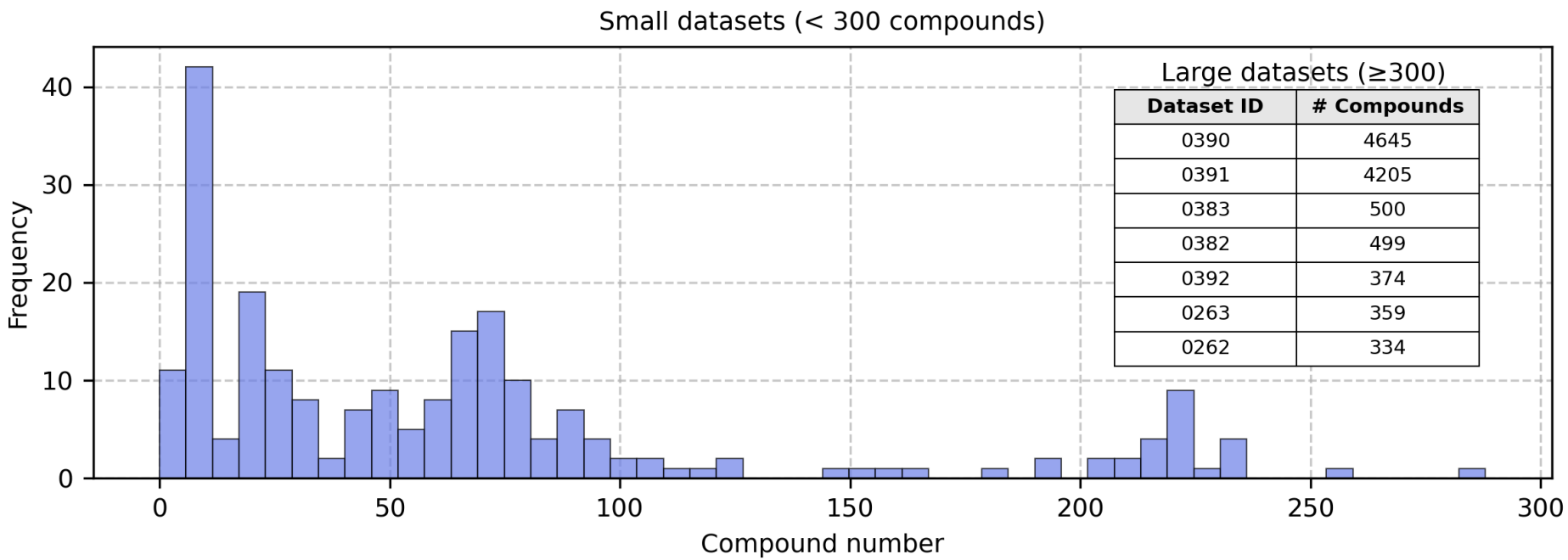


Upstream dataset
+ Downstream dataset



Downstream dataset

Results | data preprocessing



Large datasets (≥300)

Dataset ID	# Compounds
0390	4645
0391	4205
0383	500
0382	499
0392	374
0263	359
0262	334

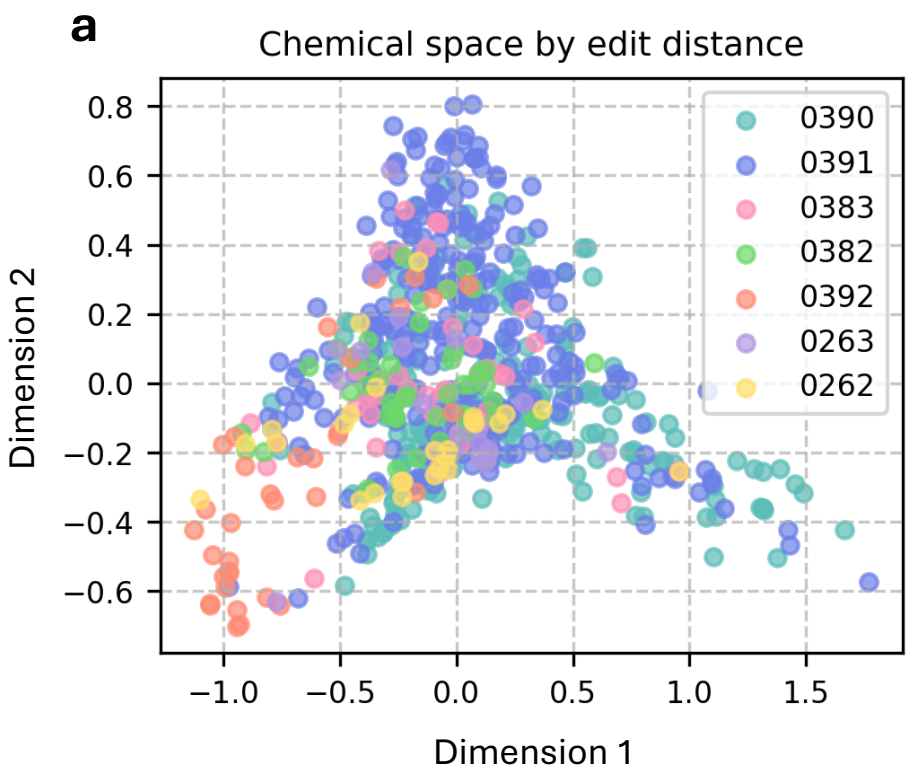
TL-difficult datasets
(TL $R_2 < 0.8$)

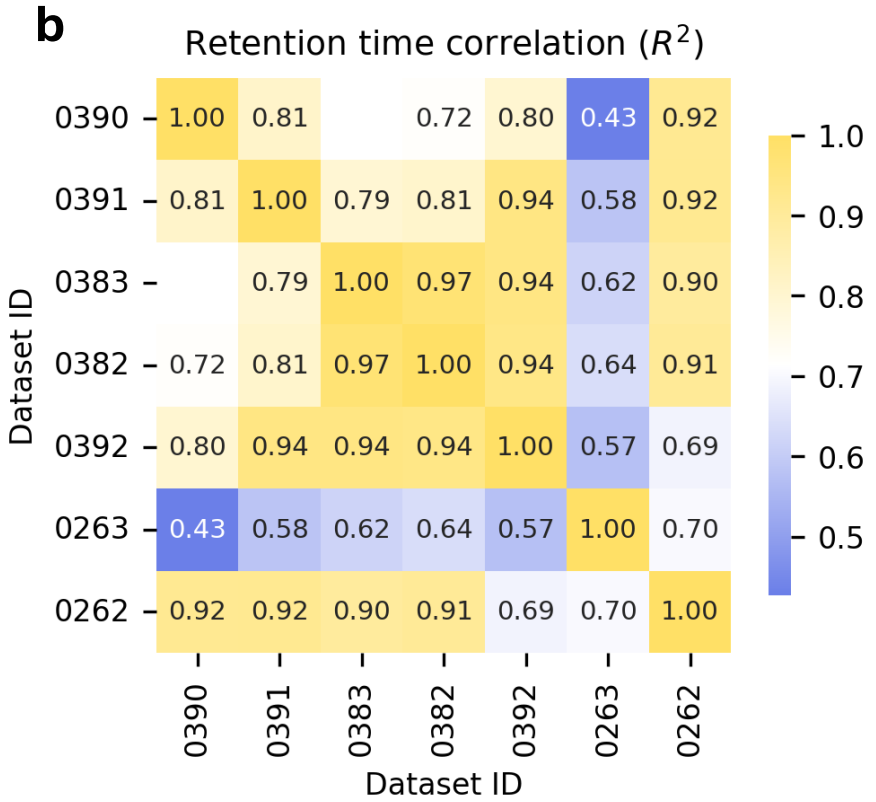
Dataset ID	# Compounds
0027	104
0183	121
0184	124
0264	96
0401	235
0420	162
0432	102
0435	191

Compounds number distribution of RepoRT database [3] after preprocessing

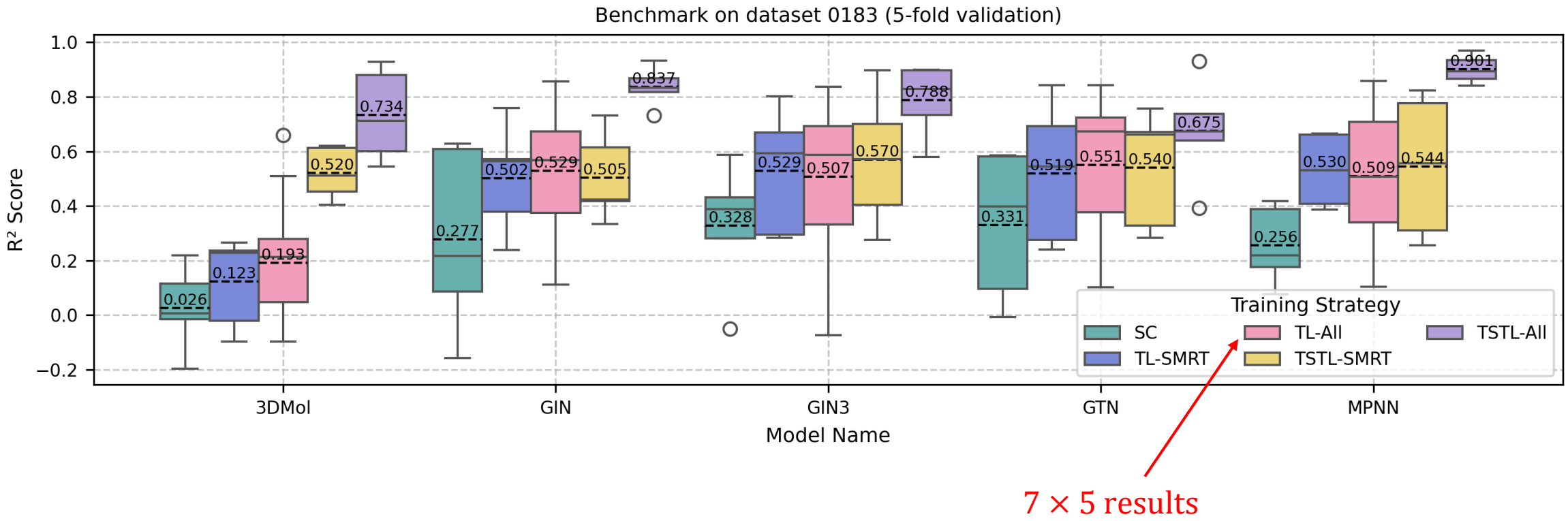
[3] Kretschmer F, et al. Nat. Methods 21.2, 153-155 (2024)

Results | diverse and correlation of up stream tasks

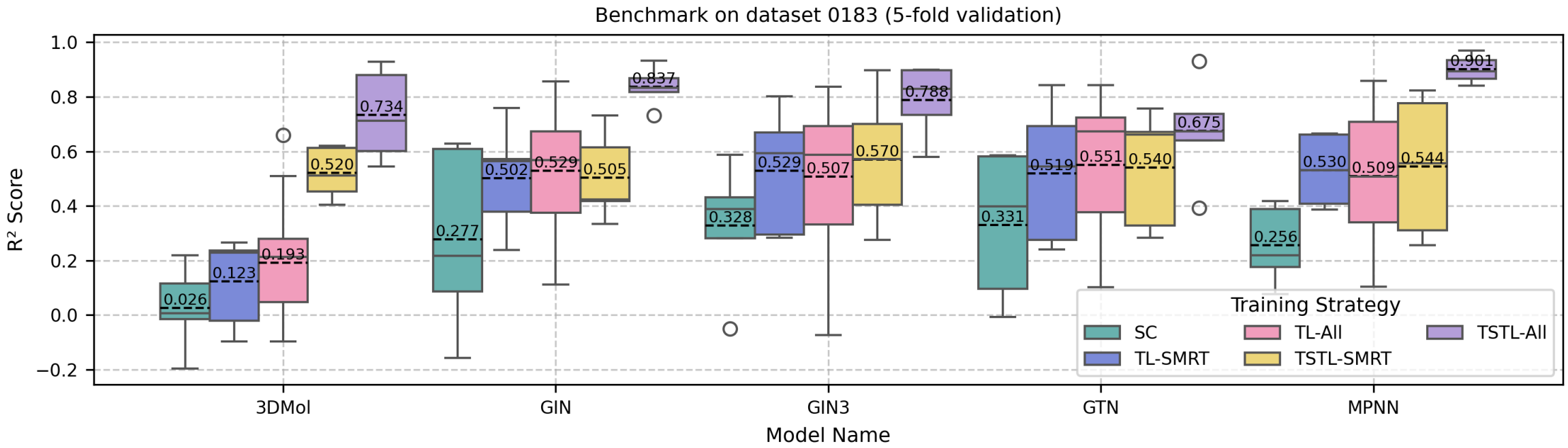




Results | performance of TSTL on retention time prediction



Results | performance of TSTL on retention time prediction



SC ⟹ TL-SMRT ⟶ TL-ALL ⟶ TSTL-SMRT ⟹ TSTL-ALL

Takeaways

- We designed Task-Specific Transfer Learning (TSTL) for training neural networks with limited data
- TSTL incorporates 2 steps: joint pre-training and a greedy integration strategy
- Experiments demonstrate that TSTL outperforms TL on all TL-difficult datasets in RepoRT database

Future work

- Apply TSTL methodology to more predictions
- Enhance integration considering experimental condition correlations

Acknowledgement

We acknowledge the Center for Bioanalytical Metrology (CBM), an NSF Industry-University Cooperative Research Center, for providing funding under grant NSF IIP-1916645. This work was also partially supported by National Science Foundation grant DBI-2011271.

Visit other works from our lab!

TP 057: AI-driven visual proteomics for blood-based Alzheimer's disease biomarker discovery by LC-MS/MS and deep neural networks. *Qingyang Xiao et al.*

WP 451: Metabolite identification by spectral searching against predicted spectral library. *Chhavi Thakur et al.*

Thank you!

Manuscript is under preparing.
Follow this GitHub repository for any updates!

CENTER for BIOANALYTIC METROLOGY
Indiana | Purdue | Notre Dame

References

[1] Domingo-Almenara, Xavier, et al. "The METLIN small molecule dataset for machine learning-based retention time prediction." Nature communications 10.1 (2019): 5811.

[2] Xue, Jun, et al. "RT-Transformer: retention time prediction for metabolite annotation to assist in metabolite identification." Bioinformatics 40.3 (2024): btae084.

[3] Kretschmer, Fleming, et al. "RepoRT: a comprehensive repository for small molecule retention times." Nature Methods 21.2 (2024): 153-155.

[4] Kretschmer, Fleming, et al. "Times are changing but order matters: Transferable prediction of small molecule liquid chromatography retention times." (2024).

[5] Kwon, Youngchun, et al. "Retention time prediction through learning from a small training data set with a pretrained graph neural network." Analytical Chemistry 95.47 (2023): 17273-17283.

[6] Hong, Yuhui, et al. "3DMolMS: prediction of tandem mass spectra from 3D molecular conformations." Bioinformatics 39.6 (2023): btad354.