Journal Name

ARTICLE TYPE

Cite this: DOI: 00.0000/xxxxxxxxxx

Supplementary Materials: 3D chemical structures allows robust deep learning models for retention time prediction

Mark Zaretckii,[a] Inga Bashkirova,[a] Sergey Osipenko,[a] Yury Kostyukevich,[a] Evgeny Nikolaev,[a] and Petr Popov*[a]

Received Date
Accepted Date

DOI: 00.0000/xxxxxxxxxx

[a] *iMolecule, Skolkovo Institute of Science and Technology, Moscow, Russia, 121205*
* *corresponding author, E-mail: p.popov@skoltech.ru*

Table 1 The MAE and MedAE performance metrics on the METLIN test sets for random and scaffold splits. Metrics are averaged using 4-x fold cross-validation.

Model	MAE(Random Split)	MedAE(Random Split)	MAE(Scaffold Split)	MedAE(Scaffold Split)
DNN	$54 \pm 1 (6.7 \pm 0.1\%)$	$37 \pm 1 (4.7 \pm 0.1\%)$	$58 \pm 1 (7.2 \pm 0.1\%)$	$41 \pm 1 (5.2 \pm 0.1\%)$
GNN	$\mathbf{39 \pm 1 (4.8 \pm 0.1\%)}$	$\mathbf{24 \pm 1 (3.0 \pm 0.1\%)}$	$\mathbf{41 \pm 1 (5.0 \pm 0.1\%)}$	$\mathbf{25 \pm 1 (3.2 \pm 0.1\%)}$
CPORT	$44 \pm 2 (5.5 \pm 0.4\%)$	$26 \pm 1 (3.4 \pm 0.1\%)$	$47 \pm 4 (5.9 \pm 0.7\%)$	$31 \pm 3 (4.0 \pm 0.5\%)$

Table 2 Grid search for optimal parameters of the CPORT model.

Conformation	Orientation	Numb. of confs.	r_{thr}(in seconds)	test MAE (seconds)
rdkit	rotated	2	15	47.82
rdkit	oriented	1	0	48.56
rdkit	oriented	2	0	48.64
rdkit	oriented	1	45	48.7
rdkit	rotated	1	15	48.87
rdkit	oriented	2	0	49.0
rdkit	oriented	1	0	49.1
rdkit	oriented	1	15	49.32
rdkit	rotated	2	0	49.34
rdkit	rotated	2	0	49.41
rdkit	rotated	2	45	49.59
MD	rotated	1	0	49.6
MD	rotated	1	15	49.66
rdkit	rotated	1	0	49.66
rdkit	rotated	1	45	49.72
rdkit	oriented	2	15	49.96
rdkit	oriented	4	0	51.03
MD	oriented	1	15	51.26
MD	oriented	1	45	51.36
MD	rotated	2	15	51.69
rdkit	oriented	2	45	51.93
rdkit	rotated	4	0	52.07
MD	rotated	1	45	52.09
MD	rotated	1	0	52.1
MD	rotated	2	0	52.2
rdkit	oriented	4	0	52.45
MD	rotated	2	0	52.48
MD	oriented	2	15	52.6
rdkit	rotated	4	15	52.63
MD	oriented	2	0	53.06
MD	rotated	4	15	53.31
MD	oriented	1	0	53.48
rdkit	rotated	4	0	53.48
MD	rotated	4	45	53.61
rdkit	oriented	4	45	54.4
MD	rotated	2	45	54.44
MD	rotated	4	0	54.45
rdkit	oriented	4	15	54.48
rdkit	rotated	4	45	55.12
MD	oriented	2	0	55.74
MD	rotated	4	0	55.75
MD	oriented	4	0	55.77
MD	oriented	2	45	55.8
MD	oriented	4	0	56.41
MD	oriented	4	15	56.56
MD	oriented	1	0	58.36
MD	oriented	4	45	59.11
rdkit	rotated	1	0	59.47

Table 3 Metrics after fine-tuning.

metrics	model	ABC	AjsTestF	AjsUoB	Bade_Publi	BDD_C18	BfG_NTS_RP1	Cao_HILIC	HILIC_tip	RPMMFDA	SNU_RIKEN_POS	Waters ACQUITY UPLC with Synapt G1 Q-TOF	in-house_c8	in-house_c18
MAE	gnn	53.53	69.31	47.04	90.12	53.83	75.74	67.96	66.36	58.89	43.25	143.78	136.09	144.17
	dnn	92.54	91.83	55.92	97.02	81.86	100.12	87.95	85.52	54.77	54.6	131.19	138.7	152.98
	cport	**48.87**	**59.17**	**40.72**	**64.88**	50.11	78.13	**64.84**	**55.47**	**36.55**	41.58	**102.14**	**92.52**	**102.88**
	XGBoost	51.98	69.7	41.86	79.57	**47.6**	**75.59**	74.09	64.48	41.35	**38.63**	103.26	119.16	138.22
MAPE	gnn	25.59	48.1	61.61	23.3	48.91	17.45	64.89	44.65	31.9	16.64	66.13	39.44	43.22
	dnn	48.88	68.72	90.07	25.34	94.82	22.32	88.36	67.3	32.35	22.89	52.2	53.89	58.19
	cport	**21.99**	**32.57**	48.58	**16.36**	**36.72**	17.73	**47.75**	**31.63**	19.01	15.5	34.81	**22.03**	28.58
	XGBoost	32.29	46.06	**42.73**	19.49	54.15	**16.27**	60.36	46.85	**18.68**	**15.33**	42.73	28.62	31.21
MedAE	gnn	33.71	42.66	32.94	76.88	**19.29**	53.85	37.47	49.37	39.86	36.77	133.57	111.7	118.14
	dnn	71.16	72.56	47.9	74.81	55.85	70.91	69.21	72.43	41.73	44.49	98.33	92.17	115.85
	cport	**29.48**	**27.28**	**20.49**	**47.93**	20.13	**46.85**	**31.13**	**28.47**	**24.34**	28.74	58.98	**63.57**	**73.66**
	XGBoost	31.87	44.43	22.01	62.3	32.32	60.02	35.16	48.01	31.52	30.48	71.94	83.75	112.13
MedAPE	gnn	14.46	23.32	35.14	14.63	16.14	11.62	40.38	22.22	12.64	13.45	36.77	14.04	14.69
	dnn	22.57	37.48	33.67	14.85	32.92	15.4	46.28	36.31	12.3	17.49	33.07	10.96	12.98
	cport	**12.23**	**15.97**	29.88	**8.93**	**13.35**	**10.21**	28.48	**18.5**	**7.41**	**10.77**	**24.57**	**8.59**	**9.02**
	XGBoost	15.15	22.54	**16.33**	12.65	26.14	12.67	**22.88**	25.24	10.57	11.14	**24.57**	10.47	13.34
Spearman	gnn	0.94	0.8	0.87	0.85	0.88	**0.89**	0.77	0.77	0.77	0.86	0.29	0.84	0.84
	dnn	0.89	0.67	0.85	0.82	0.8	0.79	0.7	0.63	0.79	0.73	0.44	0.85	0.84
	cport	**0.95**	**0.81**	0.91	**0.91**	**0.89**	0.87	0.77	**0.8**	**0.88**	0.85	0.6	**0.93**	**0.92**
	XGBoost	0.94	0.78	**0.92**	0.87	0.84	0.88	**0.84**	**0.8**	0.84	**0.89**	**0.69**	0.84	0.84

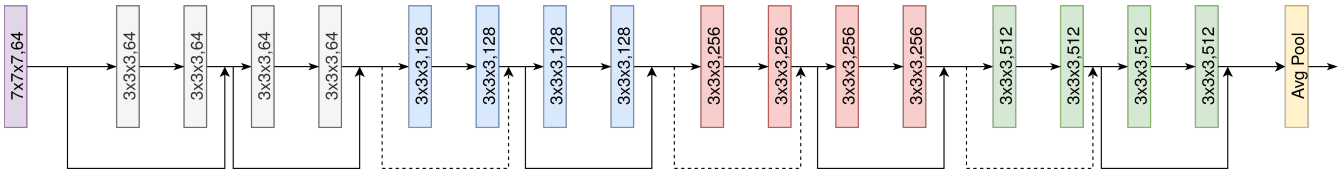

Fig. 1 The 3D convolutional neural network architecture of CPORT.

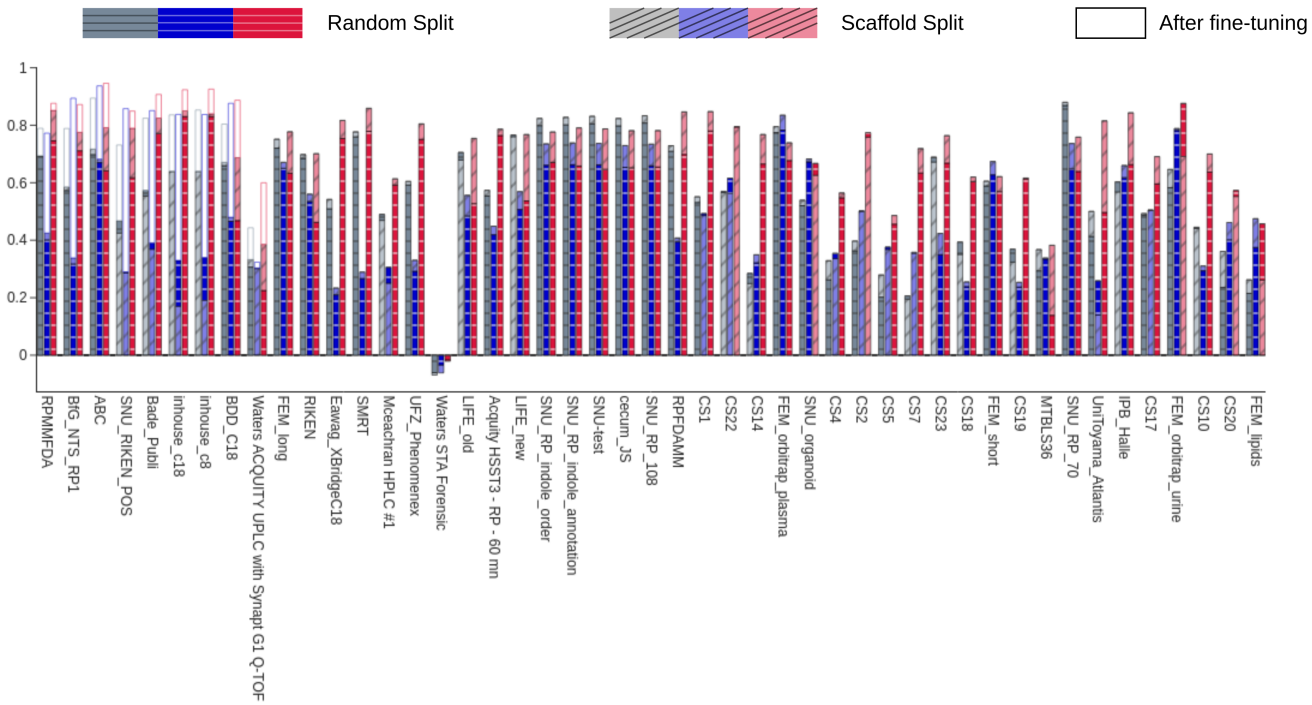

Fig. 2 The Spearman's rank correlations coefficients calculated for each external RPLC dataset. Gray, blue, and red colors correspond to the DNN, GNN, and CPORT models, respectively.

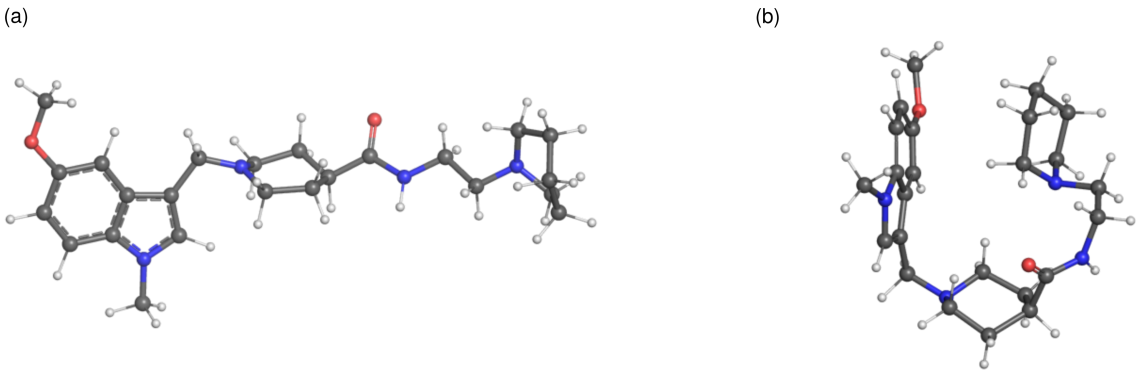

Fig. 3 Conformers of a hydrophobic molecule (a) generated with RDKit, (b) sampled from the molecular dynamics simulations in water.

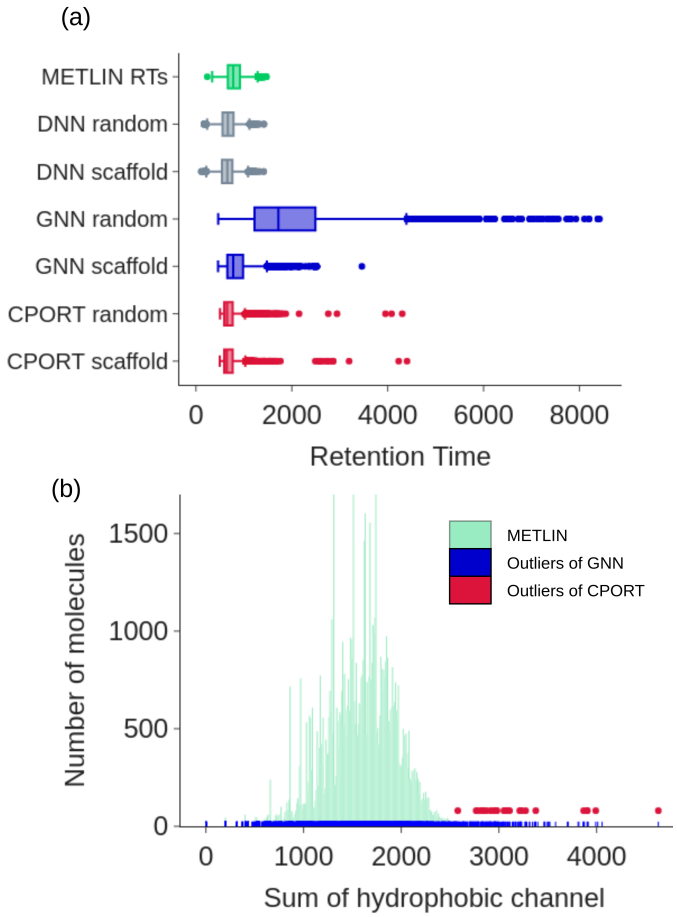

Fig. 4 (a) Predictions of models for the external datasets; (b) Sum of hydrophobicity channel for molecules from METLIN(green) dataset, and for molecules from external datasets for which GNN and CPORT models predict retention times greater than 2000 seconds.

Fig. 5 Kernel density estimations of RT values from METLIN (green), external datasets (black), predictions of the original CPORT (red), and predictions of CPORT after fine-tuning (magenta).

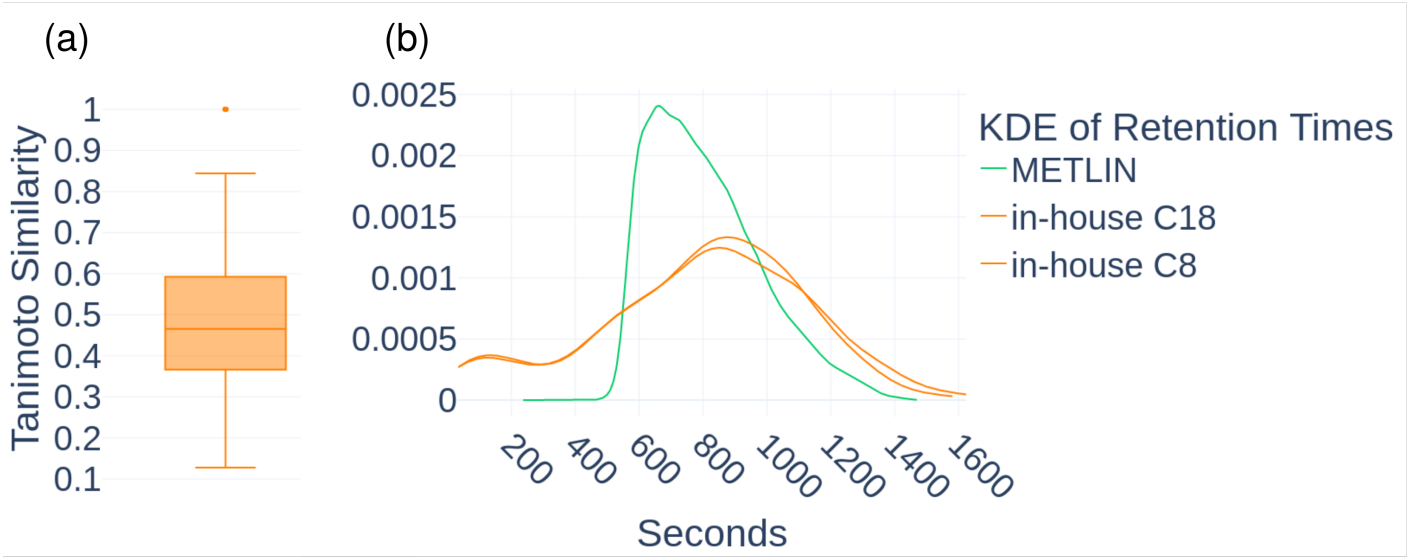

Fig. 6 (a) Boxplot of Tanimoto Similarities between molecules in the in-house and METLIN datasets; (b) Kernel Density Estimation of Retention Times of METLIN and in-house datasets.

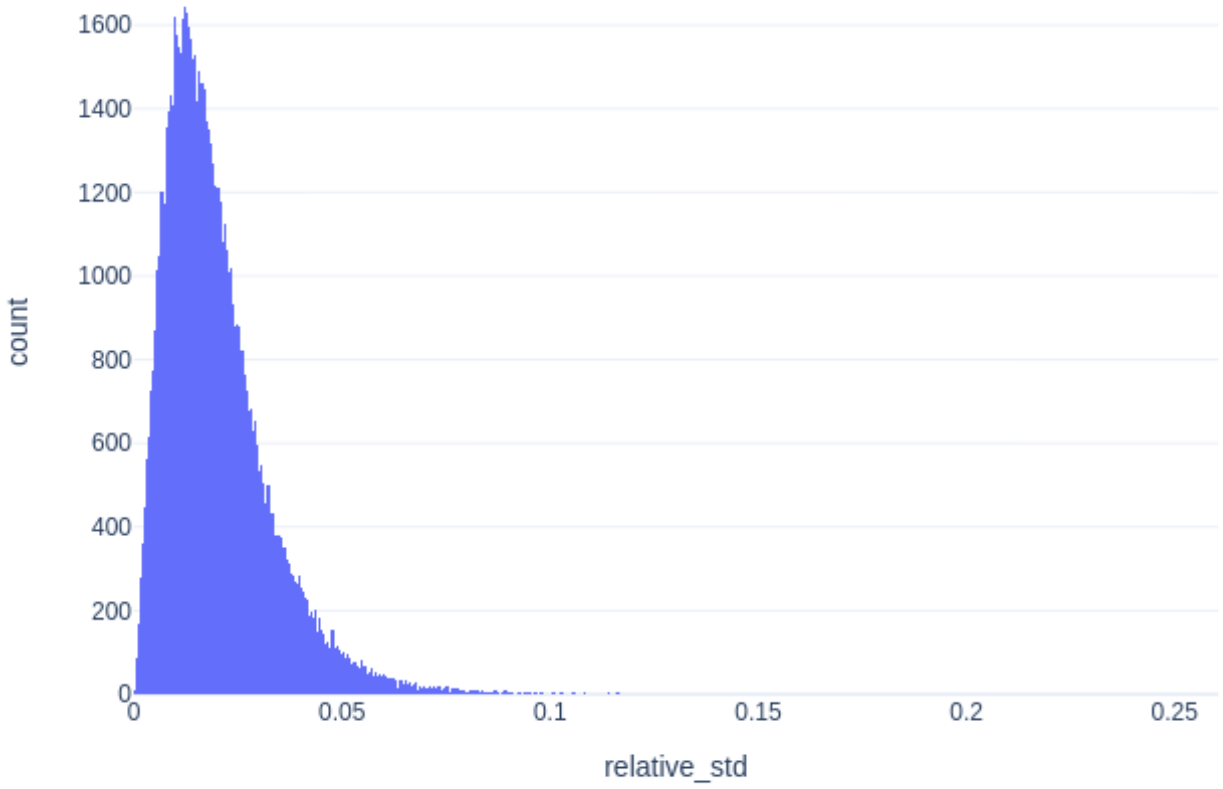

Fig. 7 Histogram of normalized standard deviation from the predictions obtained for four different conformers of each molecule m in the train and test sets.

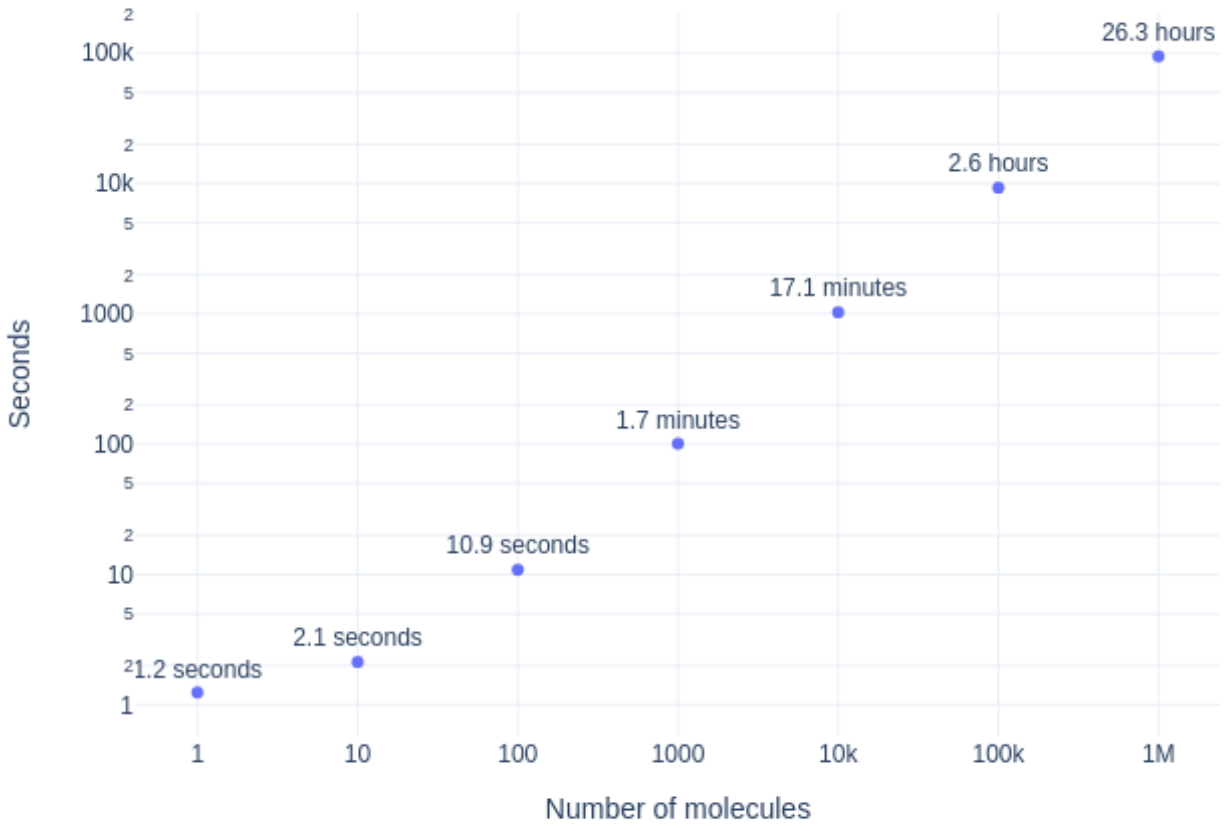

Fig. 8 Hours for processing and scoring 1000 random molecules from the METLIN dataset.

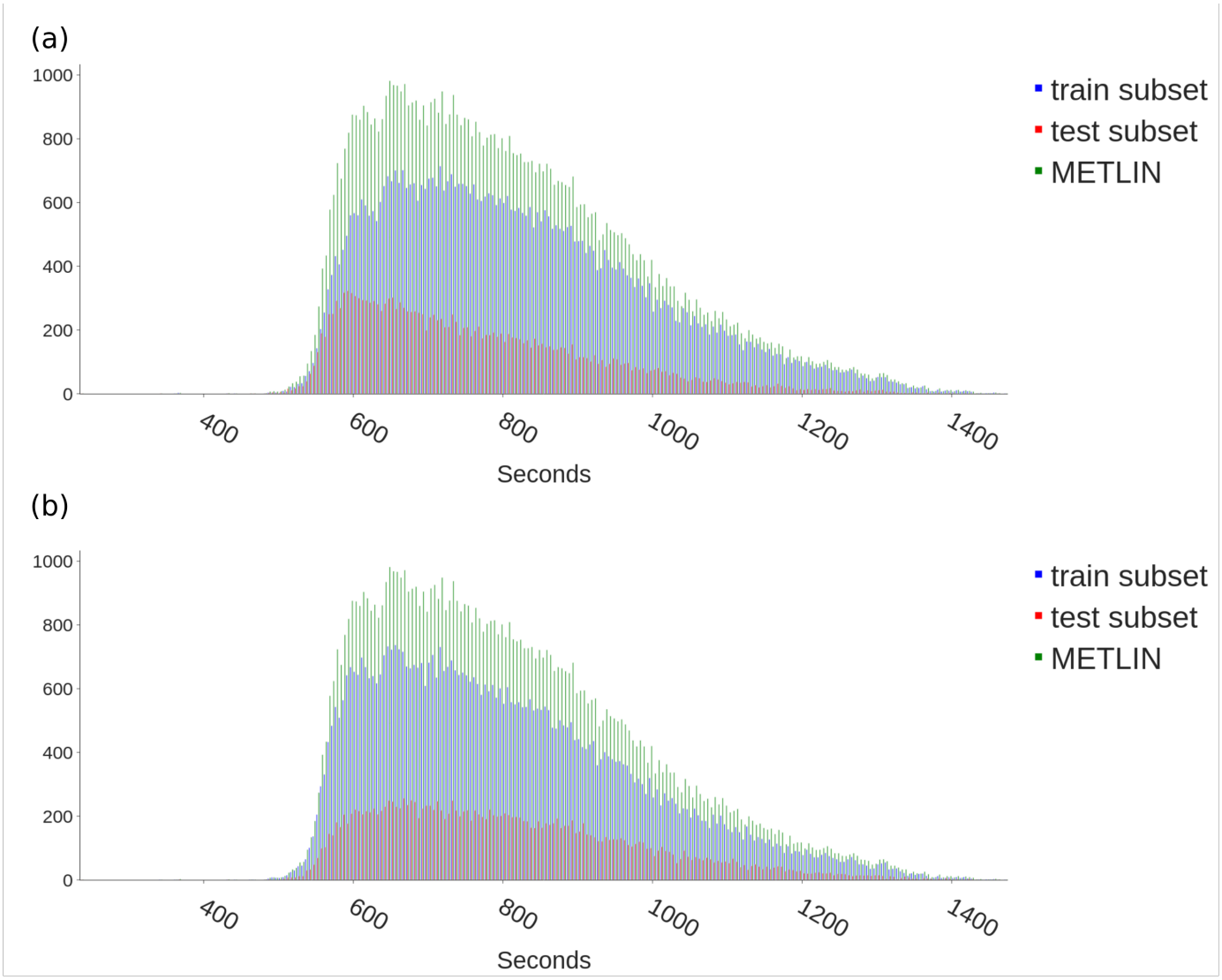

Fig. 9 Distribution of retention times in METLIN dataset and its train and test subsets after scaffold splitting using (a) the DeepChem library and (b) additional shuffling used in this work.

Table 4 Averaged Spearman's correlation rank coefficients on external datasets corresponding to the 4-fold cross-validation models trained using random and scaffold splits.

Name	cport_random	gnn_random	dnn_random	cport_scaffold	gnn_scaffold	dnn_scaffold	type
FEM_long	0.63 ± 0.22	0.65 ± 0.09	0.72 ± 0.03	$\mathbf{0.78 \pm 0.17}$	0.67 ± 0.07	0.75 ± 0.02	RPLC
FEM_short	0.57 ± 0.06	0.63 ± 0.05	0.59 ± 0.05	0.62 ± 0.06	$\mathbf{0.67 \pm 0.03}$	0.61 ± 0.03	RPLC
IPB_Halle	0.66 ± 0.17	0.62 ± 0.04	0.6 ± 0.02	$\mathbf{0.84 \pm 0.03}$	0.66 ± 0.07	0.57 ± 0.03	RPLC
MTBLS87	0.03 ± 0.32	-0.12 ± 0.14	-0.19 ± 0.06	-0.21 ± 0.11	-0.22 ± 0.2	-0.25 ± 0.03	HILIC
RIKEN	0.46 ± 0.22	0.53 ± 0.08	0.68 ± 0.02	$\mathbf{0.7 \pm 0.1}$	0.56 ± 0.08	0.7 ± 0.02	RPLC
FEM_orbitrap_plasma	0.68 ± 0.15	0.78 ± 0.04	0.77 ± 0.06	0.74 ± 0.29	$\mathbf{0.83 \pm 0.05}$	0.8 ± 0.02	RPLC
FEM_orbitrap_urine	$\mathbf{0.88 \pm 0.07}$	0.78 ± 0.03	0.58 ± 0.08	0.69 ± 0.31	0.79 ± 0.08	0.65 ± 0.05	RPLC
MTBLS36	0.14 ± 0.16	0.33 ± 0.1	0.29 ± 0.04	0.38 ± 0.1	0.34 ± 0.09	0.37 ± 0.02	RPLC
UFZ_Phenomenex	0.75 ± 0.08	0.29 ± 0.08	0.59 ± 0.03	$\mathbf{0.8 \pm 0.04}$	0.33 ± 0.06	0.6 ± 0.02	RPLC
UniToyama_Atlantis	0.5 ± 0.58	0.26 ± 0.34	0.41 ± 0.12	$\mathbf{0.82 \pm 0.07}$	0.14 ± 0.15	0.5 ± 0.04	RPLC
Eawag_XBridgeC18	0.75 ± 0.06	0.21 ± 0.04	0.51 ± 0.03	$\mathbf{0.82 \pm 0.04}$	0.23 ± 0.09	0.54 ± 0.01	RPLC
Cao_HILIC	-0.42 ± 0.26	-0.36 ± 0.1	-0.57 ± 0.04	$\mathbf{-0.69 \pm 0.04}$	-0.41 ± 0.06	-0.58 ± 0.01	HILIC
Waters ACQUITY UPLC with Synapt G1 Q-TOF	0.22 ± 0.14	0.3 ± 0.08	0.31 ± 0.0	0.39 ± 0.03	0.3 ± 0.04	0.33 ± 0.04	RPLC
KI_GIAR_zic_HILIC_pH2_7	-0.31 ± 0.21	-0.19 ± 0.14	$\mathbf{-0.58 \pm 0.01}$	-0.58 ± 0.06	-0.28 ± 0.11	-0.57 ± 0.03	HILIC
CS10	0.64 ± 0.24	0.29 ± 0.12	0.44 ± 0.13	$\mathbf{0.7 \pm 0.13}$	0.31 ± 0.09	0.44 ± 0.11	RPLC
CS7	0.63 ± 0.14	0.36 ± 0.15	0.21 ± 0.1	$\mathbf{0.72 \pm 0.05}$	0.35 ± 0.03	0.2 ± 0.12	RPLC
CS23	0.67 ± 0.09	0.35 ± 0.2	0.69 ± 0.07	$\mathbf{0.76 \pm 0.08}$	0.42 ± 0.1	0.67 ± 0.04	RPLC
HILIC_BDD_2	-0.25 ± 0.33	-0.21 ± 0.07	-0.45 ± 0.05	$\mathbf{-0.56 \pm 0.05}$	-0.29 ± 0.14	-0.47 ± 0.03	HILIC
Bade_Publi	0.77 ± 0.05	0.39 ± 0.09	0.57 ± 0.01	$\mathbf{0.82 \pm 0.02}$	0.37 ± 0.03	0.55 ± 0.03	RPLC
Acquity HSST3 - RP - 60 mn	0.76 ± 0.01	0.42 ± 0.09	0.55 ± 0.03	$\mathbf{0.79 \pm 0.03}$	0.45 ± 0.03	0.57 ± 0.03	RPLC
CS2	$\mathbf{0.77 \pm 0.05}$	0.5 ± 0.06	0.36 ± 0.07	0.76 ± 0.1	0.5 ± 0.01	0.4 ± 0.07	RPLC
CS1	0.78 ± 0.18	0.49 ± 0.11	0.53 ± 0.11	$\mathbf{0.85 \pm 0.09}$	0.49 ± 0.04	0.55 ± 0.05	RPLC
CS17	0.6 ± 0.13	0.5 ± 0.15	0.49 ± 0.07	$\mathbf{0.69 \pm 0.05}$	0.5 ± 0.1	0.49 ± 0.08	RPLC
BDD_C18	0.47 ± 0.24	0.47 ± 0.08	0.66 ± 0.02	$\mathbf{0.69 \pm 0.1}$	0.48 ± 0.13	0.67 ± 0.01	RPLC
CS20	$\mathbf{0.57 \pm 0.11}$	0.4 ± 0.19	0.24 ± 0.11	0.55 ± 0.11	0.46 ± 0.11	0.36 ± 0.2	RPLC
LIFE_old	0.53 ± 0.24	0.48 ± 0.1	0.71 ± 0.03	$\mathbf{0.75 \pm 0.04}$	0.56 ± 0.04	0.68 ± 0.02	RPLC
LIFE_new	0.54 ± 0.27	0.51 ± 0.12	$\mathbf{0.77 \pm 0.01}$	0.77 ± 0.06	0.57 ± 0.08	0.76 ± 0.02	RPLC
CS4	0.55 ± 0.12	0.35 ± 0.12	0.26 ± 0.09	$\mathbf{0.56 \pm 0.09}$	0.34 ± 0.08	0.33 ± 0.12	RPLC
CS5	0.46 ± 0.11	0.38 ± 0.1	0.2 ± 0.13	0.49 ± 0.08	0.37 ± 0.06	0.28 ± 0.15	RPLC
CS14	0.67 ± 0.13	0.32 ± 0.1	0.28 ± 0.1	$\mathbf{0.77 \pm 0.05}$	0.35 ± 0.04	0.25 ± 0.09	RPLC
CS18	0.6 ± 0.17	0.24 ± 0.04	0.39 ± 0.08	$\mathbf{0.62 \pm 0.09}$	0.25 ± 0.0	0.35 ± 0.04	RPLC
CS19	0.61 ± 0.16	0.24 ± 0.04	0.37 ± 0.08	$\mathbf{0.62 \pm 0.09}$	0.25 ± 0.02	0.32 ± 0.04	RPLC
CS22	$\mathbf{0.8 \pm 0.07}$	0.62 ± 0.1	0.57 ± 0.04	0.79 ± 0.08	0.56 ± 0.03	0.57 ± 0.04	RPLC
RPMMFDA	0.75 ± 0.11	0.4 ± 0.07	0.69 ± 0.01	$\mathbf{0.85 \pm 0.03}$	0.42 ± 0.03	0.69 ± 0.0	RPLC
RPFDAMM	0.7 ± 0.15	0.39 ± 0.08	0.71 ± 0.04	$\mathbf{0.85 \pm 0.09}$	0.41 ± 0.11	0.73 ± 0.02	RPLC
IJM_TEST	-0.04 ± 0.1	-0.35 ± 0.05	-0.25 ± 0.05	-0.19 ± 0.04	-0.41 ± 0.04	-0.23 ± 0.03	HILIC
SMRT	0.78 ± 0.09	0.27 ± 0.08	0.76 ± 0.02	$\mathbf{0.86 \pm 0.02}$	0.29 ± 0.05	0.78 ± 0.01	RPLC
ABC	0.64 ± 0.22	0.67 ± 0.07	0.7 ± 0.03	$\mathbf{0.79 \pm 0.11}$	0.68 ± 0.06	0.72 ± 0.03	RPLC
SNU-test	0.65 ± 0.2	0.66 ± 0.08	0.8 ± 0.02	0.79 ± 0.16	0.74 ± 0.06	$\mathbf{0.83 \pm 0.01}$	RPLC
SNU_RP_70	0.64 ± 0.23	0.65 ± 0.11	0.87 ± 0.03	0.76 ± 0.2	0.74 ± 0.08	$\mathbf{0.88 \pm 0.01}$	RPLC
SNU_RP_108	0.65 ± 0.2	0.66 ± 0.08	0.81 ± 0.02	0.78 ± 0.16	0.73 ± 0.06	$\mathbf{0.83 \pm 0.01}$	RPLC
HIILIC_tip	-0.41 ± 0.2	-0.32 ± 0.09	-0.52 ± 0.02	$\mathbf{-0.62 \pm 0.02}$	-0.4 ± 0.02	-0.52 ± 0.03	HILIC
SNU_RIKEN_POS	0.62 ± 0.22	0.29 ± 0.16	0.47 ± 0.06	$\mathbf{0.79 \pm 0.02}$	0.29 ± 0.05	0.42 ± 0.05	RPLC
cecum_JS	0.65 ± 0.2	0.65 ± 0.09	0.8 ± 0.02	0.78 ± 0.16	0.73 ± 0.06	$\mathbf{0.82 \pm 0.01}$	RPLC
SNU_RP_indole_annotation	0.66 ± 0.2	0.66 ± 0.08	0.8 ± 0.02	0.79 ± 0.15	0.74 ± 0.06	$\mathbf{0.83 \pm 0.01}$	RPLC
SNU_RP_indole_order	0.67 ± 0.2	0.66 ± 0.08	0.8 ± 0.02	0.78 ± 0.17	0.74 ± 0.06	$\mathbf{0.82 \pm 0.01}$	RPLC
SNU_organoid	0.67 ± 0.13	0.67 ± 0.03	0.52 ± 0.03	0.63 ± 0.22	$\mathbf{0.68 \pm 0.04}$	0.54 ± 0.02	RPLC
Meister zic-pHILIC pH9.3	-0.18 ± 0.16	-0.3 ± 0.15	-0.51 ± 0.01	-0.49 ± 0.05	-0.41 ± 0.04	$\mathbf{-0.53 \pm 0.01}$	HILIC
AjsUoB	-0.4 ± 0.29	-0.38 ± 0.11	-0.6 ± 0.03	$\mathbf{-0.7 \pm 0.04}$	-0.44 ± 0.06	-0.61 ± 0.01	HILIC
AjsTestF	-0.39 ± 0.21	-0.32 ± 0.1	-0.53 ± 0.02	$\mathbf{-0.63 \pm 0.01}$	-0.41 ± 0.02	-0.52 ± 0.03	HILIC
Waters STA Forensic	-0.02 ± 0.02	-0.03 ± 0.05	-0.06 ± 0.04	-0.02 ± 0.01	-0.06 ± 0.05	-0.07 ± 0.02	RPLC
BfG_NTS_RP1	0.71 ± 0.08	0.32 ± 0.05	0.57 ± 0.02	$\mathbf{0.78 \pm 0.05}$	0.34 ± 0.04	0.59 ± 0.01	RPLC
Mceachran HPLC 1	0.59 ± 0.02	0.31 ± 0.07	0.49 ± 0.03	$\mathbf{0.61 \pm 0.02}$	0.25 ± 0.05	0.47 ± 0.03	RPLC
FEM_lipids	0.46 ± 0.31	0.37 ± 0.13	0.22 ± 0.15	0.26 ± 0.61	0.47 ± 0.12	0.26 ± 0.09	RPLC
in-house_c8	0.83 ± 0.03	0.34 ± 0.03	0.64 ± 0.02	$\mathbf{0.84 \pm 0.03}$	0.19 ± 0.09	0.64 ± 0.01	RPLC
in-house_c18	0.83 ± 0.02	0.33 ± 0.03	0.64 ± 0.02	$\mathbf{0.85 \pm 0.02}$	0.17 ± 0.09	0.64 ± 0.01	RPLC

Table 5 Ablation studies

Name	MAE(MAPE)	MedAE(MedAPE)
Full model	44 (5.5 %)	26 (3.4 %)
without hydrophobicity channel	+11 (1.5 %)	+12 (1.5 %)
without aromaticity channel	+25 (3.3 %)	+27 (3.3 %)
without h-bond donor channel	+16 (2.0 %)	+16 (2.1 %)
without h-bond acceptor channel	+26 (2.7 %)	+25 (3.3 %)
without positive ionizable channel	+1 (0.1 %)	+3 (0.3 %)
without negative ionazible channel	+11 (1.6 %)	+11 (1.5 %)
without occupancy channel	+9 (1.1 %)	+10 (1.3 %)

Table 6 RDKit Descriptors

MaxEStateIndex	MinEStateIndex	MaxAbsEStateIndex	MinAbsEStateIndex	qed	MolWt	HeavyAtomMolWt	ExactMolWt	NumValenceElectrons	NumRadicalElectrons
MaxPartialCharge	MinPartialCharge	MaxAbsPartialCharge	MinAbsPartialCharge	FpDensityMorgan1	FpDensityMorgan2	FpDensityMorgan3	BalabanJ	BertzCT	Chi0
Chi0n	Chi0v	Chi1	Chi1n	Chi1v	Chi2n	Chi2v	Chi3n	Chi3v	Chi4n
Chi4v	HallKierAlpha	Ipc	Kappa1	Kappa2	Kappa3	LabuteASA	PEOE_VSA1	PEOE_VSA10	PEOE_VSA11
PEOE_VSA12	PEOE_VSA13	PEOE_VSA14	PEOE_VSA2	PEOE_VSA3	PEOE_VSA4	PEOE_VSA5	PEOE_VSA6	PEOE_VSA7	PEOE_VSA8
PEOE_VSA9	SMR_VSA1	SMR_VSA10	SMR_VSA2	SMR_VSA3	SMR_VSA4	SMR_VSA5	SMR_VSA6	SMR_VSA7	SMR_VSA8
SMR_VSA9	SlogP_VSA1	SlogP_VSA10	SlogP_VSA11	SlogP_VSA12	SlogP_VSA2	SlogP_VSA3	SlogP_VSA4	SlogP_VSA5	SlogP_VSA6
SlogP_VSA7	SlogP_VSA8	SlogP_VSA9	TPSA	EState_VSA1	EState_VSA10	EState_VSA11	EState_VSA2	EState_VSA3	EState_VSA4
EState_VSA5	EState_VSA6	EState_VSA7	EState_VSA8	EState_VSA9	VSA_EState1	VSA_EState10	VSA_EState2	VSA_EState3	VSA_EState4
VSA_EState5	VSA_EState6	VSA_EState7	VSA_EState8	VSA_EState9	FractionCSP3	HeavyAtomCount	NHOHCount	NOCount	NumAliphaticCarbocycles

Table 7 Grid parameters for XGBoost

parameter	values
learning_rate	[0.1, 0.01, 0.05]
max_depth	[2, 4, 6, 8]
n_estimators	[50, 100, 200]
num_boost_round	[50, 100, 200, 500, 1000]