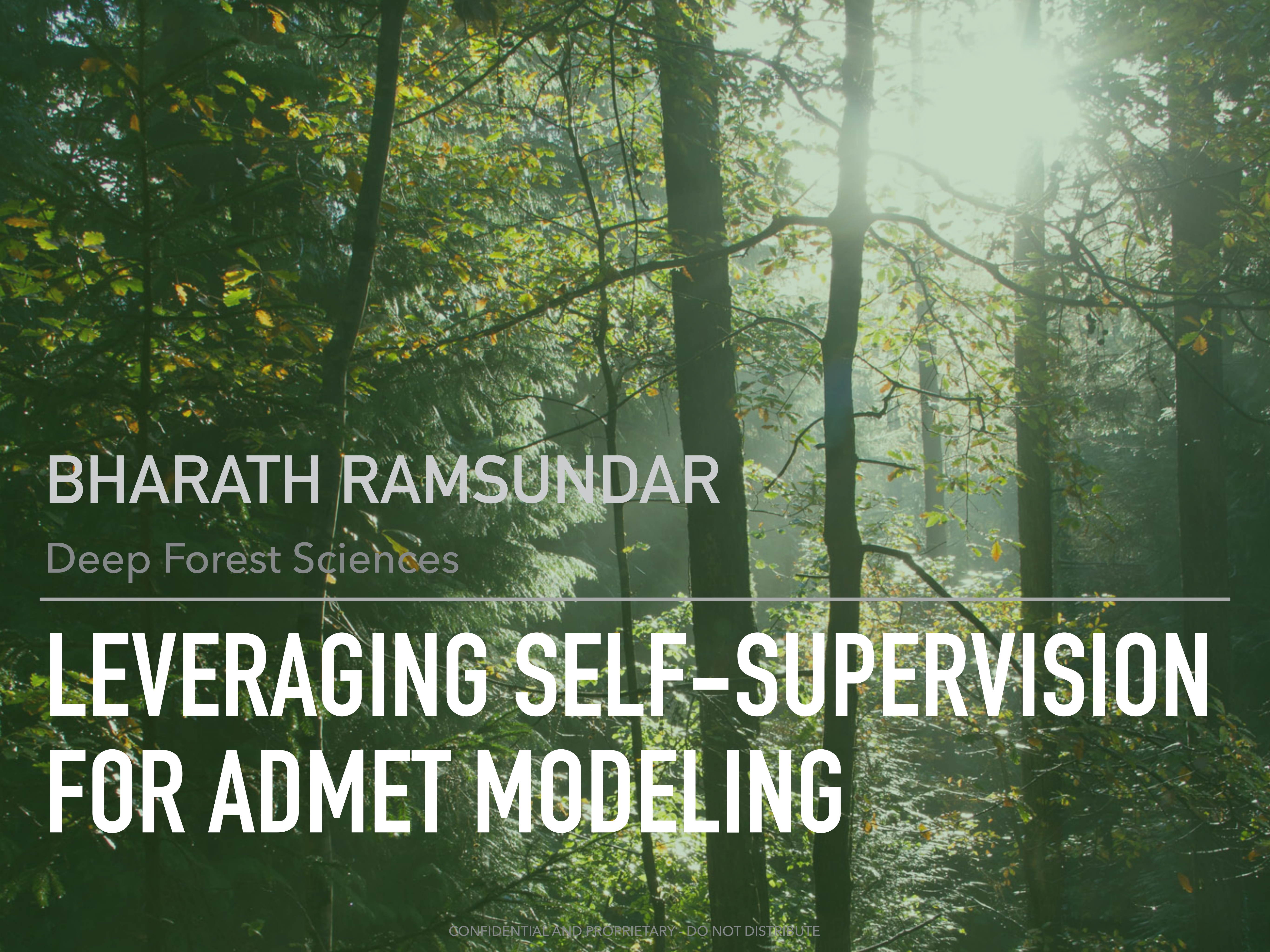
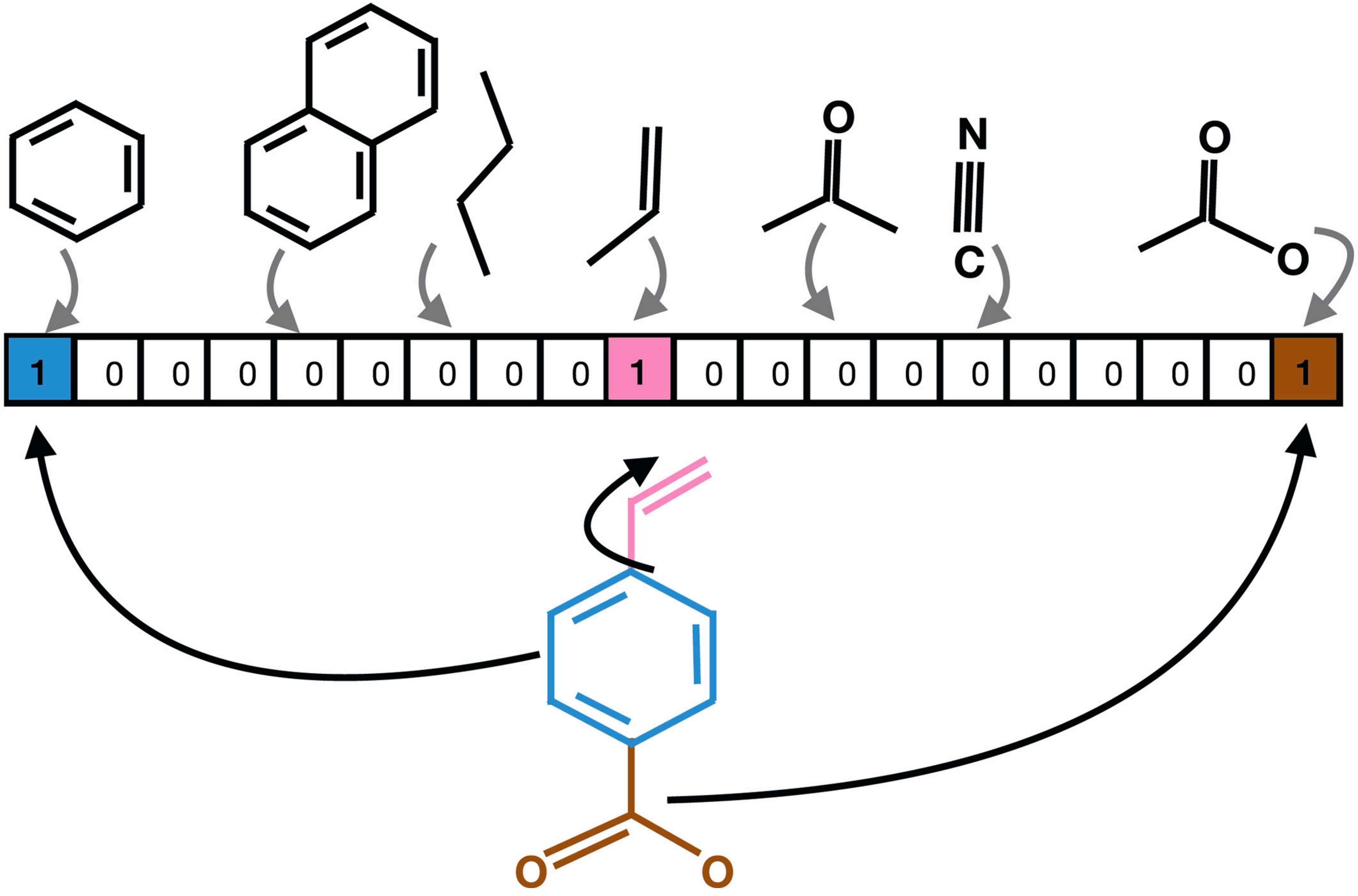


BHARATH RAMSUNDAR

Deep Forest Sciences

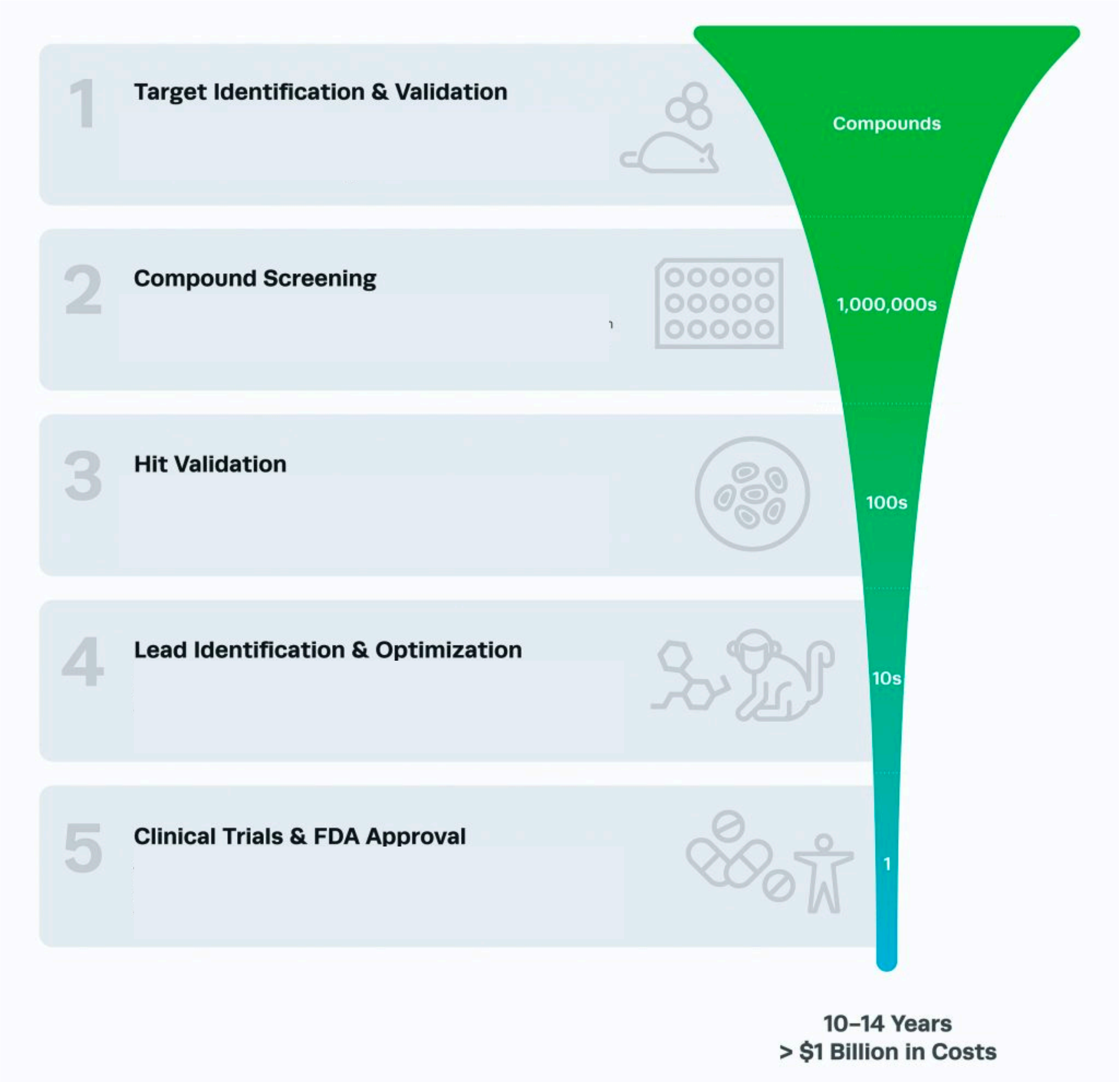
LEVERAGING SELF-SUPERVISION FOR ADMET MODELING

MACHINE LEARNING 101: TURN MOLECULES INTO VECTORS



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

LOW DATA IS A FUNDAMENTAL CHALLENGE FOR DRUG DISCOVERY



1 Target Identification & Validation

2 Compound Screening

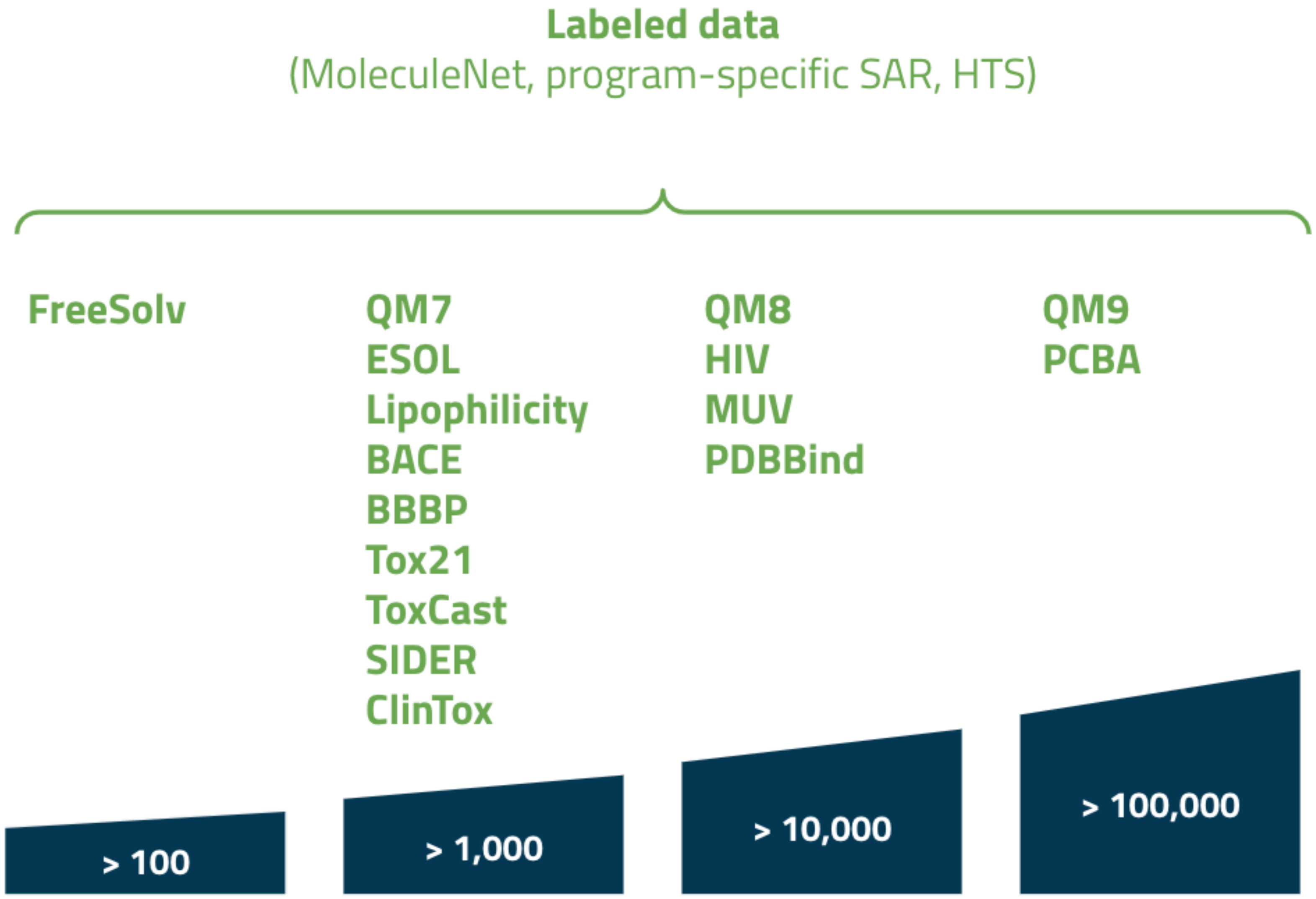
3 Hit Validation

4 Lead Identification & Optimization

5 Clinical Trials & FDA Approval

10–14 Years
> $1 Billion in Costs

ASSAY READOUTS ARE EXPENSIVE



Labeled data

(MoleculeNet, program-specific SAR, HTS)

FreeSolv

> 100

QM7
ESOL
Lipophilicity
BACE
BBBP
Tox21
ToxCast
SIDER
ClinTox

> 1,000

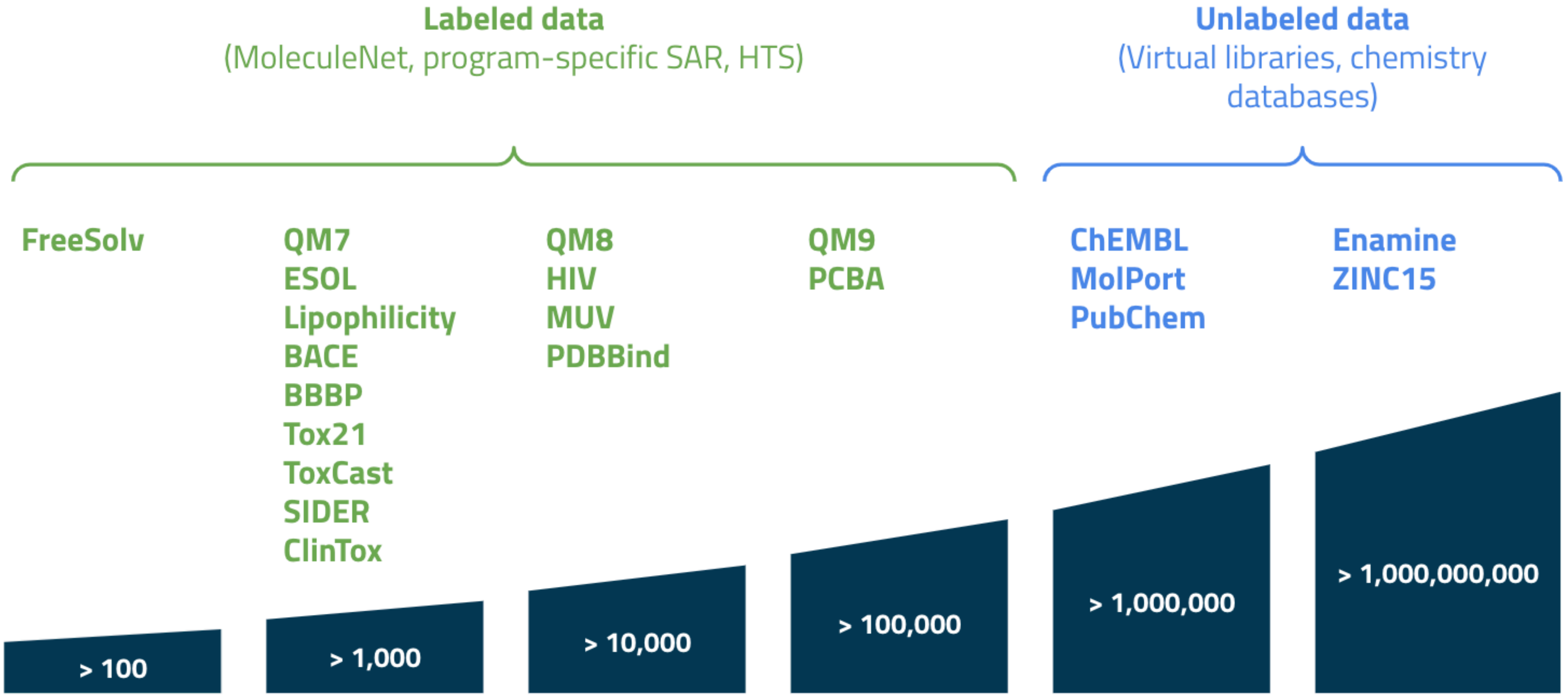
QM8
HIV
MUV
PDBBind

> 10,000

QM9
PCBA

> 100,000

ASSAY READOUTS ARE EXPENSIVE, BUT CHEMICAL STRUCTURES ARE UNLIMITED

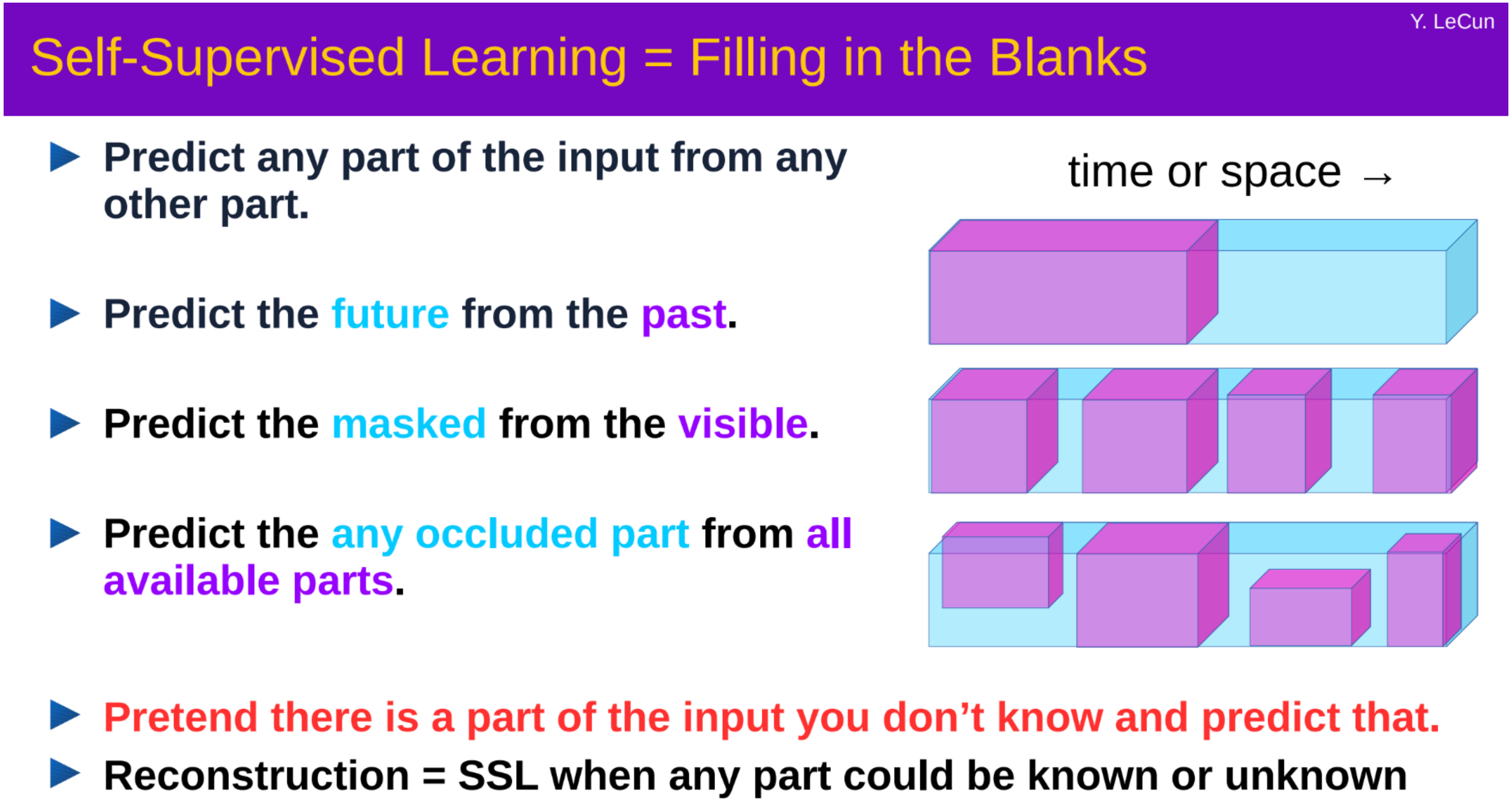


Labeled data
(MoleculeNet, program-specific SAR, HTS)

Unlabeled data
(Virtual libraries, chemistry databases)

FreeSolv	QM7 ESOL Lipophilicity BACE BBBP Tox21 ToxCast SIDER ClinTox	QM8 HIV MUV PDBBind	QM9 PCBA	ChEMBL MolPort PubChem	Enamine ZINC15
> 100	> 1,000	> 10,000	> 100,000	> 1,000,000	> 1,000,000,000

WHAT IS SELF-SUPERVISION?



Y. LeCun

Self-Supervised Learning = Filling in the Blanks

- ▶ Predict any part of the input from any other part.
- ▶ Predict the future from the past.
- ▶ Predict the masked from the visible.
- ▶ Predict the any occluded part from all available parts.

time or space →

- ▶ Pretend there is a part of the input you don't know and predict that.
- ▶ Reconstruction = SSL when any part could be known or unknown

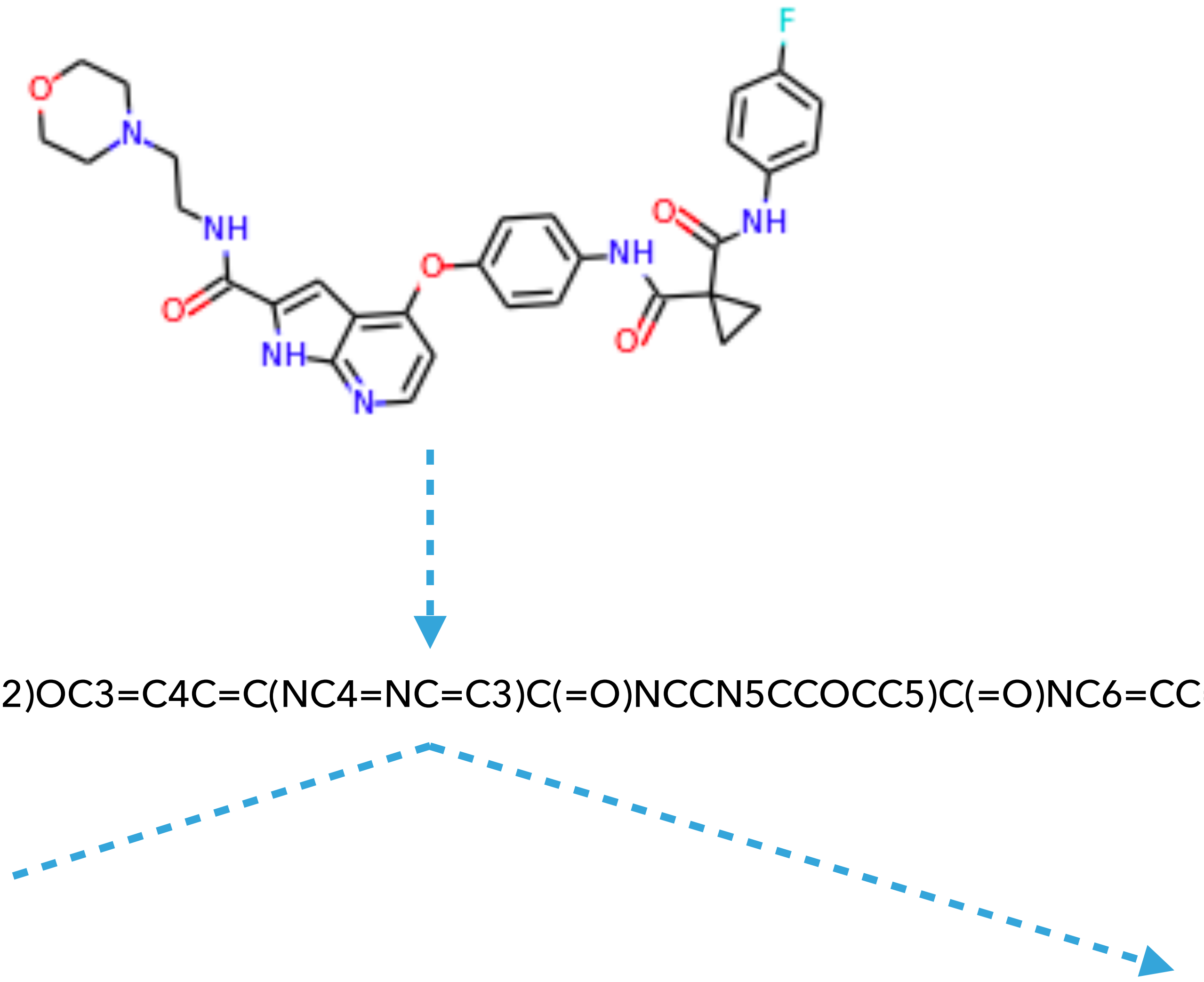
https://drive.google.com/file/d/1r-mDL4IX_hzZLDBKp8_e8VZqD7fOzBkF/view, Y. LeCun

HOW CAN WE SELF-SUPERVISE CHEMICAL MODELS?

- Chemical self-supervision is an active area of research with several different methods proposed in the literature.
  - Property prediction based approaches (predict chemical properties from structure)
  - In Fill Based Approaches (remove characters from SMILES and in-fill)
  - Graph Based Approaches (mutual information, node/edge in-fill)
  - Mutual Information Based Approaches (align 2D and 3D embeddings)

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

CHEMBERTA: PRETRAIN DIRECTLY FROM SMILES REPRESENTATION



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

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

SELF-SUPERVISION IMPROVES WITH MORE STRUCTURES SEEN

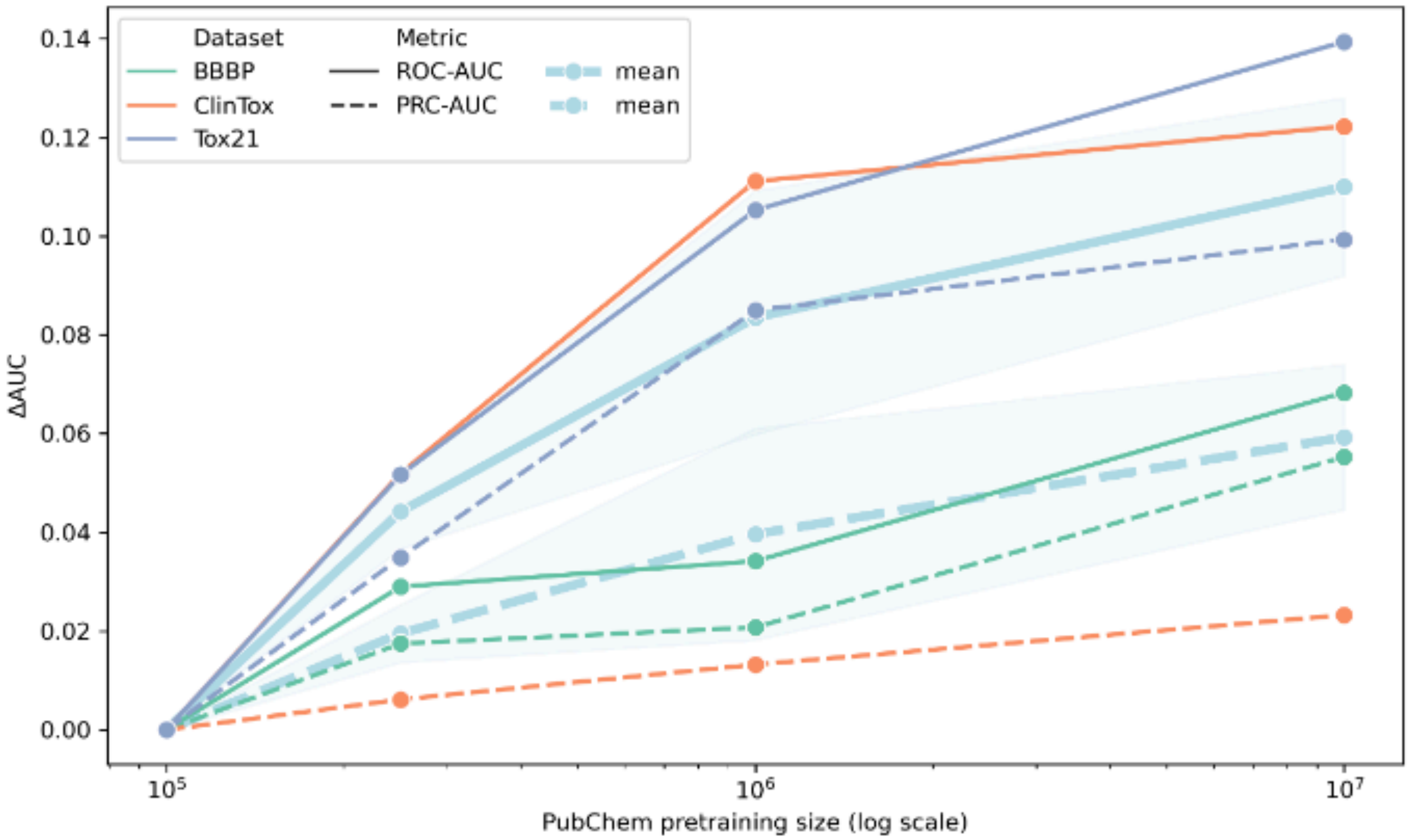
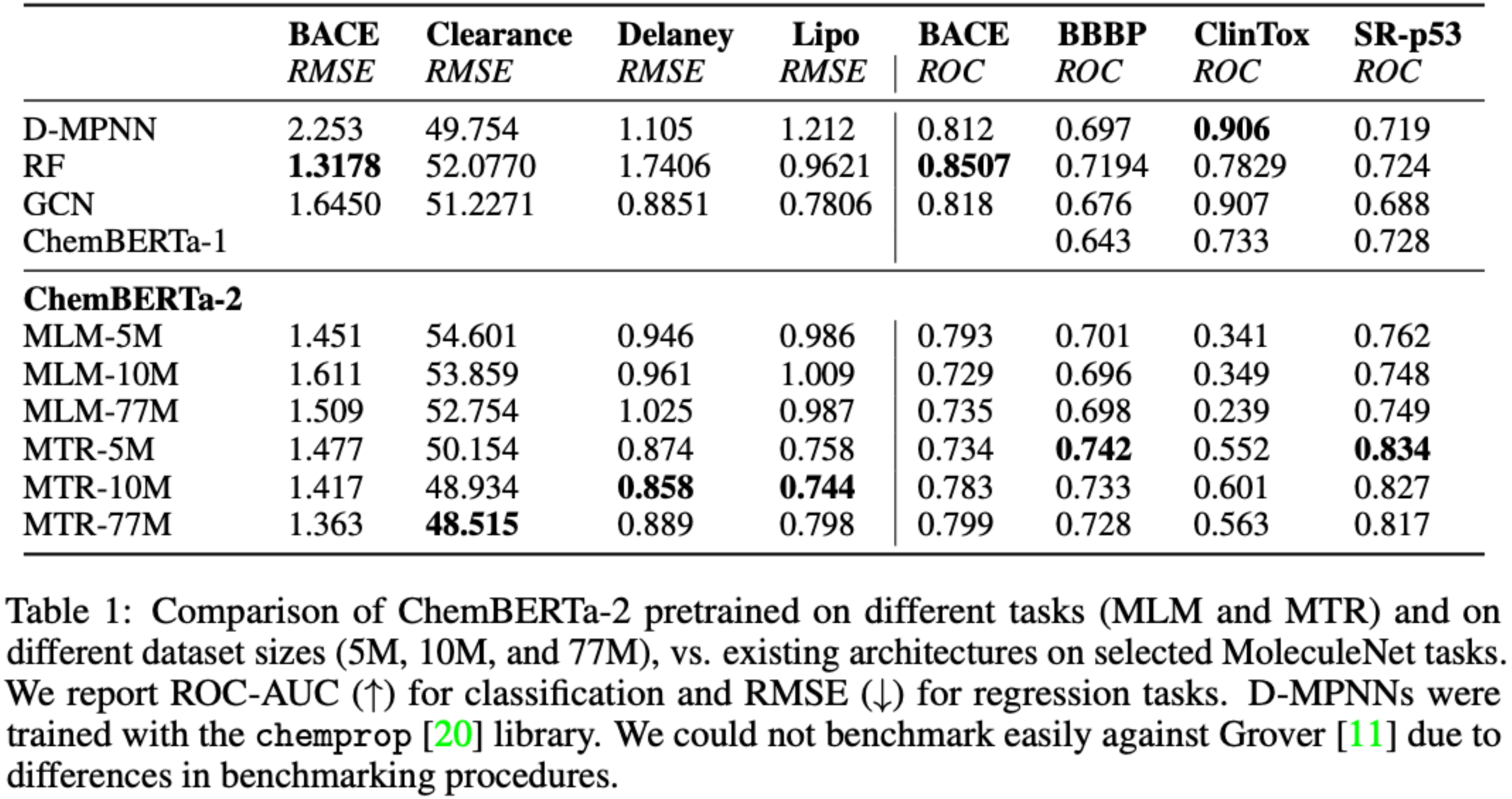


Figure 1: Scaling the pretraining size (100K, 250K, 1M, 10M) produces consistent improvements in downstream task performance on BBBP, ClinTox, and Tox21. (HIV was omitted from this analysis due to resource constraints.) Mean ΔAUC across all three tasks with a 68% confidence interval is shown in light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

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

CHEMBERTA OFFERS STRONG ADMET PERFORMANCE

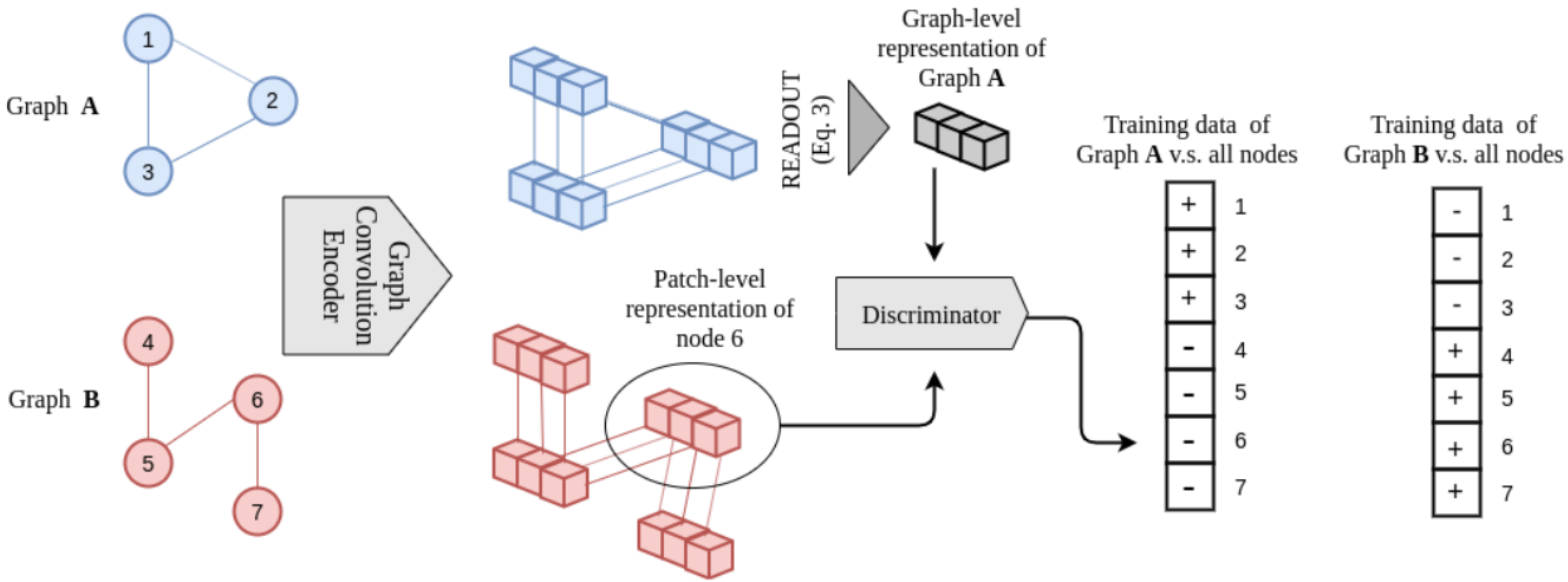


	BACE *RMSE*	Clearance *RMSE*	Delaney *RMSE*	Lipo *RMSE*	BACE *ROC*	BBBP *ROC*	ClinTox *ROC*	SR-p53 *ROC*
D-MPNN	2.253	49.754	1.105	1.212	0.812	0.697	**0.906**	0.719
RF	**1.3178**	52.0770	1.7406	0.9621	**0.8507**	0.7194	0.7829	0.724
GCN	1.6450	51.2271	0.8851	0.7806	0.818	0.676	0.907	0.688
ChemBERTa-1						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.

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

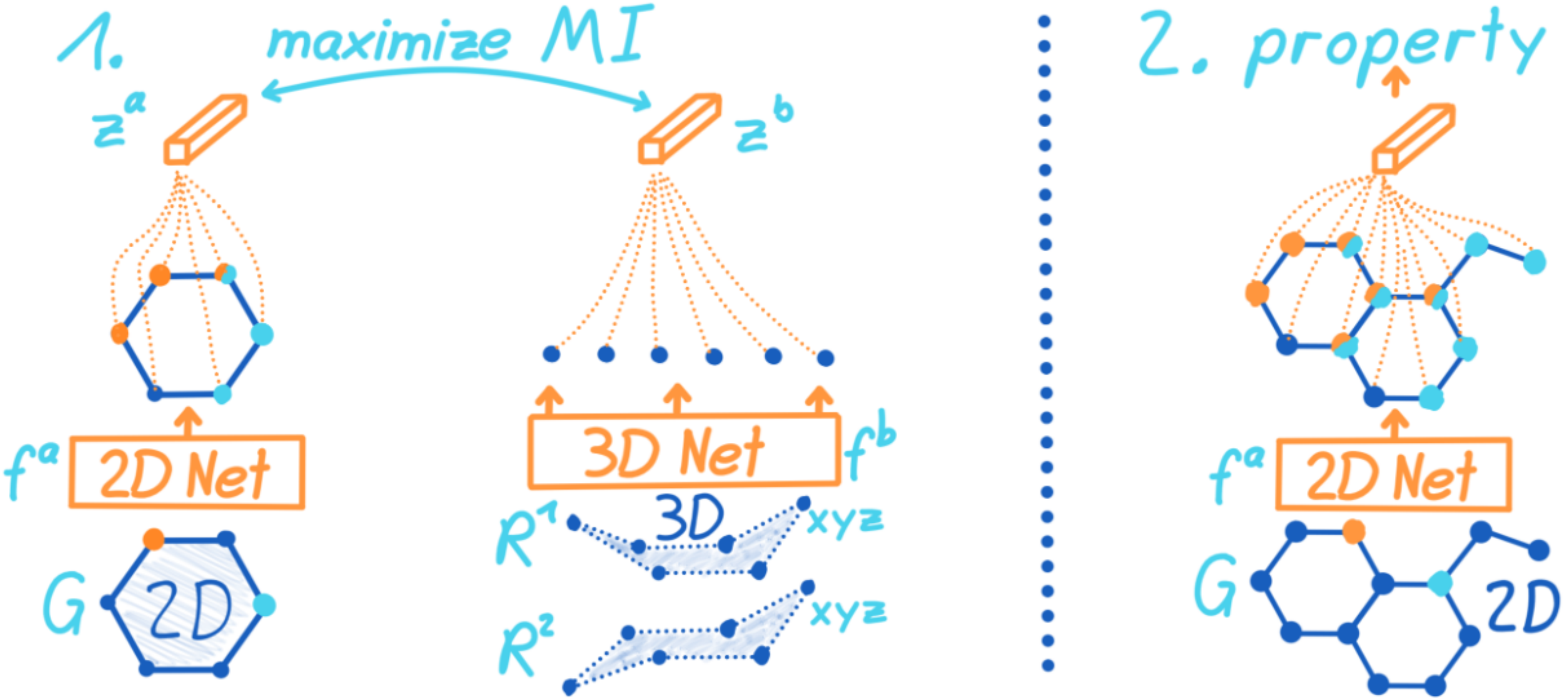
INFOGRAPH: GRAPH-LEVEL REPRESENTATION LEARNING VIA MUTUAL INFORMATION


Graph A
1
2
3
Graph B
4
5
6
7
Graph Convolution Encoder
READOUT (Eq. 3)
Graph-level representation of Graph A
Patch-level representation of node 6
Discriminator
Training data of Graph A v.s. all nodes
+ 1
+ 2
+ 3
- 4
- 5
- 6
- 7
Training data of Graph B v.s. all nodes
- 1
- 2
- 3
+ 4
+ 5
+ 6
+ 7

Sun, Fan-Yun, et al. "Infograph: Unsupervised and semi-supervised graph-level representation learning via mutual information maximization." arXiv preprint arXiv:1908.01000 (2019).

INFOMAX: ALIGNING 2D AND 3D REPRESENTATIONS WITH MUTUAL INFORMATION

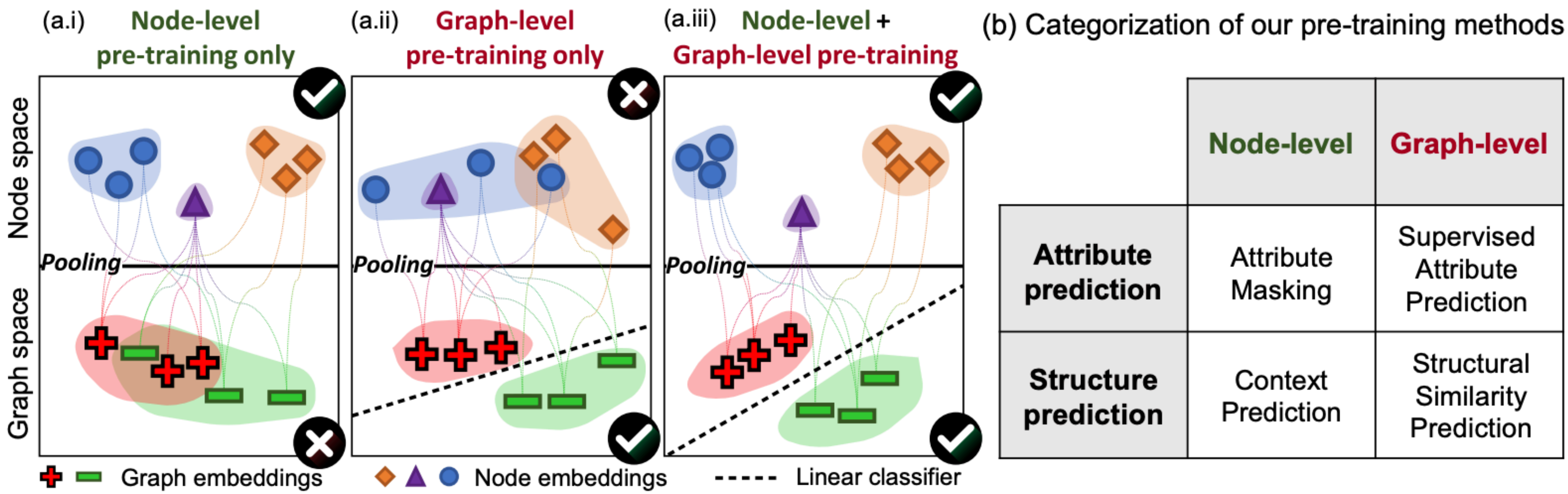
Maximizes the mutual information between learned 3D summary vectors and the representations of a graph neural network. The pretrained GNN is then finetuned for property prediction.



Stärk, Hannes, et al. "3d infomax improves gnns for molecular property prediction." International Conference on Machine Learning. PMLR, 2022.

SNAP: GRAPH AND NODE LEVEL PRETRAINING STRATEGIES

Maximizes the mutual information between the graph-level representation and the representations of substructures of different scales by discriminating if a subgraph belongs to another graph


(a.i) Node-level pre-training only
(a.ii) Graph-level pre-training only
(a.iii) Node-level + Graph-level pre-training
Node space
Graph space
Pooling
Pooling
Pooling
✚ ▬ Graph embeddings
◆ ▲ ● Node embeddings
- - - - Linear classifier

(b) Categorization of our pre-training methods

	Node-level	Graph-level
Attribute prediction	Attribute Masking	Supervised Attribute Prediction
Structure prediction	Context Prediction	Structural Similarity Prediction

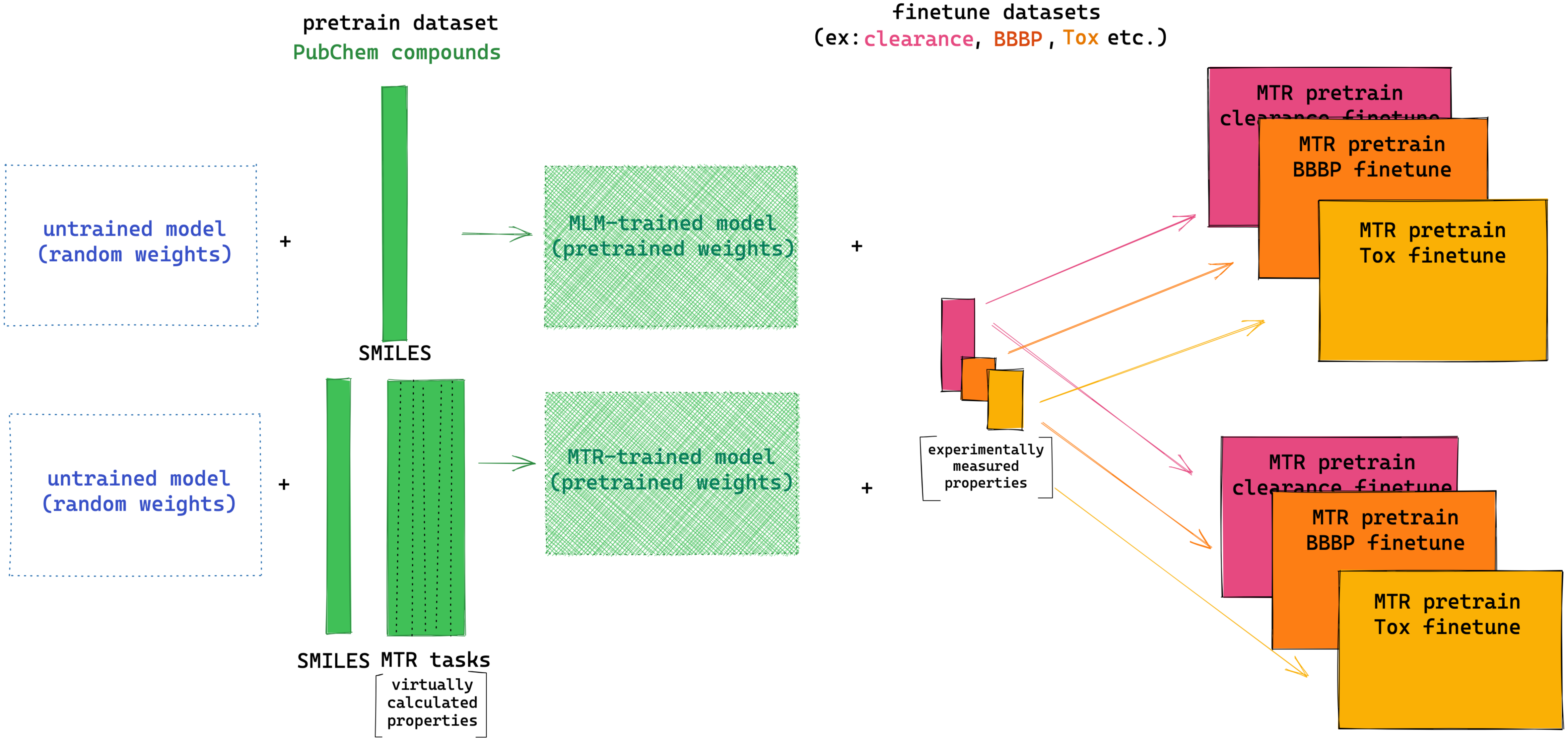
Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with one-shot learning. ACS central science, 3(4), 283-293.

PRELIMINARY ADMET BENCHMARKING RESULTS (UNTUNED)

	Lipo (RMSE ↓)	Tox21 (ROC-AUC ↑)	BBBP (ROC-AUC ↑)
ChemBERTa-MLM (250K)	0.929	0.482	-
ChemBERTa-MLM (1M)	0.927	-	-
ChemBERTa-MTR (1M)	0.964	0.5	0.490
InfoGraph (250K)	0.893	0.656	0.661
InfoGraph (1M)	0.893	0.669	0.670
InfoMax3D (250K)	0.869	0.652	0.661
Snap (250K)	0.855	0.675	0.646
Snap (1M)	0.869	0.680	0.637
RF	0.962	-	0.719

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

HYPERPARAMETER TUNING FOR PRETRAINING REMAINS CHALLENGING



pretrain dataset
PubChem compounds

untrained model
(random weights)

+

SMILES

MLM-trained model
(pretrained weights)

+

untrained model
(random weights)

+

SMILES MTR tasks

virtually calculated properties

MTR-trained model
(pretrained weights)

+

finetune datasets
(ex: clearance, BBBP, Tox etc.)

experimentally measured properties

MTR pretrain
clearance finetune

MTR pretrain
BBBP finetune

MTR pretrain
Tox finetune

MTR pretrain
clearance finetune

MTR pretrain
BBBP finetune

MTR pretrain
Tox finetune

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

CONCLUSIONS

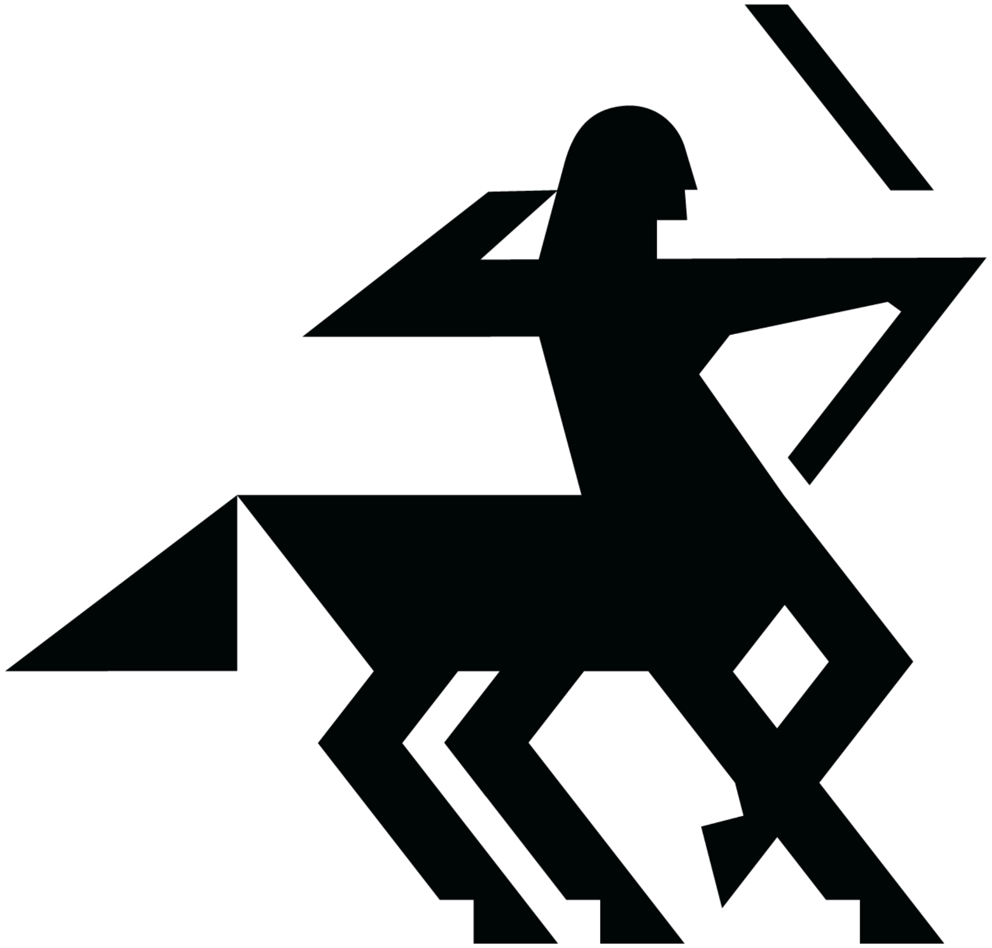
- Lack of data limits ADME modeling
- Chemical self-supervision allows for systematic incorporation of chemical priors which may help lower data needs.
- Multiple choices of self-supervision curricula. Not clear which is the best yet, but under active research.
- Hyperparameter tuning these models is challenging and requires large amounts of compute.

CHIRON MAKES MACHINE LEARNING IN DRUG DISCOVERY PRACTICAL

CONTACT US

bharath@deepforestsci.com

@rbhar90



CONTACT US

bharath@deepforestsci.com

🐦 @rbhar90