The Promise of Foundation Models and Generative AI

Payel Das
Principal Researcher and Master Inventor
Generative AI Research Lead
Manager, Trustworthy AI
IBM Research
daspa@us.ibm.com
payel791@
https://www.linkedin.com/in/payeldas/

Foundation models are...

- **Pre-trained** on unlabeled datasets of different modalities (e.g., language, time-series, tabular)
- Leverage **self-supervised learning**
- Learn **generalizable & adaptable data representations** which can be effectively used in **multiple downstream tasks** (e.g., text generation, machine translation, classification for languages)

Note: while transformer architecture is most prevalent in foundation models, definition not restricted by model architecture

In recent years, Large Language Models (LLMs) have taken the field of AI by the storm

Stanford University
Human-Centered
Artificial Intelligence

Language Processing, Machine Learning

How Large Language Models Will Transform Science, Society, and AI

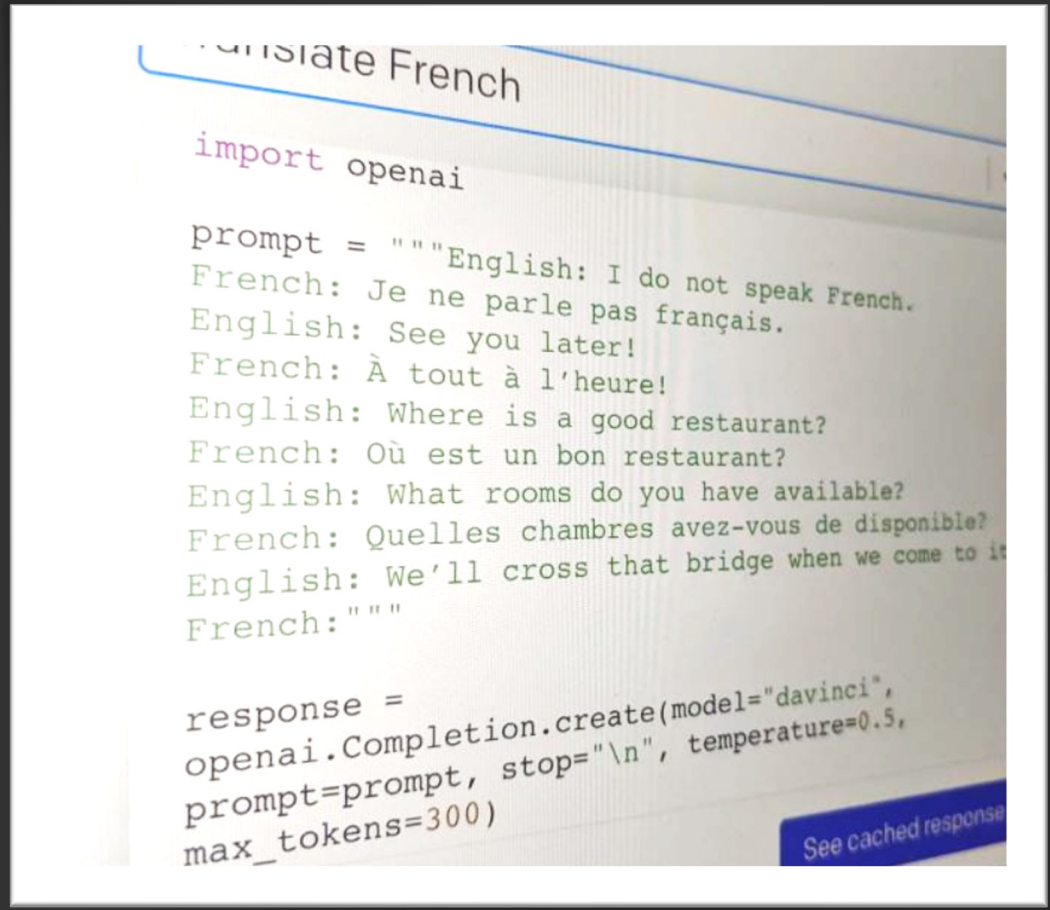
Scholars in computer science, linguistics, and philosophy explore the pains and promises of GPT-3.

Feb 5, 2021 | Alex Tamkin and Deep Ganguli

Hugging Face

Large Language Models: A New Moore's Law?

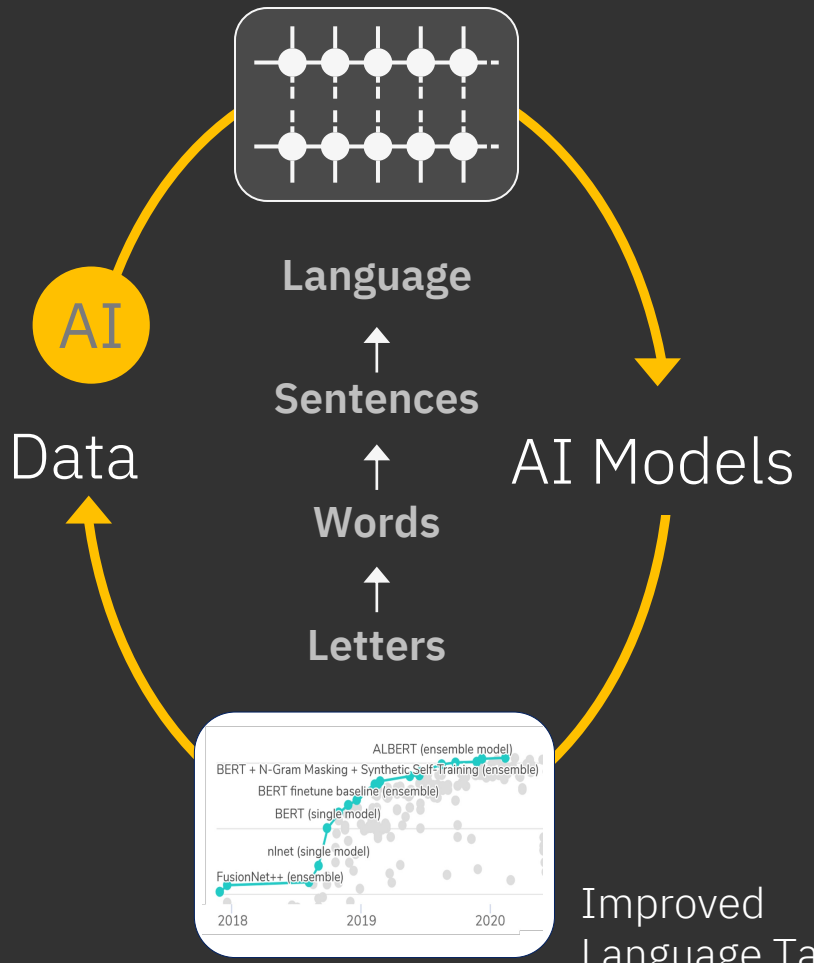
Published October 26, 2021.



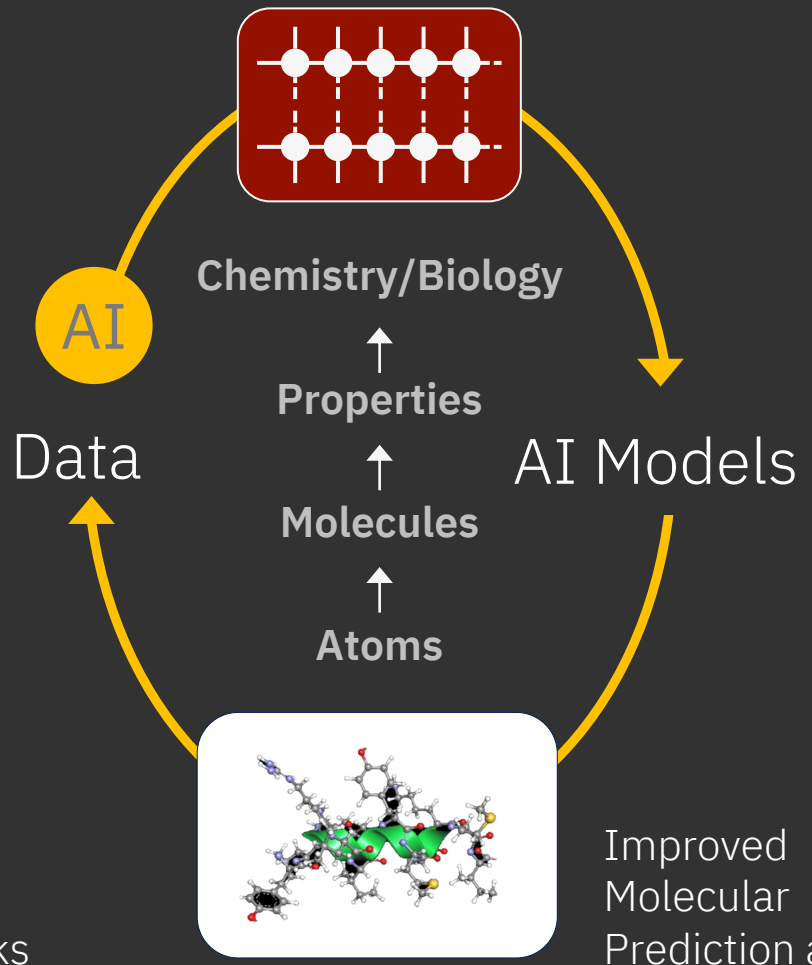
GPT-3 can translate language, write essays, generate code, and more — all with limited to no supervision.

IBM

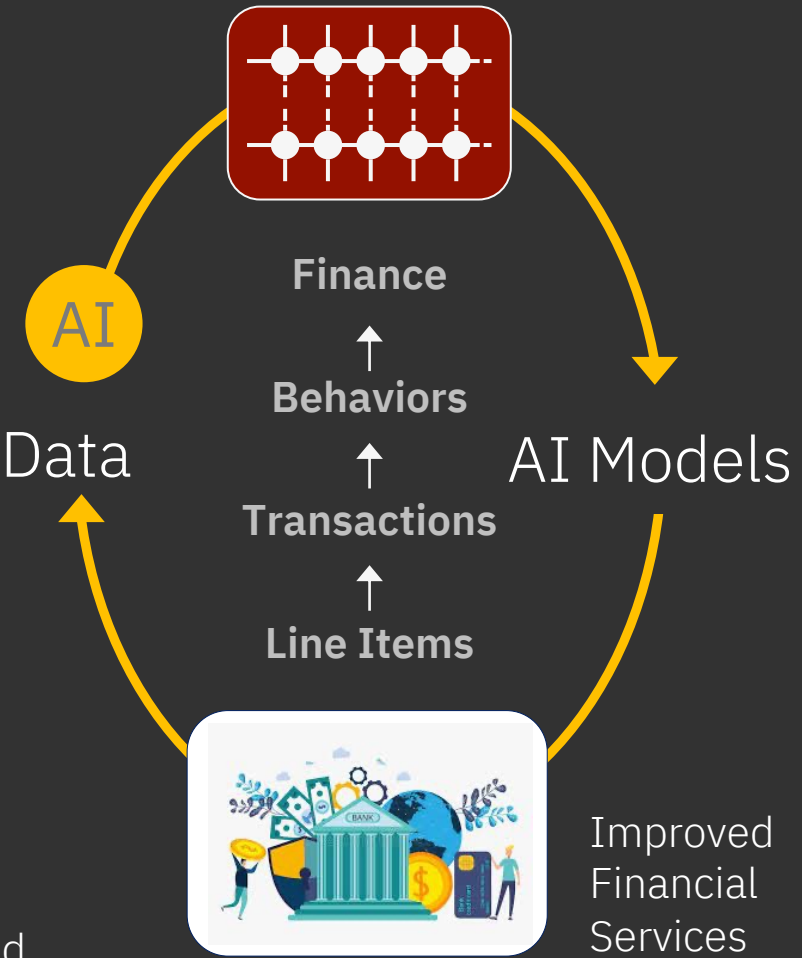
The same AI breakthroughs happening in language are impacting other scientific and enterprise applications



Improved Language Tasks



Improved Molecular Prediction and Design Tasks



Improved Financial Services Tasks

IBM

Can molecular foundation models be useful?

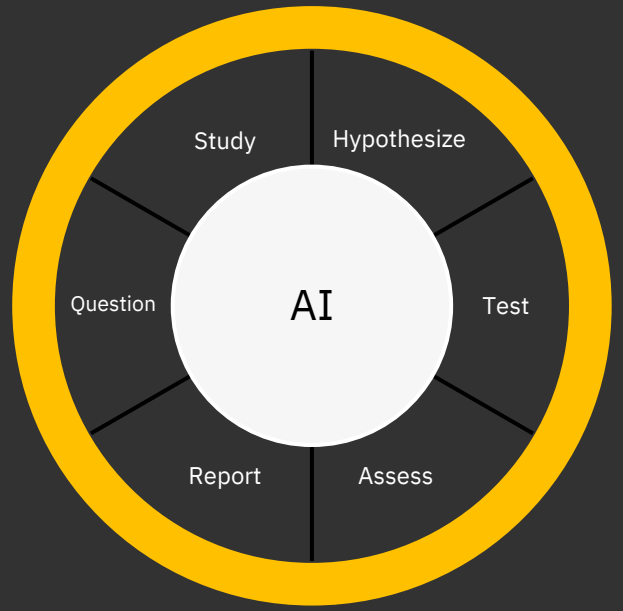
2020 – COVID-19

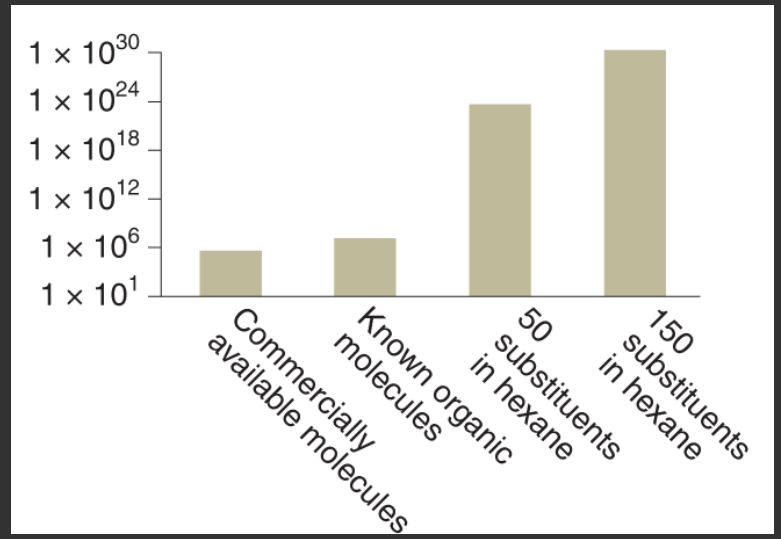
2014 – *Ebola*

2013 – *MERS*

2003 – *SARS*

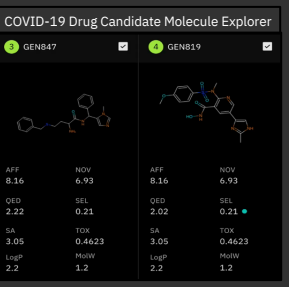
2022 – *Monkey Pox*





The perils of machine learning in designing new chemicals and materials.
Nat Mach Intell 4, 314–315 (2022).

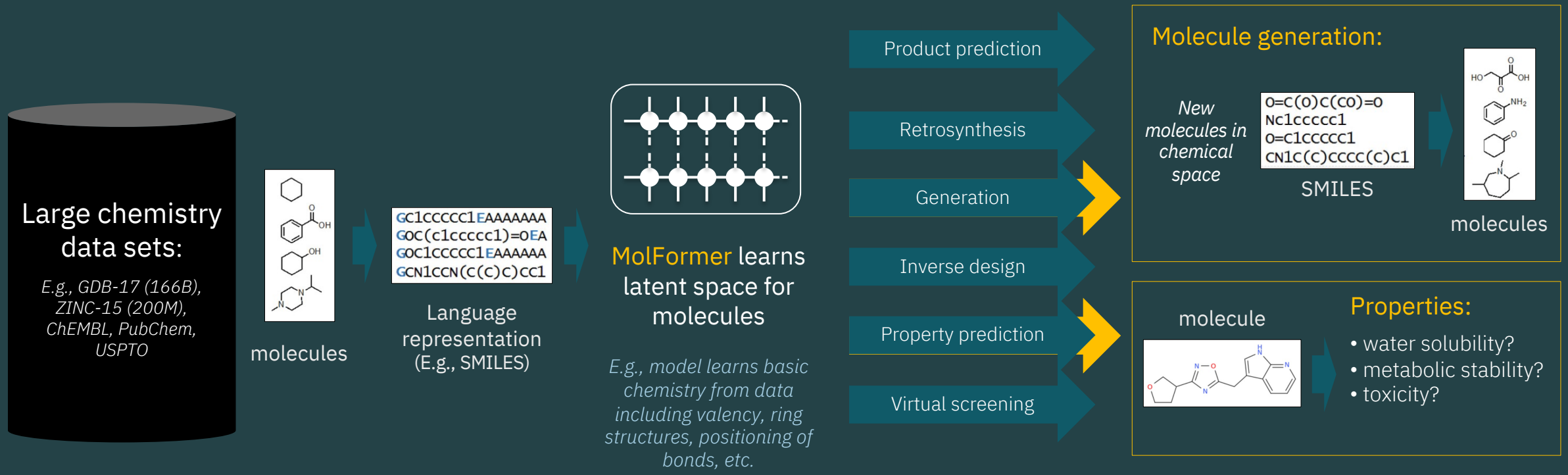
COVID-19 Pandemic Response



Science is the engine that will develop proven therapies

Foundation models for molecules – property prediction and generation

Foundation Models learn the language of chemistry/biology from data and can power up a multitude of discovery tasks – We call them MoLFormer



IBM Research, **CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models.** *NeurIPS*, 2020.
IBM Research, **Accelerating Antimicrobial Discovery with Controllable Deep Generative Models and Molecular Dynamics.** *Nature Biomed. Eng.* 2021.
IBM Research, **Augmenting Molecular Deep Generative Models with Topological Data Analysis Representations**. *ICASSP* 2022.
IBM Research, **Optimizing Molecules using Efficient Queries from Property Evaluations.** Nature Machine Intelligence 2021.
IBM Research, **Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design**. ICML 2021.
IBM Research, **Data-Efficient Graph Grammar Learning for Molecular Generation**. ICLR 2022.
IBM Research, **Biological Sequence Design with GFlowNets**. ICML 2022.
IBM Research, **Active learning of deep surrogates for PDEs...** npj Comp Mat 2020.
IBM Research, Protein Representation Learning by Geometric Structure Pretraining. ICLR 2023.

IBM

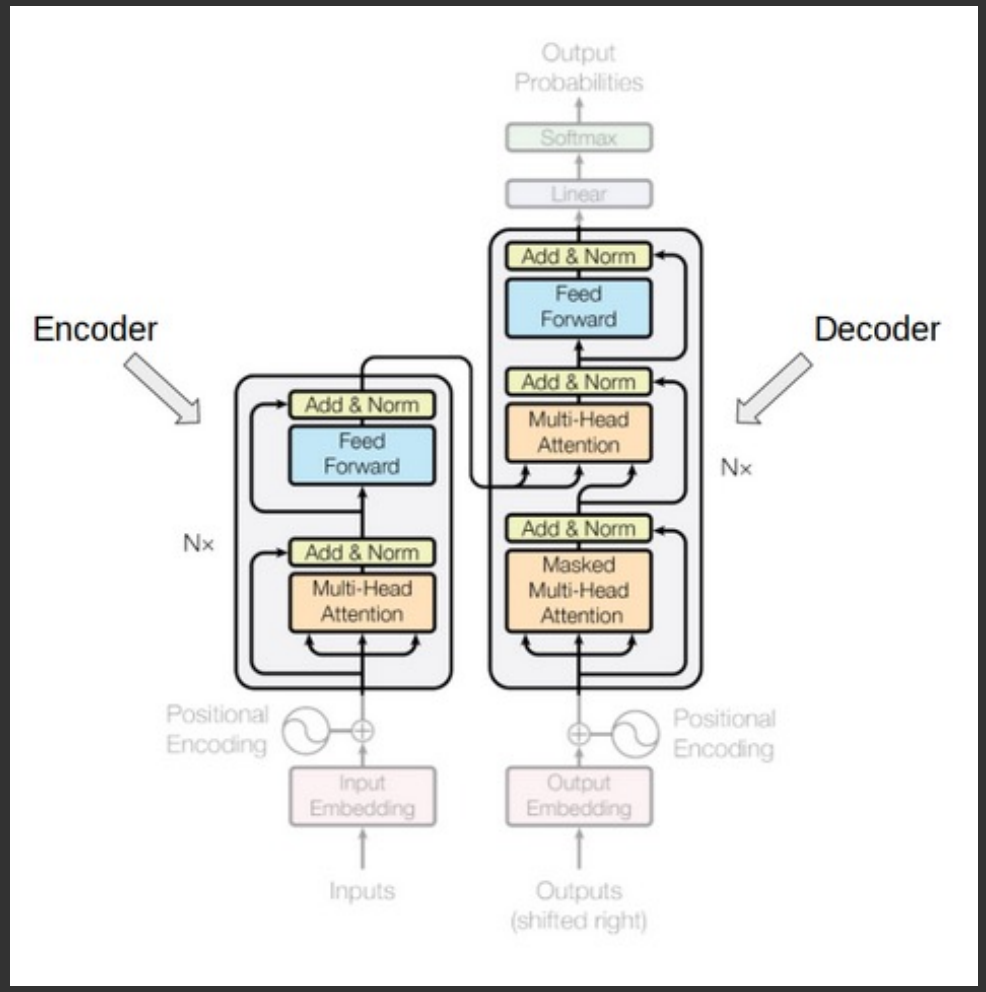
MolFormer: Foundational transformer for chemistry/biology applications

MoLFormer-XL – a specific example from MoLFormer family

- Trained on up to over **a billion** molecular text strings (SMILES), with relatively limited hardware resources (16 V100 GPUs).

- Scalable and fast to train linear time attention transformers as encoders and decoders
- Relative position embeddings facilitate learning on SMILES

- State-of-the-art, universal chemical language model for wide ranges of **70+ molecular property prediction**
- Shows **emergent behavior**, such as geometry, taste, etc.



IBM Research, **Large Scale Molecular Language Representations Capture Important Structural Information.** Nat Mach Intell 4, 1256–1264 (2022).

IBM

Molformer performs comparabiliy than existing GNNs and language models on quantum chemical property regression of QM9 benchmark

	Graph-Based			Geometry-Based			SMILES-Based	
Measure	A-FP	123-gnn	GC	CM	DTNN	MPNN	MoLFormer-XL	ChemBERTa
α	0.492	**0.27**	1.37	0.85	0.95	0.89	**0.3327**	0.8510
C_v	0.252	**0.0944**	0.65	0.39	0.27	0.42	**0.1447**	0.4234
G	0.893	**0.0469**	3.41	2.27	2.43	2.02	**0.3362**	4.1295
gap	0.00528	**0.0048**	0.01126	0.0086	0.0112	0.0066	**0.0038**	0.0052
H	0.893	**0.0419**	3.41	2.27	2.43	2.02	**0.2522**	4.0853
ϵ_{homo}	0.00358	**0.00337**	0.00716	0.00506	0.0038	0.00541	**0.0029**	0.0044
ϵ_{lumo}	0.00415	**0.00351**	0.00921	0.00645	0.0051	0.00623	**0.0027**	0.0041
μ	0.451	0.476	0.583	0.519	**0.244**	**0.358**	0.3616	0.4659
$\langle R^2 \rangle$	26.839	22.90	35.97	46.00	**17.00**	28.5	**17.0620**	86.150
U_0	0.898	**0.0427**	3.41	2.27	2.43	2.05	**0.3211**	3.9811
U	0.893	**0.111**	3.41	2.27	2.43	2.00	**0.2522**	4.3768
ZPVE	0.00207	**0.00019**	0.00299	0.00207	0.0017	0.00216	**0.0003**	0.0023
Avg MAE	2.6355	**1.9995**	4.3536	4.7384	2.3504	3.1898	**1.5894**	8.7067
Avg std MAE	0.0854	**0.0658**	0.1683	0.1281	0.1008	0.1108	**0.0567**	0.1413

IBM

Comparison of MoLFormer with existing baselines on classification and regression benchmarks

Dataset	BBBP	Tox21	ClinTox	HIV	BACE	SIDER
Tasks	1	12	2	1	1	27
RF	71.4	76.9	71.3	78.1	**86.7**	**68.4**
SVM	72.9	**81.8**	66.9	**79.2**	86.2	68.2
MGCN [56]	**85.0**	70.7	63.4	73.8	73.4	55.2
D-MPNN [57]	71.2	68.9	**90.5**	75.0	85.3	63.2
Hu, et al. [58]	70.8	78.7	78.9	80.2	85.9	65.2
N-Gram [44]	91.2	76.9	85.5	**83.0**	87.6	63.2
MolCLR [24]	73.6	79.8	93.2	80.6	**89.0**	68.0
MoLFormer-XL	**93.7**	**84.7**	**94.8**	82.2	88.21	**69.0**

Dataset	QM9	QM8	ESOL	FreeSolv	Lipophilicity
GC	4.3536	0.0148	0.970	1.40	0.655
A-FP	2.6355	0.0282	0.5030	0.736	0.578
MPNN	3.1898	0.0143	0.58	1.150	0.7190
MoLFormer-XL	**1.5894**	**0.0102**	**0.2787**	**0.2308**	**0.5289**

DEMONSTRATIONS

Real time inference from MoLFormer-XL

https://multimodal-ai-demo.bx.cloud9.ibm.com/

IBM Research

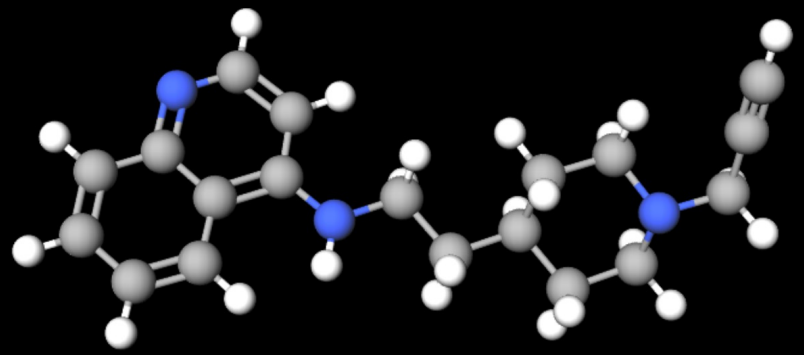
IBM Research Molecular Explorer

Cloud Based Real Time Molecular Screening Platform with MolFormer

To help researchers virtually navigate the chemical space and screen molecules of interest, here we present a cloud-based real-time platform enabled by our large-scale chemical language model, MolFormer.

The platform leverages molecular embedding inferred from MolFormer and retrieves nearest neighbors from PubChem for a list of input chemicals. To assist with automating chemistry, drug discovery and material design tasks, we also show in the platform the molecular attributes of the retrieved nearest neighbors as metadata, such as physicochemical properties (estimated using RDKIT), bioactivities (Enamine BioActivity), odor (Olfactionbase), and ease of synthesis (Enamine Real).

Results are for research use only.

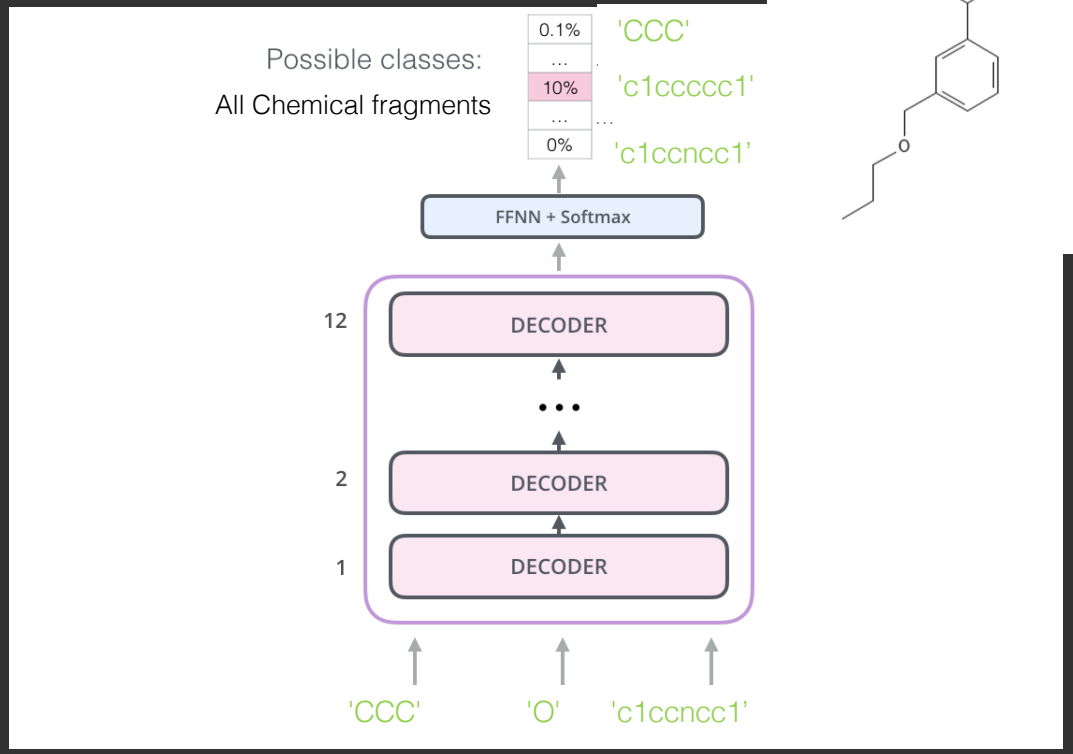


IBM Research, **Cloud-Based Real-Time Molecular Screening Platform with MolFormer .** ECML PKDD (2022). https://2022.ecmlpkdd.org/wp-content/uploads/2022/09/sub_1455.pdf

MolGPT: Foundational transformer for molecule generation

- A large-scale and efficient molecular language generator
- Efficient training on over **a billion** molecular text strings (SMILES)

- Scalable and fast to train linear time attention transformers
- Distributed training using Pytorch Lightning
- Enjoys fast inference due to operating on text



	Validity	Uniqueness	Novelty
MolGPT (ours)	0.95	0.99	0.99
MegaMolBert (Nvidia)	0.75	0.85	0.51

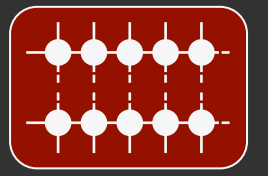
IBM Research, **To be submitted.**

IBM

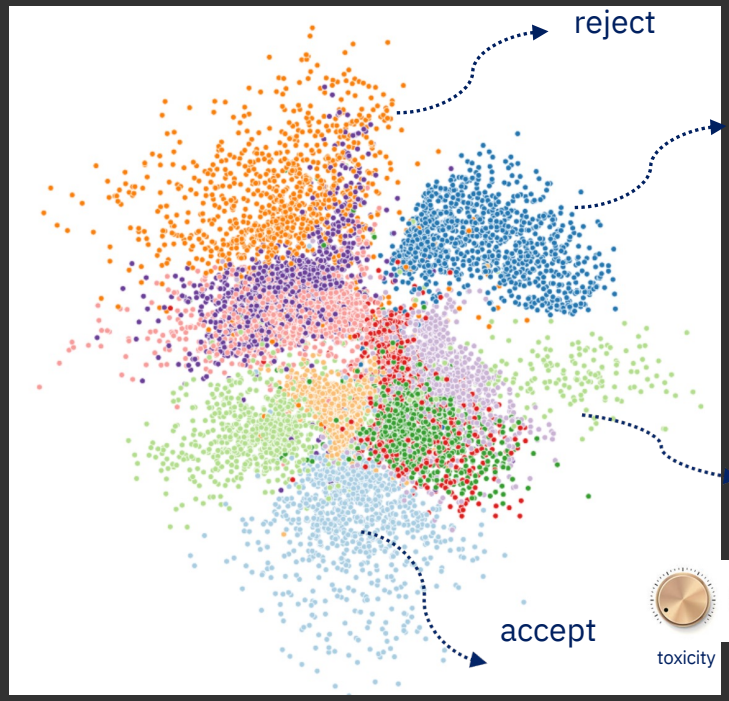
Large-scale unsupervised pretraining, novel sampling, and optimization methods enable **controllable generation of novel artifacts with desired properties**

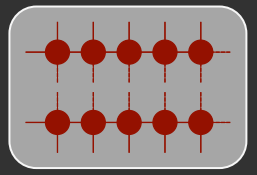
Large chemistry data sets:

E.g., GDB-17 (166B), ZINC-15 (200M), ChEMBL, PubChem, USPTO

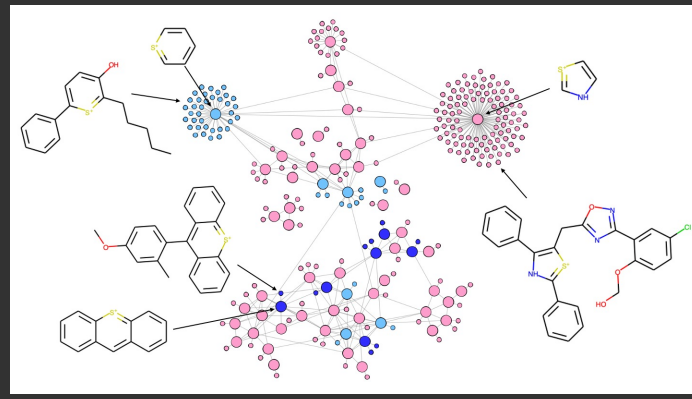


encoder





decoder



generated molecules

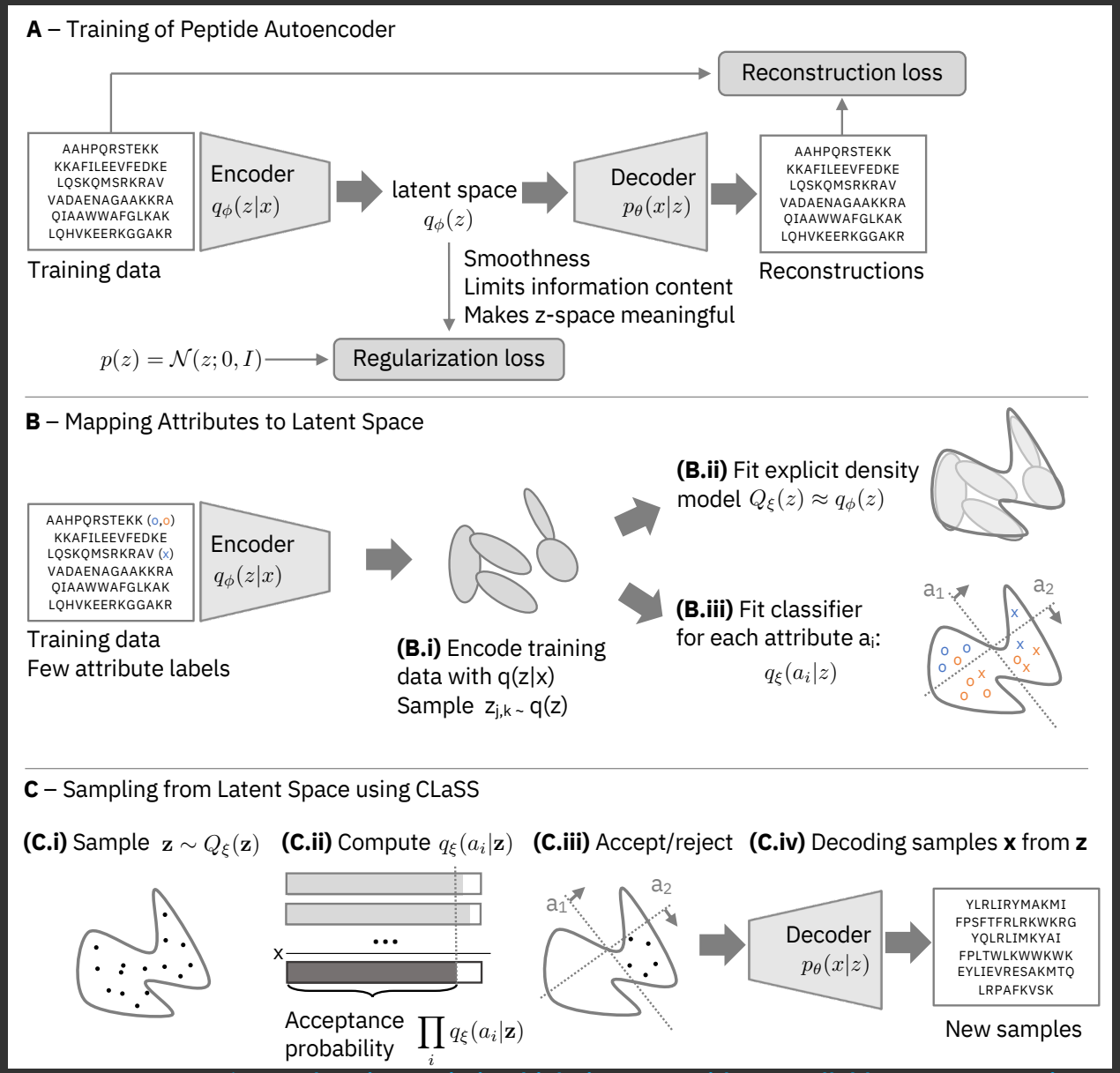
Molformer models learn latent space of molecules that preserves important structural information and learns similarity between datapoints

+

Highly-controlled sampling of the latent space

results in valid generations, provides access to novel chemical scaffolds as per expert adjudication

Conditional Latent (attribute) Space Sampling -CLaSS



Adding Property Controls On A Generative Foundation Model

$$p(\mathbf{x}|\mathbf{a}) = \int \mathrm{d}z\, p(\mathbf{z}|\mathbf{a}) p(\mathbf{x}|\mathbf{z})$$

$$p(\mathbf{z}|\mathbf{a}) = \frac{p(\mathbf{a}|\mathbf{z}) q_\phi(\mathbf{z})}{p(\mathbf{a})}$$

$$= \frac{q_\phi(\mathbf{z}) \prod_i p(a_i|\mathbf{z})}{p(\mathbf{a})}$$

IBM Research, **Accelerating Antimicrobial Discovery with Controllable Deep Generative Models and Molecular Dynamics.** *Nature Biomed. Eng.*, March 2021.

IBM

Generative AI for Molecular Generation (COVID-19)

https://covid19-mol.mybluemix.net/

Explore novel drug candidates for COVID-19

To help researchers generate potential new drug candidates for COVID-19, we have applied our novel AI generative frameworks to three COVID-19 targets and have generated 3000 novel molecules. We are sharing these molecules under an Creative Commons here.

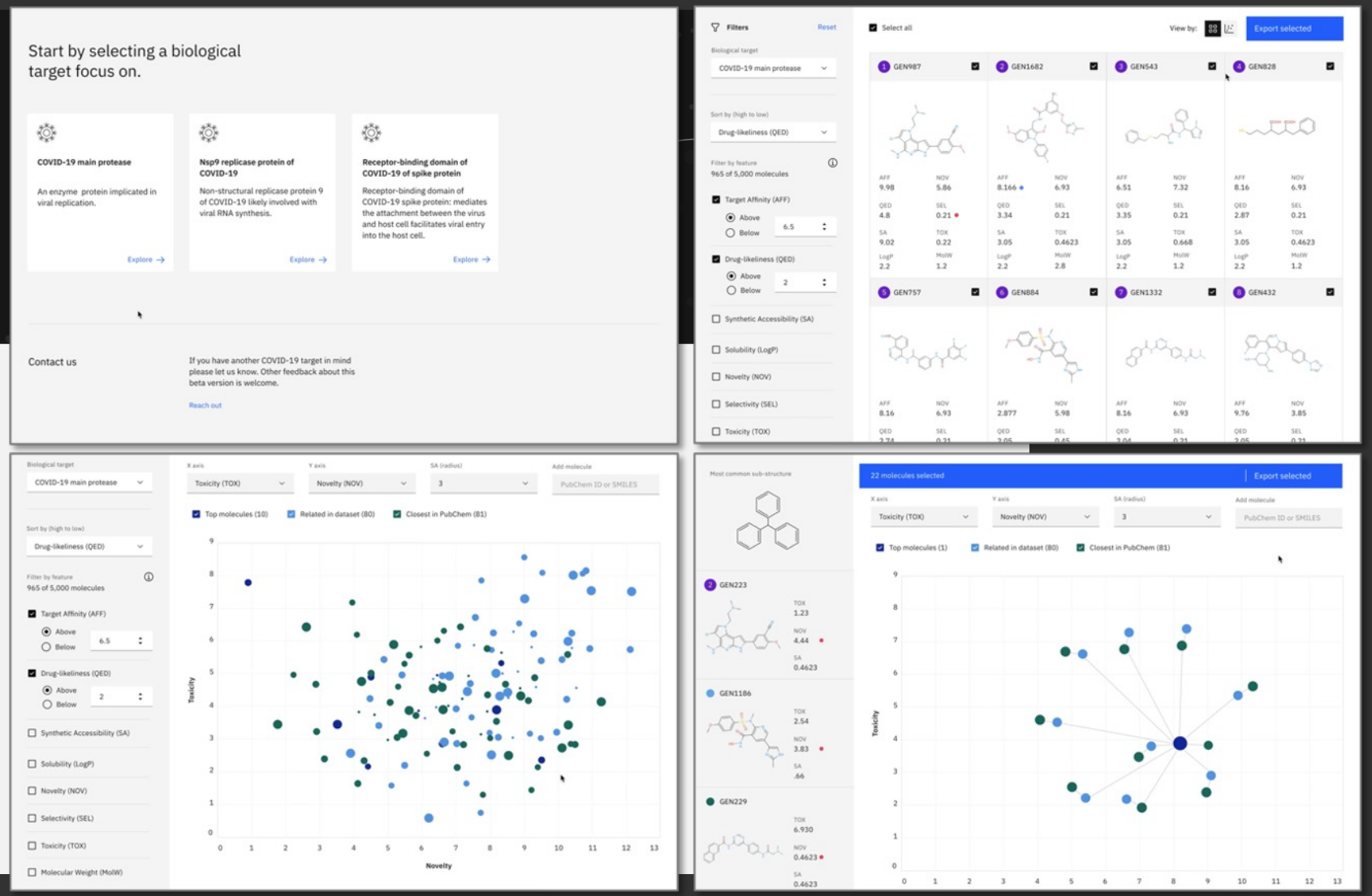
Launch exploration

How it works

1 / 4

Select a biological target and filter generated molecules by important characteristics

Out of 1,000 AI-generated potential candidates for a specific COVID-19 target, the top 10 candidates are selected and ranked using a specific filtering criterion on an attribute of interest. The molecules are displayed along with a list of attributes. These top candidates can be further filtered and sorted with additional criteria to select the most promising molecules.

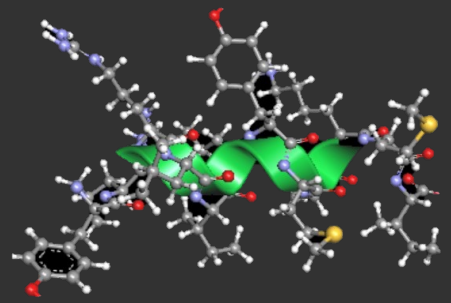


IBM Research, **CogMol: Target-Specific and Selective Drug Design for COVID-19 Using Deep Generative Models**. NeurIPS 2020.
Recognized by Harvard Belfer Center Technology and Public Purpose (TAPP) Project 2021.

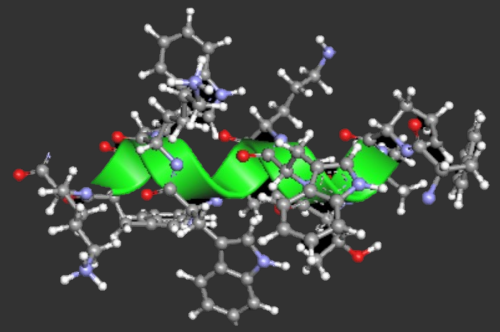
IBM

Evaluation of our foundation models for chemistry and biology have resulted in groundbreaking molecular discovery

(A) Two novel AI-designed antimicrobials with high broad-spectrum potency, low toxicity, and low resistance onset, validated in wet lab.

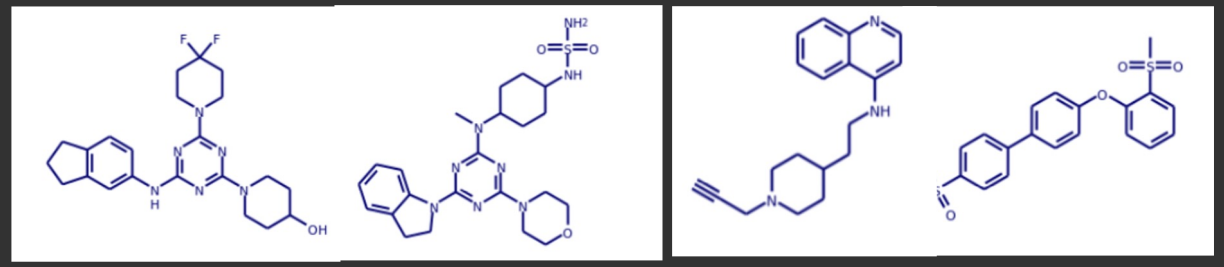


YLRLIRYMAKMI-CONH2
(YI12, 12 amino acids)



FPLTWLKWWKWKK-CONH2
(FK13, 13 amino acids)

(B) Four novel drug-like inhibitor molecules against two distinct SARS-CoV-2 targets, the main protease (Mpro) and the receptor binding domain (RBD) of the spike protein.



GXA70 GXA112 GEN727 GEN725

Mpro Spike RBD

4-6 weeks and 10-50% success rate with generative AI, compared to 2-4 years and <1% success rate with existing methods.

IBM Research, **Accelerating Antimicrobial Discovery with Controllable Deep Generative Models and Molecular Dynamics.**
Nature Biomed. Eng., March 2021.

IBM Research, **Accelerating Inhibitor Discovery for Multiple SARS-CoV-2 Targets with a Single, Sequence-Guided Deep Generative Framework.**
(Under Review)

IBM

Emergent behavior of FMs due to data and neural scaling

Emergent Behavior in Foundation Models :
Case study – MoLFormer-XL

Dataset	BBBP	HIV	BACE	SIDER	Clintox	Tox21
10% ZINC + 10% PubChem	91.5	81.3	86.6	68.9	94.6	84.5
10% ZINC + 100% PubChem	92.2	79.2	86.3	69.0	94.7	84.5
100% ZINC	89.9	78.4	87.7	66.8	82.2	83.2
MoLFormer-Base	90.9	77.7	82.8	64.8	61.3	43.2
MoLFormer-XL	**93.7**	**82.2**	**88.2**	**69.0**	**94.8**	**84.7**

Dataset	QM9	QM8	ESOL	FreeSolv	Lipophilicity
10% Zinc + 10% Pub	1.7754	0.0108	0.3295	0.2221	0.5472
10% Zinc + 100% Pub	1.9093	**0.0102**	**0.2775**	**0.2050**	0.5331
100% Zinc	1.9403	0.0124	0.3023	0.2981	0.5440
MoLFormer-Base	2.2500	0.0111	0.2798	0.2596	0.6492
MoLFormer-XL	**1.5984**	**0.0102**	0.2787	0.2308	**0.5298**

IBM Research, **Large Scale Molecular Language Representations Capture Important Structural Information.** Nat Mach Intell 4, 1256–1264 (2022).

IBM

MoLFormer appears compatible to geometric GNNs or better

QM9 Task	SchNet[41]	DimeNet[37]	MOLFORMER-XL
U_0_atom	0.0140	**0.0080**	0.0827
U_atom	0.0190	**0.0079**	0.0974
H_atom	0.0140	**0.0081**	0.0947
G_atom	0.0140	**0.0089**	0.0888

Task	DimeNet[37]	GeomGCL[36]	GEM[38]	MOLFORMER-XL
ESOL (RMSE)	0.633	0.575	0.798	**0.2787**
FreeSolv (RMSE)	0.978	0.866	1.877	**0.2308**
Lipophilicity (RMSE)	0.614	0.541	0.660	**0.5289**

IBM Research, **Large Scale Molecular Language Representations Capture Important Structural Information.** Nat Mach Intell 4, 1256–1264 (2022).

IBM

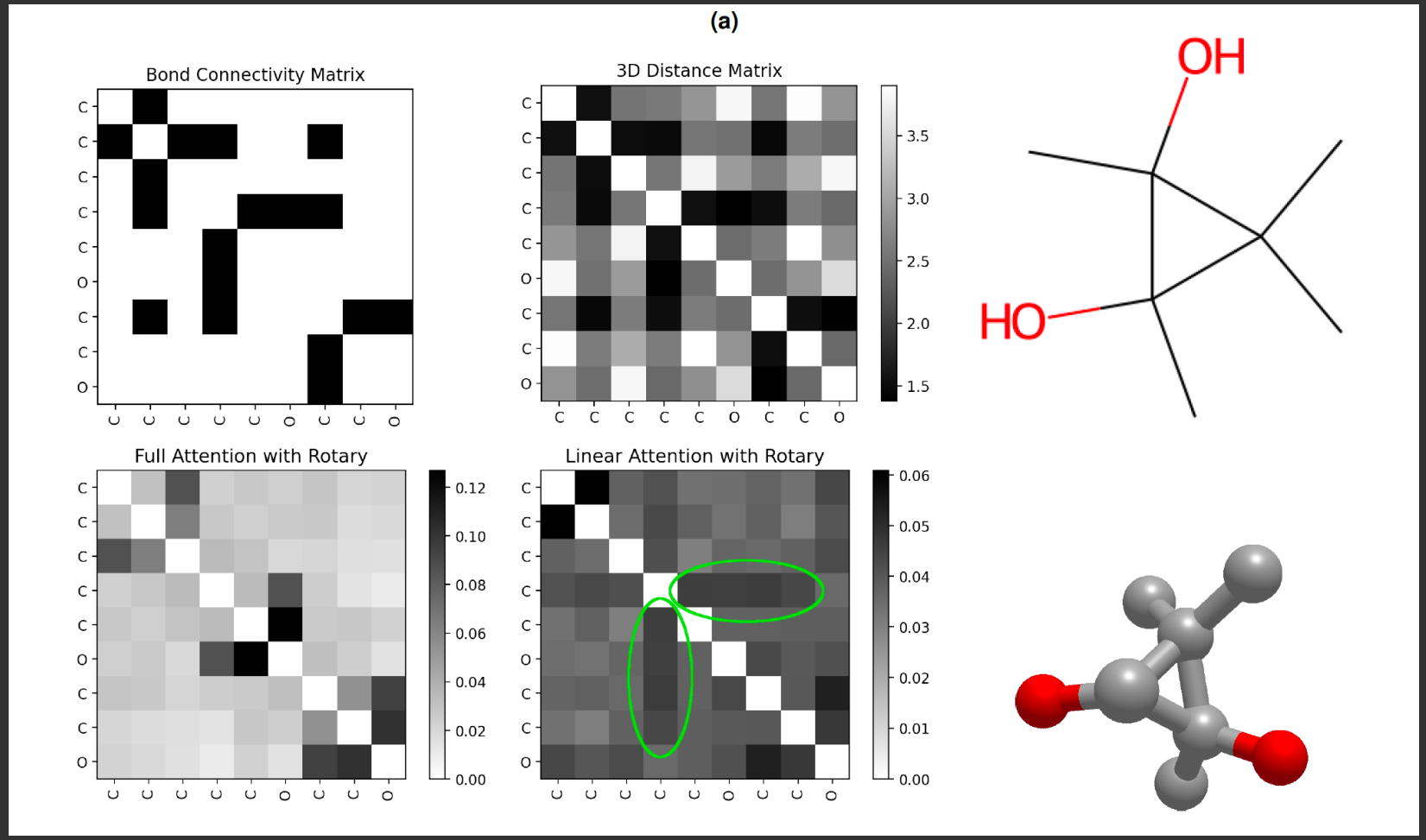
Through lens of MoLFormer attention visualization – correlation with spatial distances

Distance-Category	Attention	1	3	5	7	9	11
Short	Full (✓ Rotary)	**0.615**	0.604	0.603	**0.615**	0.601	**0.598**
	Linear (✓ Rotary)	0.596	0.597	0.602	0.597	0.600	0.594
Medium	Full (✓ Rotary)	0.716	0.724	0.724	0.716	**0.727**	0.727
	Linear (✓ Rotary)	**0.729**	**0.728**	**0.724**	**0.727**	0.726	**0.730**
Long	Full (✓ Rotary)	0.204	0.207	0.208	0.205	0.208	0.210
	Linear (✓ Rotary)	**0.211**	**0.210**	**0.210**	**0.211**	**0.209**	**0.210**

IBM Research, **Large Scale Molecular Language Representations Capture Important Structural Information.** Nat Mach Intell 4, 1256–1264 (2022).

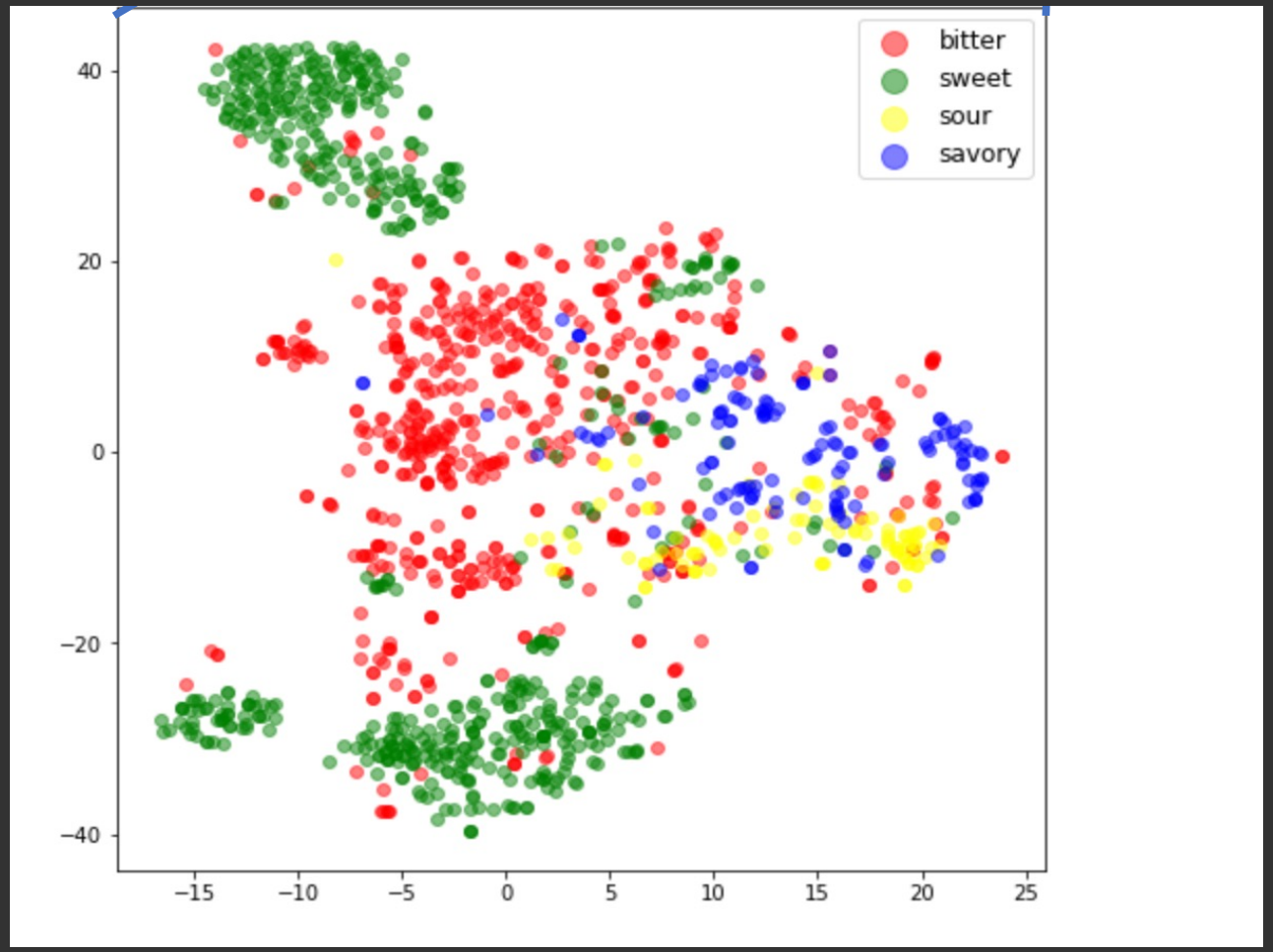
IBM

Molformer indeed captures sufficient structural information



IBM Research, **Large Scale Molecular Language Representations Capture Important Structural Information.** Nat Mach Intell 4, 1256–1264 (2022).

MoLFormer Learns molecular taste without labels

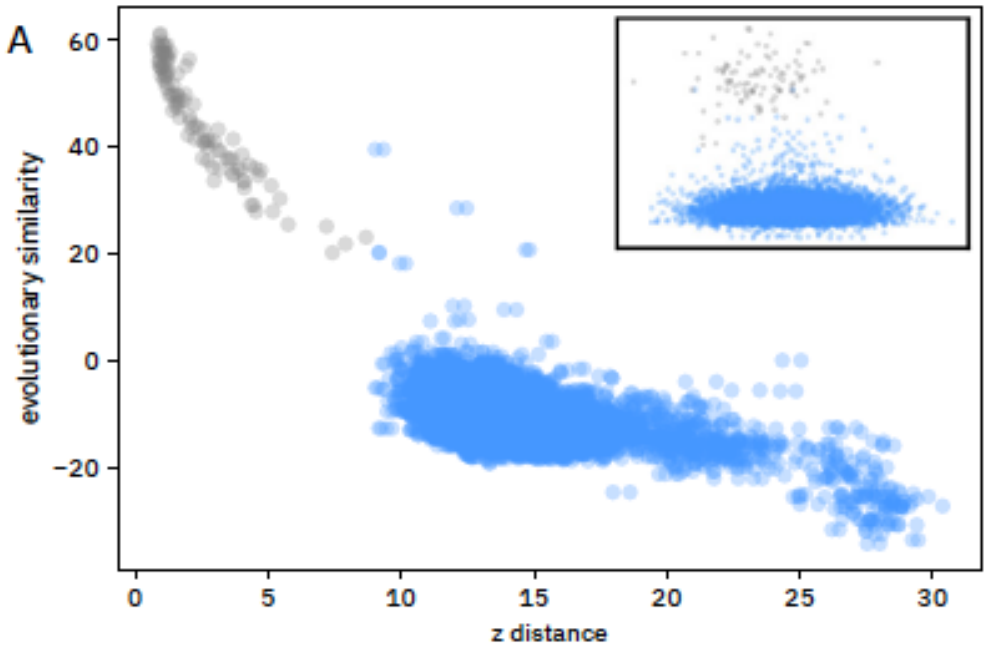


IBM Research, **Cloud-Based Real-Time Molecular Screening Platform with MolFormer .** ECML PKDD (2022). https://2022.ecmlpkdd.org/wp-content/uploads/2022/09/sub_1455.pdf

IBM

Emergent Behavior in Foundation Models :
Case study: Peptide Generative AE

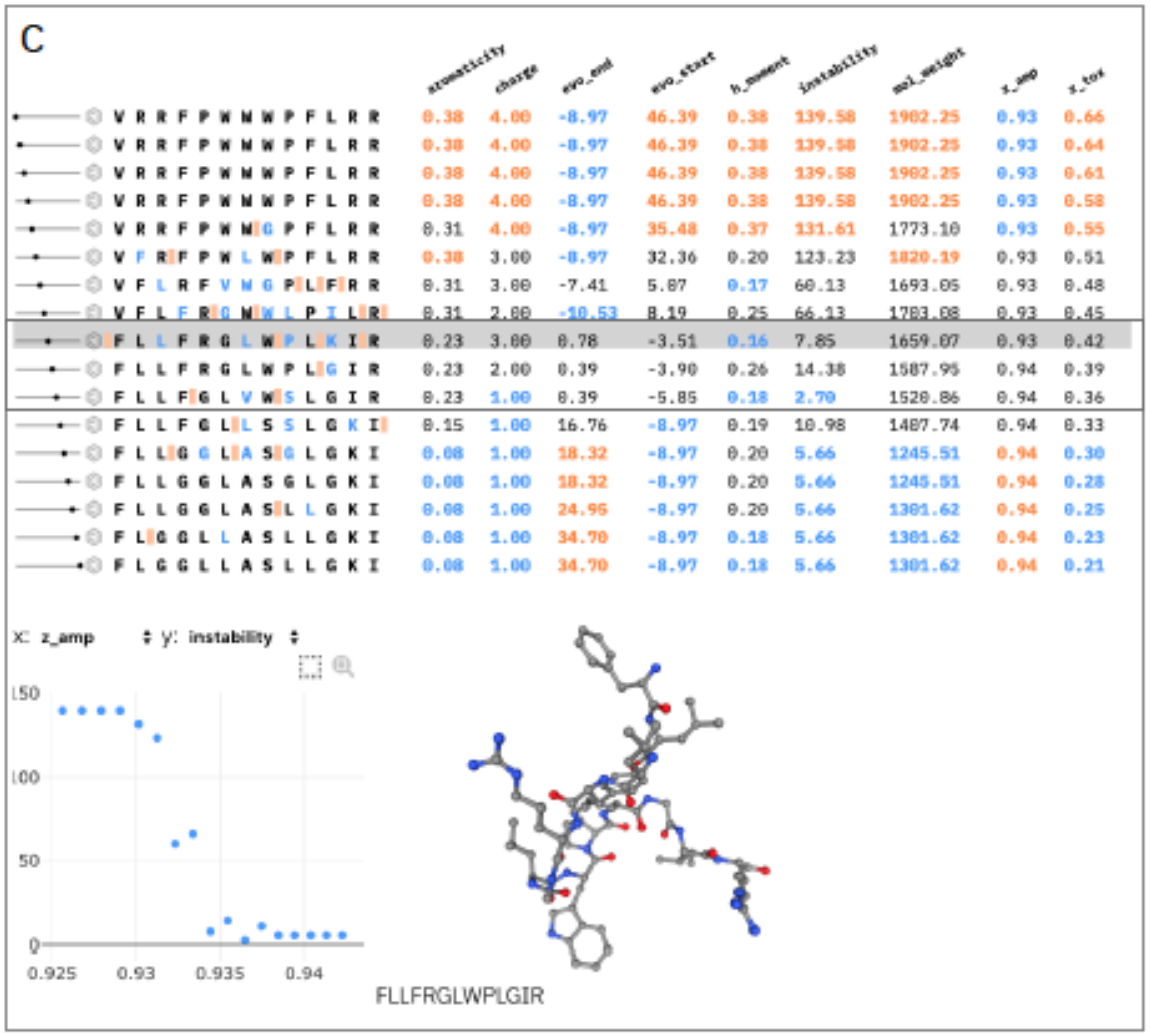
A



B

Classifier	AMP	Toxic
WAE z-classifier	87.4	68.9
Sequence classifier	88.0	93.7

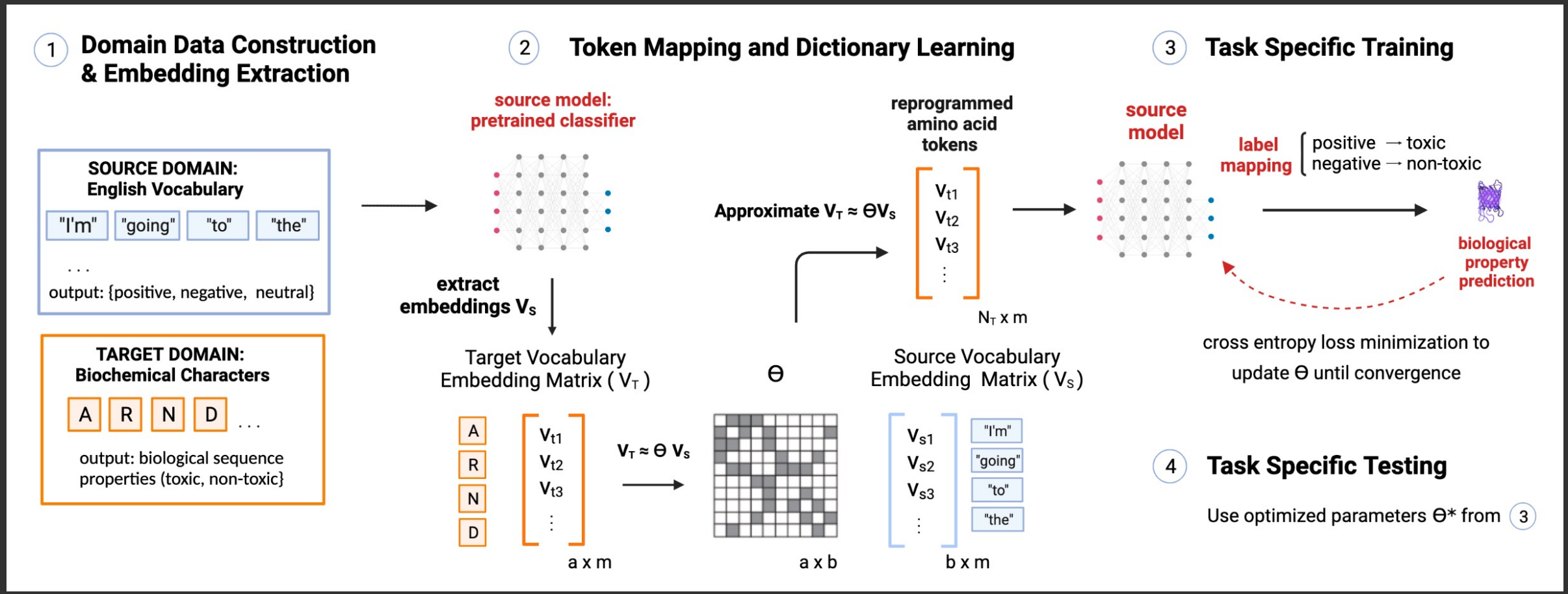
C



IBM Research, **Accelerating Antimicrobial Discovery with Controllable Deep Generative Models and Molecular Dynamics.** *Nature Biomed. Eng.*, March 2021.

IBM

Emergent behavior leads to domain-specific model reprogramming
From natural language to biology

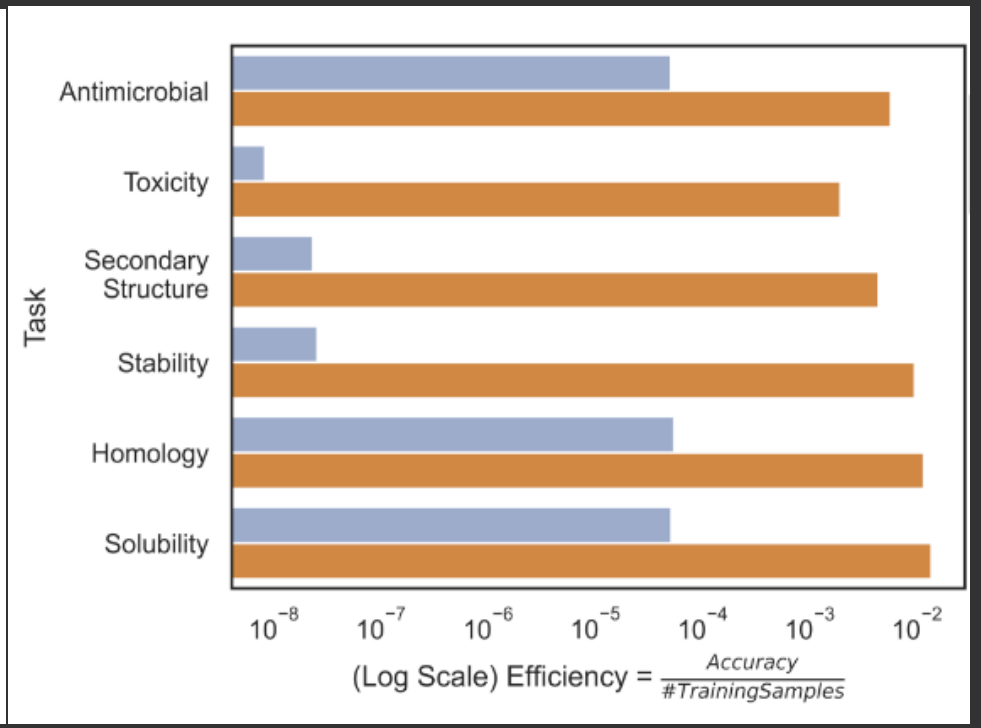


IBM Research, Reprogramming Pretrained Language Models for Protein Sequence Representation Learning. Under review.

IBM

Reprogrammed FMs are accurate, data-efficient, and robust

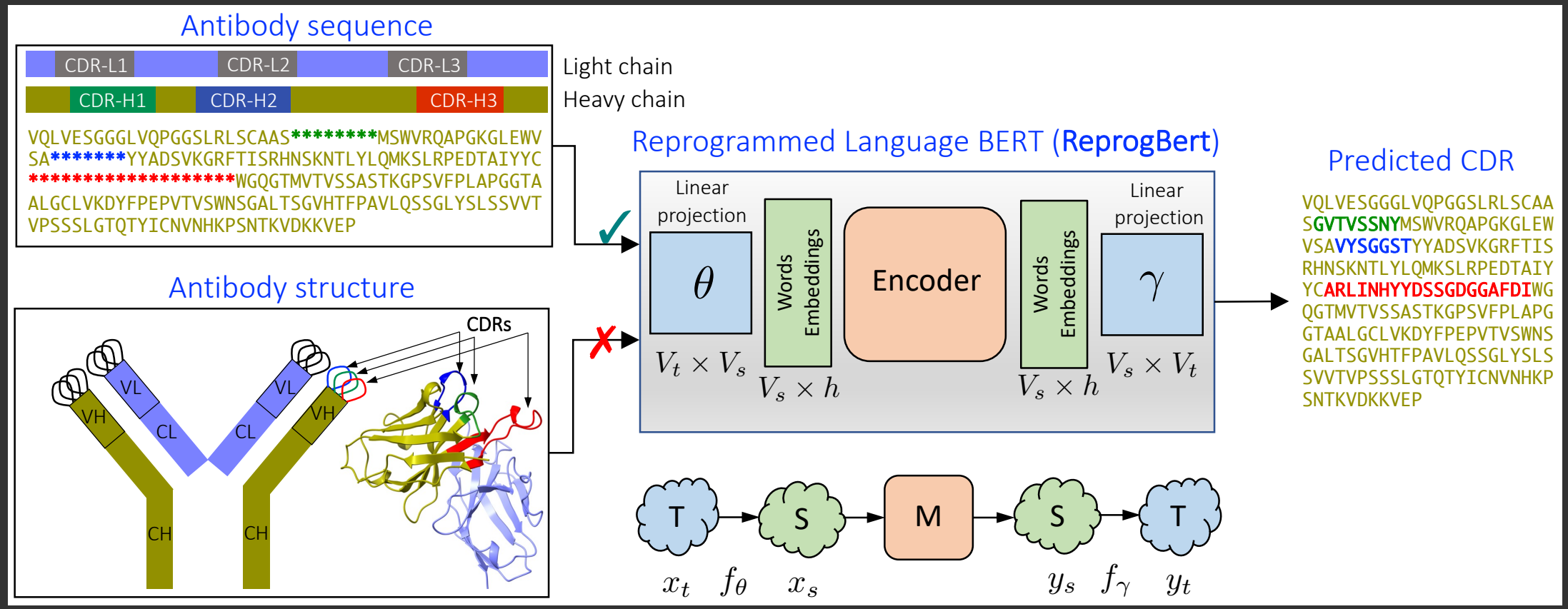
Protein Downstream Task	R2DL			Pretraining			Supervised		
	Training Samples	Accuracy	Efficiency	Training Samples	Accuracy	Efficiency	Training Samples	Accuracy	Efficiency
Secondary Structure	8678	0.841	9.70E-05	3.10E+07	0.801	2.58E-08	8678	0.623	7.18E-05
Stability	21446	0.849	3.96E-05	3.10E+07	0.738	2.38E-08	21446	0.660	3.08E-05
Homology	12312	0.241	1.96E-05	3.10E+07	0.265	8.56E-09	12312	0.245	1.99E-05
Solubility	16253	0.943	5.80E-05	1.70E+06	0.872	5.13E-07	16253	0.856	5.27E-05
Antibody Affinity	4000	0.9456	2.36E-04	-	-	-	4000	0.928	2.32E-04
Antimicrobial	6489	0.900	1.39E-04	1.70E+06	0.883	5.19E-07	6489	0.874	1.35E-04
Toxicity	8153	0.961	1.18E-04	1.70E+06	0.937	5.51E-07	8153	0.689	8.45E-05



IBM Research, Reprogramming Pretrained Language Models for Protein Sequence Representation Learning. Under review.

IBM

Efficient CDR design via sequence infilling with FM reprogramming



- We introduce additional amino acid embeddings (target domain), together with the linear matrices θ and γ to project from one domain to another.
- During CDR infilling training, only the θ and γ and protein embeddings are fine-tuned, the source English language model remains unmodified.

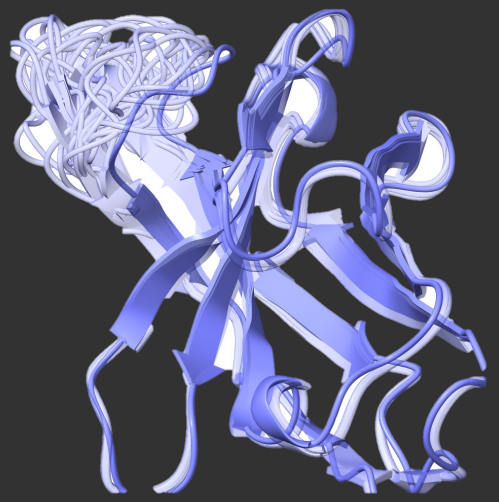
IBM Rsearch, Reprogramming Large Pretrained Language Models for Antibody Sequence Infilling. Under review. arXiv:2210.07144

IBM

Performance on antibody heavy-chain CDR design

	SabDab CDR-H3									
	PPL	PPL-ProGen	RMSD	RMSD-AF	RMSD-IF	TM-AF	TM-IF	AAR	AAR>30%	DIV
LSTM	9.20	–	–	–	–	–	–	–	–	–
AR-GNN	9.44	–	3.63	–	–	–	–	–	–	–
Refine-GNN	8.38	7.2	2.50	5.62	3.43	85.0	94.0	28.2	no	25.7
ProtBert	–	6.8	–	5.40	3.39	85.2	94.0	41.5	yes	14.5
EnglishBert	–	5.9	–	5.53	3.26	84.9	94.0	35.6	yes	59.8
ReprogBert	–	5.4	–	5.54	3.44	85.1	94.0	32.6	yes	67.4

Refine-GNN: Jin, et al, Iterative refinement graph neural network for antibody sequence-structure co-design. ICLR 2022



ReprogBert model upholds structural integrity, sequence recovery, and naturalness.

High novelty and diversity of the generated sequences are achieved.

Can handle multiple CDR infilling at once.

Generated antibodies also show antigen specificity and improved virus neutralization in silico

Lightweight training, while leveraging information from large out-of-domain language pretraining.

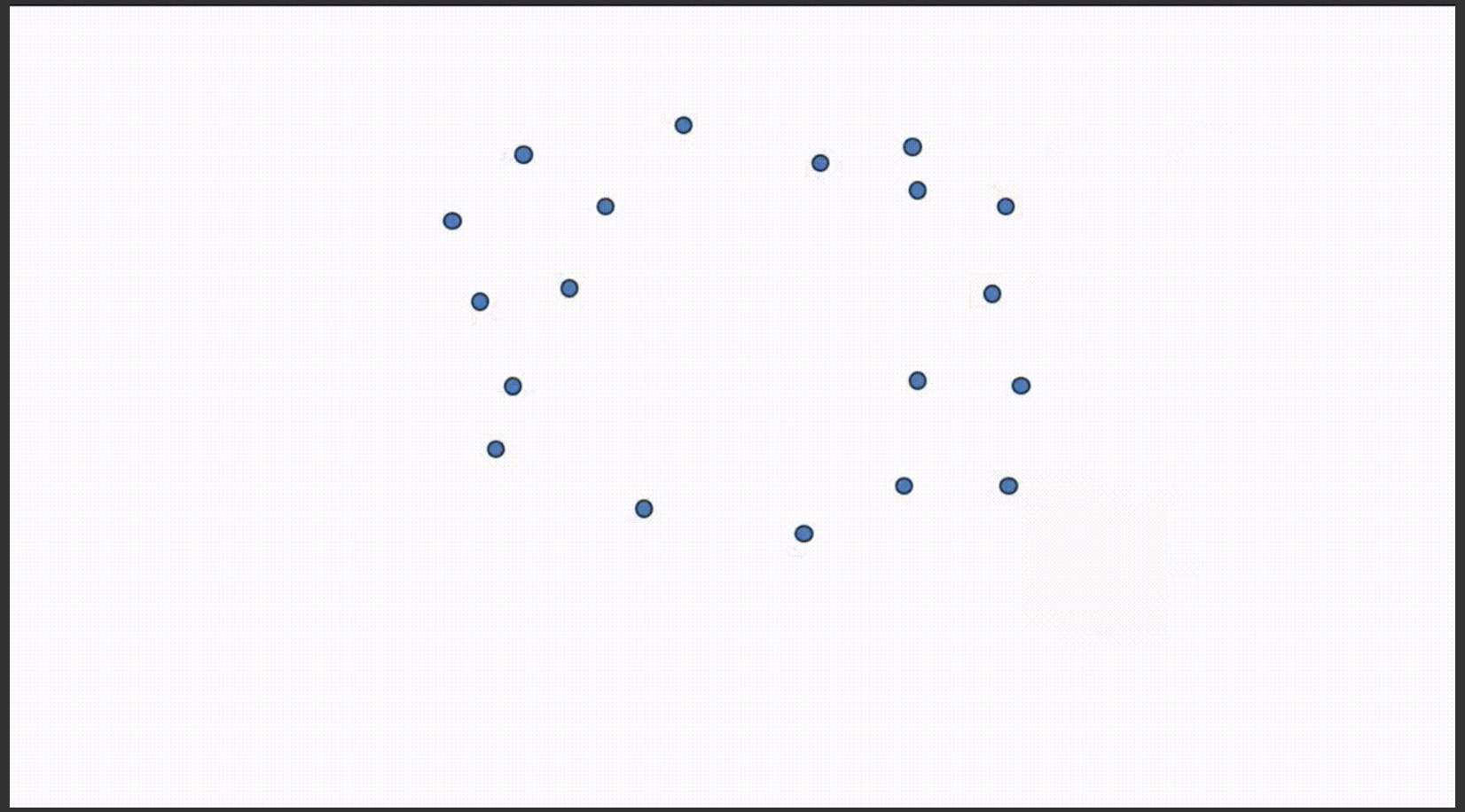
	Neutralization Score	
Model	CoV-AbDab	CoV-AbDab + SabDab
Original	–	69.3
LSTM	–	72.0
AR-GNN	–	70.4
Refine-GNN	–	75.2
ProtBert	72.7	74.7
EnglishBert	70.5	71.0
ReprogBert	75.6	76.7

IBM Rsearch, Reprogramming Large Pretrained Language Models for Antibody Sequence Infilling. Under review. arXiv:2210.07144

How can we explicitly include geometry in molecular FMs?

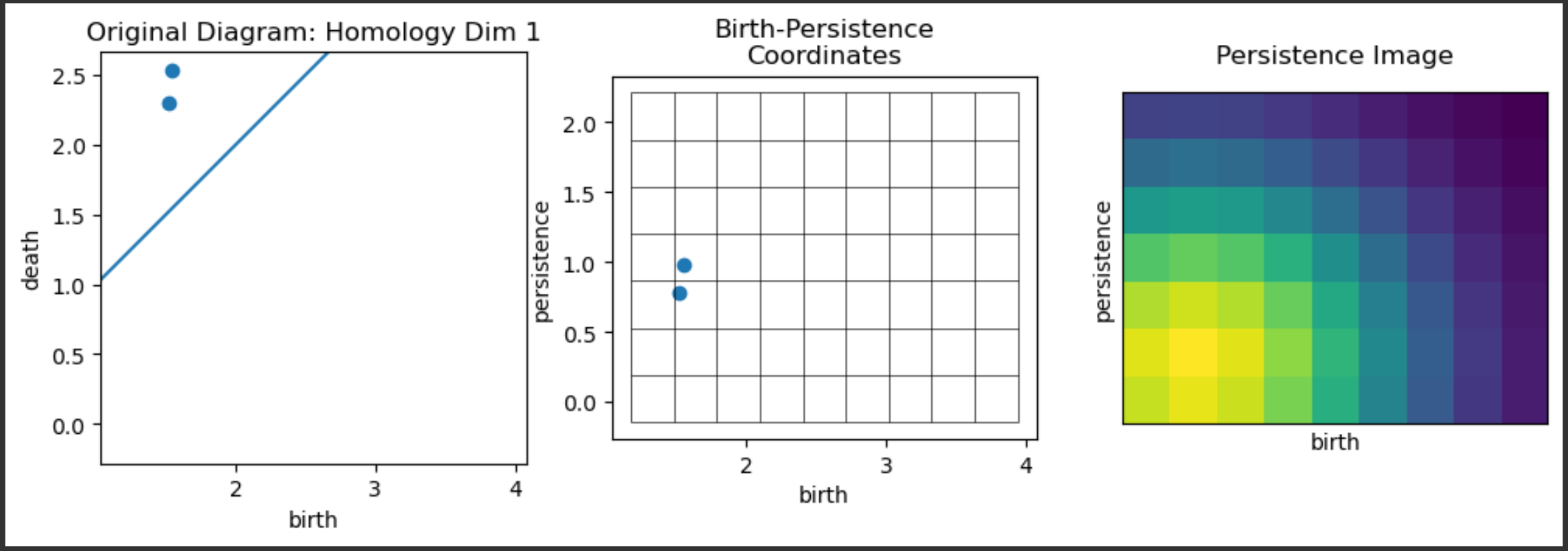
Capturing molecular geometry with topological information

Background: Topological Data Analysis (TDA)

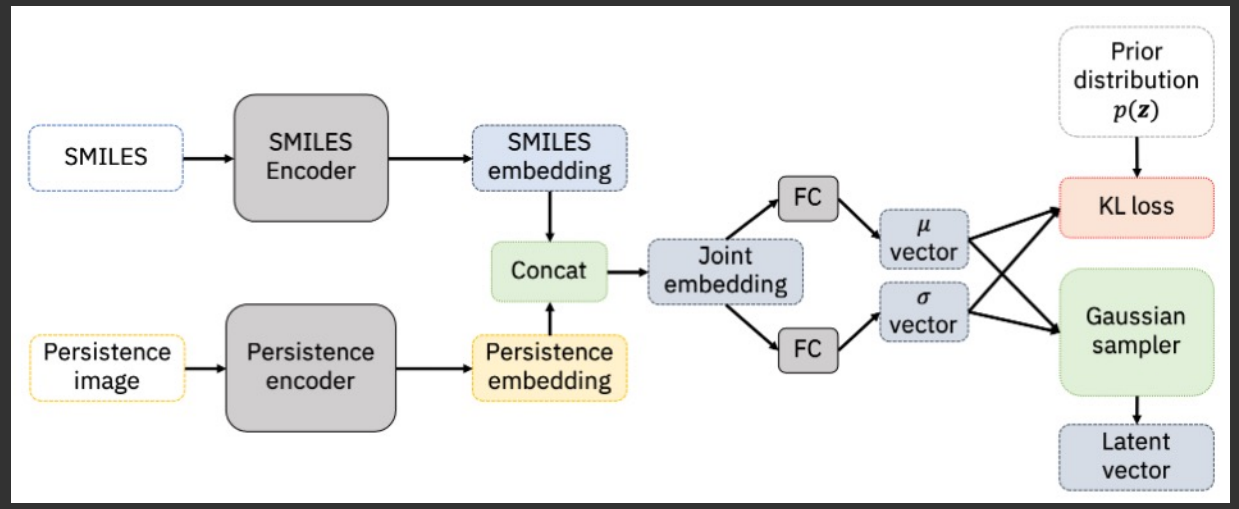


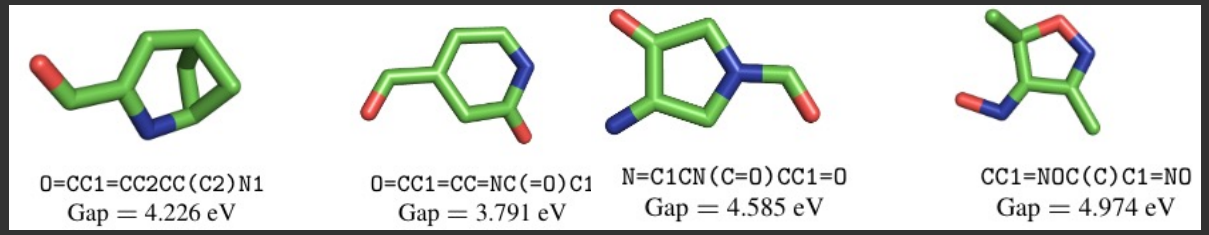
Augmenting molecular generative models with geometric (TDA) information

Background: Persistence images



Augmenting molecular generative models with geometric (TDA) information





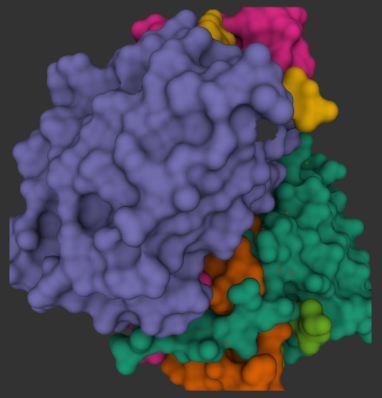
persistent homology information for robust modeling the global geometry -- invariant to translation, rotation, and node label permutation.

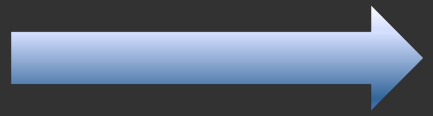
	QM9	SMILES	3D	3D + q	GVAE*	CGVAE†	MPGVAE*	MolGAN*	G-SchNet†
Validity	1.000	0.819	0.840	0.852	0.810	**1.000**	0.91	0.98	0.771
Ring size									
R3	0.470	0.479	0.462	**0.470**	0.560	0.430	0.552	0.385	0.623
R4	0.586	0.490	0.561	**0.582**	0.333	0.692	0.647	0.247	0.657
R5	0.495	0.409	0.482	**0.483**	0.218	0.902	0.526	0.325	0.430
R6	0.158	0.169	0.155	**0.157**	0.110	0.649	0.104	0.115	0.133
Sum	1.709	1.600	**1.731**	1.734	1.222	2.673	1.828	1.072	1.843
χ^2	–	0.003	**0.000**	**0.000**	0.040	0.056	0.005	0.017	0.008

IBM Research, **Augmenting Molecular Deep Generative Models with Topological Data Analysis Representations.** ICASSP 2022.

IBM

Inverse Folding





“How can we find "good" amino acid sequences (i) that fold to a desired "target" structure as a native conformation of lowest accessible free energy and (ii) that will not simultaneously fold to many other conformations of the same free energy?” Yue & Dill, PNAS 1992.

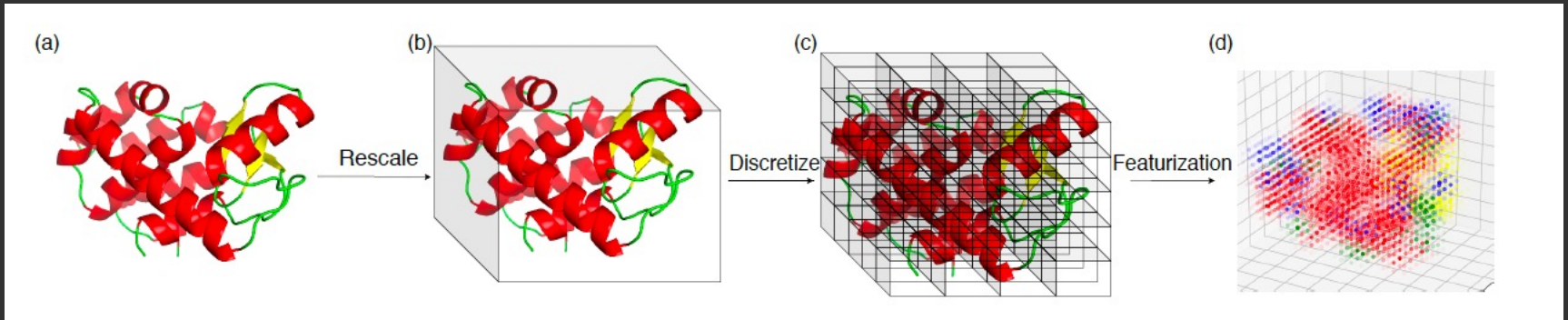
Physics-based models are expensive.

ML/DL models focus on high recovery with respect to input and does not handle conformational flexibility.

The goal is to sample diverse sequences --- overlooked in most ML studies.

IBM

3D Geometry-Aware *Diverse and Novel* Protein Sequence Design



(a) Consider a one hot-encoding $\mathbf{t_j} \in \{0,1\}^4$ of four types of secondary structures : 1. helix, 2. beta strands, 3. loop, 4. turn

(b) Scale in/out the structure into a fixed size box with ratio $\mathbf{r}$.

(c) Discretize the cubic space into $2\text{Å}\times2\text{Å}\times2\text{Å}$ voxels.

(d) For each voxel i, we sum up the contributions from all residues as:

$$\mathbf{y_i} = \sum_{\mathbf{j=1}}^{\mathbf{N}} \exp\left(-\frac{||\mathbf{c_j r} - \mathbf{v_i}||_2^2}{\sigma^2}\right) \cdot \mathbf{t_j}$$

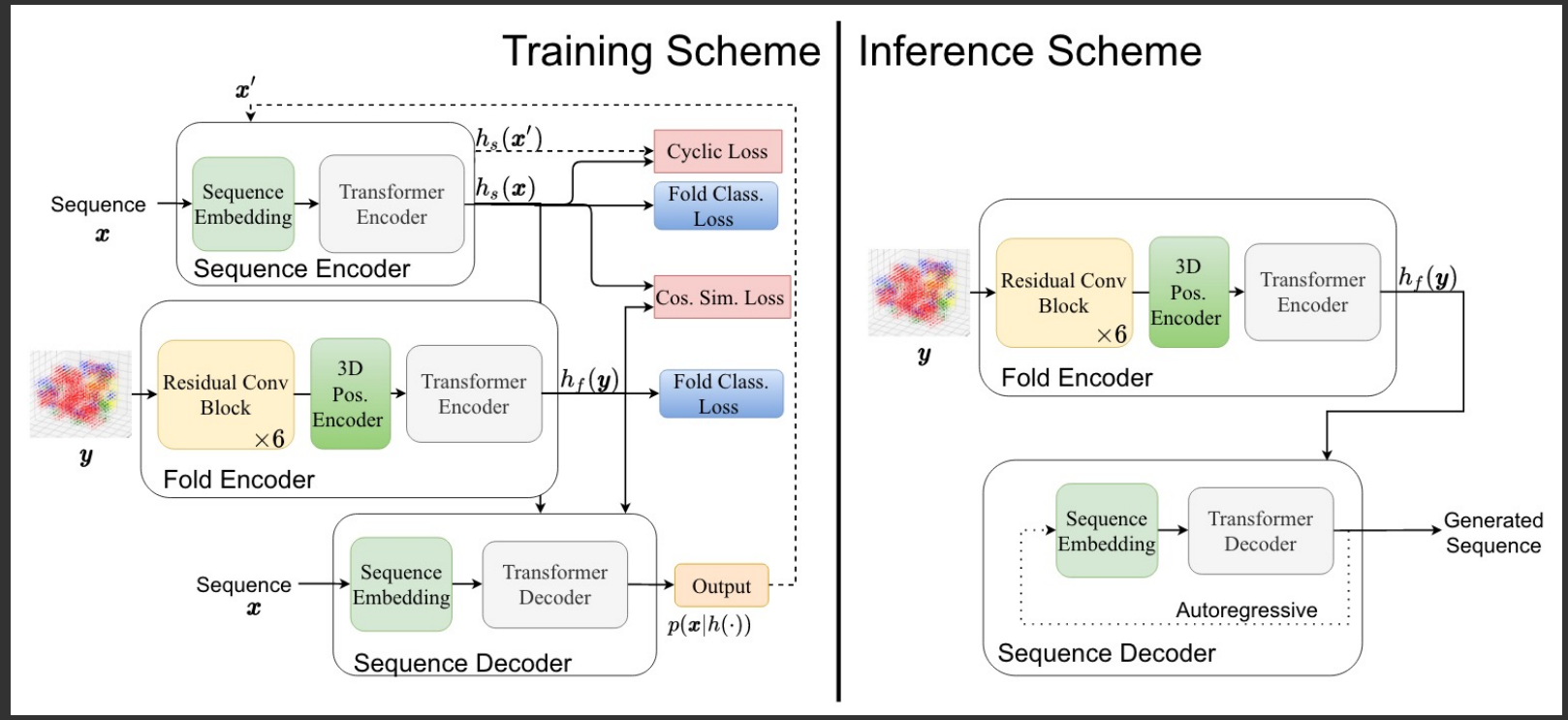
Where $\mathbf{c_j}$ is the coordinate of residue j, and $\mathbf{v_i}$ is the coordinate of the center of voxel i.

IBM Research, **Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design**. ICML 2021.
IBM Research, **Benchmarking deep generative models for diverse antibody sequence design.** NeurIPS LMRL 2021. https://arxiv.org/abs/2111.06801.

IBM

3D Geometry-Aware *Diverse and Novel* Protein Sequence Design

Goal: Learn a joint sequence–fold embedding



Two Intra-modal losses: fold classification: $\mathbf{FC_f}$ and $\mathbf{FC_s}$

Two reconstruction losses: fold2seq and seq2seq: $\mathbf{RE_f}$ and $\mathbf{RE_s}$

One Inter-modal loss: cosine similarity: $\mathbf{CS}$

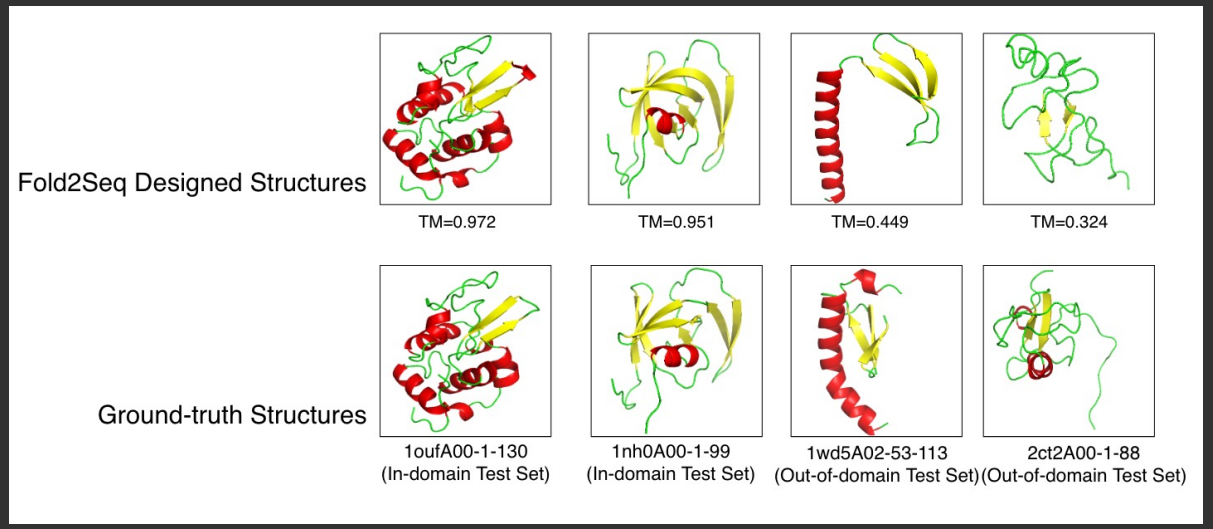
One cyclic sequence loss $\mathbf{CY}$

Full loss objective: $\mathbf{L = \lambda_1 RE_f + \lambda_2 RE_s + \lambda_3 FC_f + \lambda_4 FC_s + \lambda_5 (CY - CS)}$

IBM Research, **Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design**. ICML 2021.
IBM Research, **Benchmarking deep generative models for diverse antibody sequence design.** NeurIPS LMRL 2021. https://arxiv.org/abs/2111.06801.

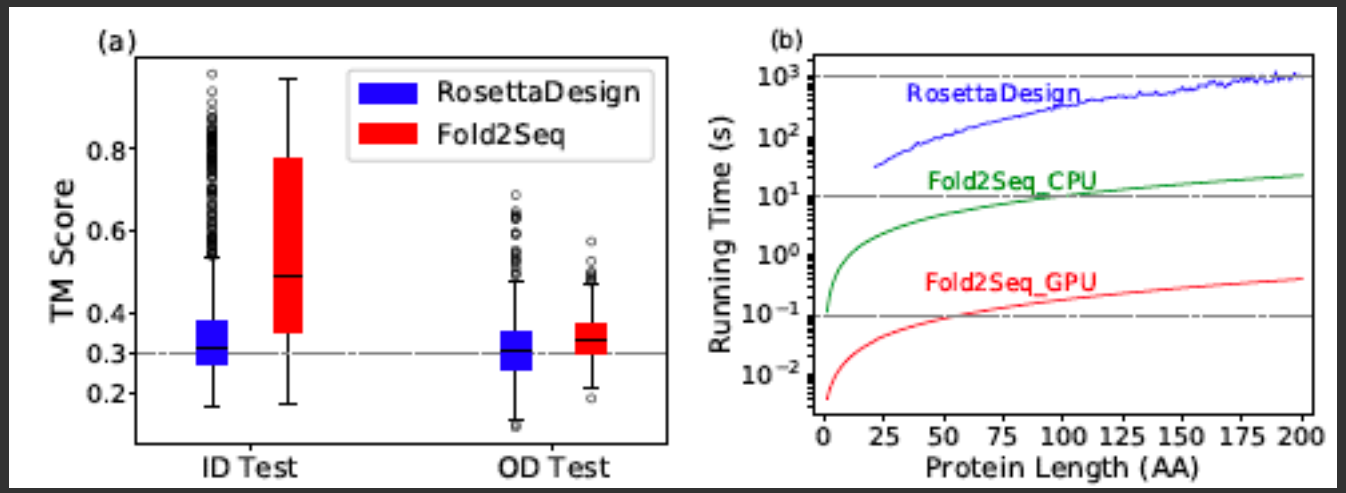
IBM

Fold2Seq Performance on Geometry-Aware Protein Sequence Design



Fold2Seq works in real-world settings – with inputs such as incomplete structure, low-resolution structure, or NMR structural ensemble.

Maintains fold consistency, while providing broad sequence diversity



	ID Test		OD Test	
Subset	$\lvert\mathcal{S}_i\rvert \leqslant 3$	$\lvert\mathcal{S}_i\rvert > 3$	$\lvert\mathcal{S}_i\rvert \leqslant 3$	$\lvert\mathcal{S}_i\rvert > 3$
$\#cov^{\mathrm{f}}_{\mathrm{fold}}(i) > cov^{\mathrm{g}}_{\mathrm{fold}}(i)$	104	53	13	8
Total #folds	118	78	18	10
Ratio	0.88	0.68	0.72	0.80

Ingraham, et al, **Generative models for graph-based protein design**. NeurIPS 2019.

IBM Research, **Fold2Seq: A Joint Sequence(1D)-Fold(3D) Embedding-based Generative Model for Protein Design**. ICML 2021.

IBM

Take Home

Large pre-trained models are emerging as a promising tool to be integrated in molecular prediction & design workflows.

While designing those models and methods, integration of domain knowledge and physics at each stage can help boost performance and efficiency.

Benchmarks and metrics are good for consistency and reproducibility, but we need to go beyond what currently exist and work with the community to create and validate new ones that are more realistic and relevant.

Emergent behavior with data and neural scaling – new paradigm of learning

Geometry can be implicitly and/or explicitly included in FMs efficiently with proper coarse-graining.

IBM

Acknowledgement

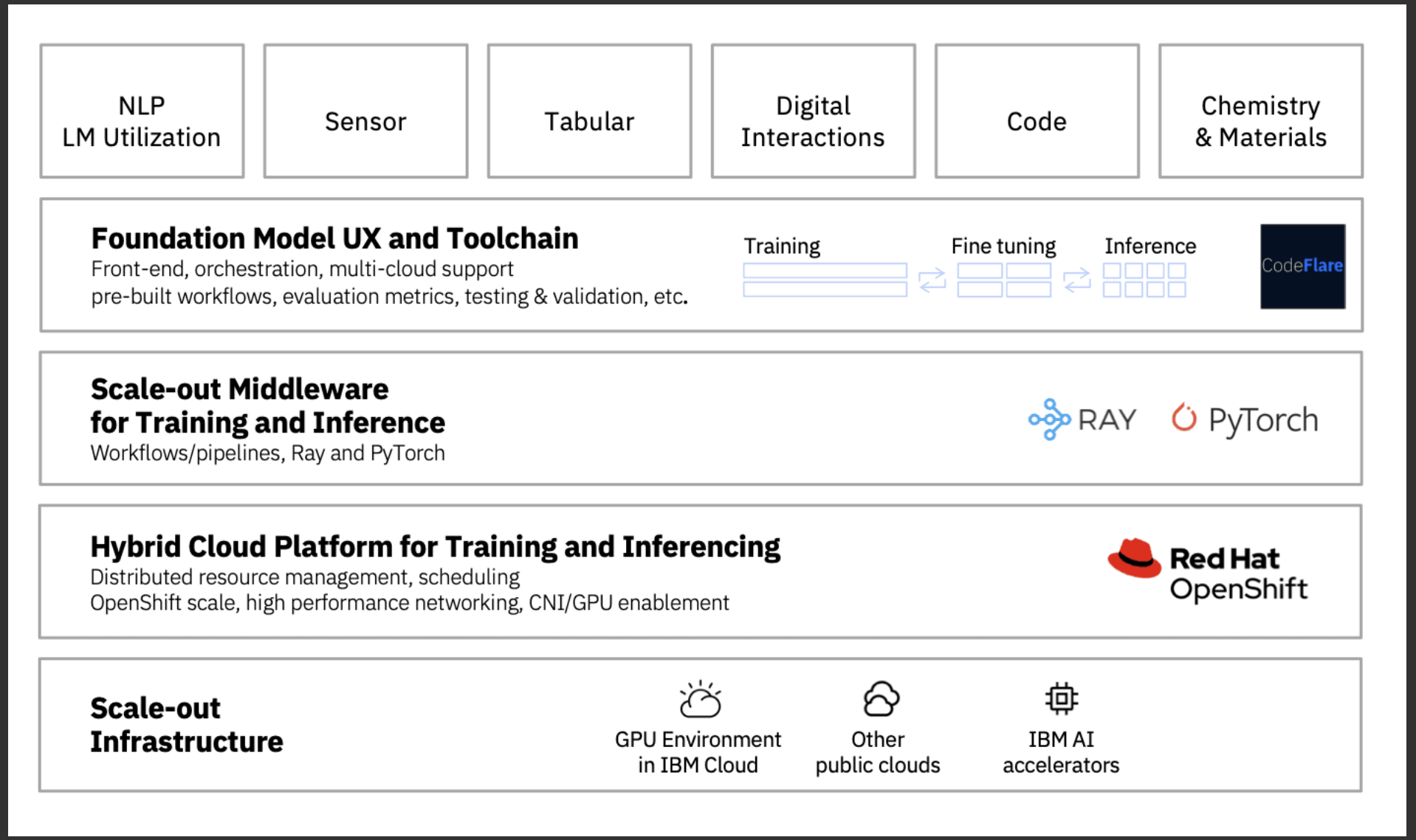
IBM Team: Aleksandra Mojsilovic, Pin-Yu Chen, Cicero Dos Santos, Enara Vijil, Tom Sercu, Inkit Padhi, Kahini Wadhawan, Flaviu Cipcigan, Jason Crain, Matteo Manica, Youssef Mroueh, Hendrik Strobelt, Brian Quanz, Ben Hoover, Kar Wai Lim, Karthik Shanmugam, Hamid Dadkhahi, Jesus M Rios, Igor Melnyk, Jim Hedrick, Yue Cao, Ria Vinod, Oscar Chang, Thanh Ngyuen, Joey Tatro, Eleni Litsa, Devleena Das, Amit Dhurandhar, Samuel Hoffman, Aurelie Lozano, Pierre Dognin, Brian Belgadore

Collaborators: Yoshua Bengio (MILA), Jian Tang (MILA), Giuseppe Romano (MIT), Lydia Kavraki (Rice U), John Dordick (RPI), Steven Johnson (MIT), S Subramanian (ANL), Yi-Yan Yang (A*Star, Singapore), Sansom group (Oxford, UK), Walsh and Stuart (Diamond UK), Weinstein & Schwartz (Weill Cornell), Yang Shen (TAMU), Rongie Lai (RPI)

IBM

Realizing the Value of Foundation Models

End to End Cloud Native and Customizable Stack



IBM