## Supplementary Information: Language Models can learn Complex Molecular Distributions

D. Flam-Shepherd et al.

Supplementary Discussion We experiment with additional graph generative model baselines on all tasks. These include HierVAE [1], GCPN [2], GRAPH AF [3], GENRIC [4], CNF [5], Molecular RNN (MRNN) [6], GRAPHVAE [7], NAT-GRAPHVAE [8], MOLGAN [9], GRAPH NVP [10], DGMG [11], MOLMP [12], GRAPHINVENT [13]. From these models, most of the single shot generative models do not scale- from MOLGAN, GRAPHVAE, NAT-GRAPHVAE, GRAPH NVP. None of these models including GCPN were able to acheive better than 1% valid, unique and novel- meaning they are unable to generate molecules from the training distribution. Furthermore, all of the autoregressive graph generative models (DGMG, MOLMP, GRAPH INVENT) were unable to handle the larger molecules- even in the LogP and Multi-distribution tasks. Training on these larger datasets exacerbate the stability issues [11] these models suffer from-making them unable to stably train to completion. The baselines that were able to train could only handle the LogP and multi-distribution tasks, these include: two discrete normalizing flow models CNF [5] and GRAPH AF [3], GENRIC which employs a Markov chain, MRNN or Molecular RNN which uses RNNs to generate atoms and bonds and HIERVAE which extends JTVAE to larger common motifs or substructures. All baselines have high scoring standard metrics (Table II) but their wasserstein distance metrics are much further from the Train Oracle than the RNNs (Table I). HIERVAE and MRNN stand out and are higher scoring than GENRIC, CNF and GRAPH AF-HIERVAE even beats the SF-RNN on SA and NP but not the SM-RNN. Indeed, from the distribution plot in Figure 1a for the LogP task we can see that MRNN and HIERVAE are closer to the training distribution than the additional baselines but nearly as close as the RNNs. For the multi-distribution task, the closest are MRNN and CNF, shown in the distribution plot in Supplementary Fig. 1c- where MRNN learns all of the modes (but poorly) while CNFentirely misses the CEP mode. In contrast the RNNs, perfectly learn all four modes (Supplementary Fig. 1b).

Task	Task Samples		$\mathbf{SA}$	QED	MW	BCT	NP
	TRAIN	0.020	0.0096	0.0029	1.620	7.828	0.013
	SM-RNN	0.095	0.0312	0.0068	3.314	21.12	0.054
	SF-RNN	0.177	0.2903	0.0095	6.260	25.00	0.209
$^{5}$ P	HIERVAE	0.661	0.0464	0.0710	51.73	141.9	0.079
Γõ	MRNN	0.769	1.2321	0.0710	58.27	142.9	0.898
	GRAPHAF	3.534	1.8820	0.2413	164.7	664.4	1.206
	CNF	2.773	3.4727	0.1879	37.87	174.7	1.456
	GENRIC	2.764	1.3626	0.1092	81.41	308.0	1.286
	TRAIN	0.048	0.0158	0.0020	2.177	14.15	0.010
	SM-RNN	0.081	0.0246	0.0059	5.483	21.19	0.012
	SF-RNN	0.286	0.1791	0.0227	11.35	68.81	0.079
lti	HIERVAE	2.356	0.2151	0.1024	157.7	687.0	0.175
Μu	MRNN	1.519	0.6644	0.0593	97.92	400.1	0.598
	GRAPHAF	3.140	1.9122	0.1174	106.1	971.7	0.723
	CNF	2.378	2.0793	0.0991	61.87	436.7	1.070
	GENRIC	1.623	2.0029	0.0827	105.7	445.3	0.787

Supplementary Table I. Wasserstein distance metrics for LogP, SA, QED, MW, BT and NP between molecules from the training data and generated by the additional baselines and RNNs for all three tasks. Values closer to TRAIN are better.

Task Metric  SM-RNN SF-RNN HIERVAE MRNN GRAPHAF CNF GENR
--

$\operatorname{LogP}$	validity	0.941	1.000	1.000	1.000	1.000	1.000	1.000
	unique	0.987	1.000	1.000	0.999	0.906	1.000	0.886
	novelty	0.721	0.871	1.000	0.994	1.000	1.000	0.993
Multi	valid	0.969	1.000	1.000	0.999	1.000	1.000	0.997
	unique	0.996	0.989	0.938	0.999	0.985	1.000	0.912
	novelty	0.937	0.950	1.000	1.000	1.000	1.000	0.998

Supplementary Table II. Standard Metrics. From molecules generated by all models (Closer to 1.0 is better).



Supplementary Fig. 1. Additional Baselines a For the LogP task, the histogram and KDE of penalized logP of training molecules along with KDEs of molecular weight of molecules generated from additional baselines model that could generate samples. b The histogram and KDE of molecular weight of training molecules along with KDEs of molecular weight of molecular weight of training molecules along with KDEs of molecular weight of molecules with the training data and RNNs. c The histogram and KDE of molecular weight of training molecular weight of molecular weight of training molecular weight of training molecular weight of molecular weight of training molecular weight of molecular weight of training molecular weight of training molecular weight of molecular weight of training molecular weight

 ${\bf e}~$  Molecules with more than 10 rings.

 $\label{eq:supplementary Fig. 2. Penalized LogP Task a-e Training molecules with different properties.$ 

	Generated Molecules									
TRAIN	୦୦ମ୍ବ୦୦	9.00	~~000	no de	220	<u>~~~~~</u>	molo	ogod	وبطحو	Arost.
	and	<sup>రార్థ</sup> ికి	o <sup>dq</sup> ~~qd <sub>o</sub>	\$\$	agra	aza	appedro	ordia	ardizk	ంసార్రం
	Super	Ronatio	oraçio	<del>с.</del>	æfte o	فيتضع	o data	<u>ई</u> ठेन्द	ka go	ofre
	or of o	p <sup>and</sup> q	away	Sugo	o sto	రాహిం	لكطرحه	დიკაც	~9~00+	Ent
	arat	مهم	ᡬᢌᡊᢓ	હત્ર્દ્વ	8.00	કેન્સ્	محمرم	all a	ه مکتمه	مگره
E	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	ᠹᡔ᠆ᡇ	ኊቝ	ó~~o	Go	outo	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	بهني	K.,48	₹¥
	्र्मूख्	39¢	<del>2</del> 200	<i>કેન્</i> કેન્દ્ર	õra	<del>3</del> 2-	Jergen	معدل	<del>ç</del> êr	office
GVA	-9 <mark>-</mark> 9-9-9	- <b>O</b> -	(Jp	¢.	بمكر	محجم	Fairs	Jorpo	÷\$	\$6g2
C	<u> </u>	Lann	عسطور	र्म्ह्र् <del>स्ट्रि</del> स	Der	mag	20	Ś	ञ्च्क	and
	rapa.	nal	or and	्रक्ट्यूर	aging	<i>b</i> y		~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	-9,0,-0	Err
JTVAE	and the second	8sta	+01010	princ	Sr.p	ogo	angras	10-10-10	Raft	ound
	and	gorord	ofqa	3.05	of the	26	arta	40,00	xoring	afr
	Sard	-0-0-0	ዹኇ	ogaa	×24.90	010100	30	ano	From Control	-
	for	g.g.ro	xorda	379	ograo	क्ष	g.g.s	9,00	and	for
	৾৾৵ঀৢ৾৾ঀ	A-B	prop	niz	5.85	++++++040	quara	0.0020	Sec.	and and
	argo	Atta-8	Êa	°~~¢	$\sqrt{2}$	ઝ્ટ્રેલ	कत्वर्	۲.3° ۵ ۵	grata	or do
z	822	to any to	tà	æ.g.	aourop.	ᡐᡔᡠ	ago à	ထိုလ်ခ	ممرو	6,000
F-RN	6 b b b b b b b b b b b b b b b b b b b	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	0°5.0.	ቝጜ	55	Star	ውሪንያ	drep.	రారాధ	సాచాని
$\mathbf{S}$	જ્યુઝ	orato	مؤه	مريكم	opog	Lordon	ممغه	ထထ	~ <del>~~</del> ~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	ade
	જુર્સ્ટ	જાજ્યરી	.0-0 <sup>37</sup> Q.	প্রপন্ধ	der to the second secon	ځېکې لو	ᢐᡇᡐ	Britick	azero	ጭ <sub>ተ</sub> ትዮ
M-RNN	وكمبرغ	લ્સુર	anag	~ <del>§</del> হ্হ~এ	¢,	inter of the second	مېتونه	der de	مهم	and the
	and	jandary	40°20	مرکونرم	~ હુવ	ڡٛۄؠؠٮڡ	migo	afra	gard	ondo
	and	చారాళ్	maggi	4-3 4-4	Ext.d	grave	and a	ౚం౪ౢౢౢౢౢౢౢౢౢౢౢ	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	queo
S	asia	prodera	~~~~g	ዮራሪ	}	ኆ፟፝፞፞፞፞፞፞፞፞፞፞	4000 4000 4000	8 pro	to to t	0-0-60 0-
	Inos.	Raps	de la	age of	20-corto	ogo	grant	gonal	منكب	ordy or

 $\label{eq:supplementary Table III. Penalized LogP Task \ \ \mbox{Molecules generated from each model}.$ 



 ${\bf b}$  Molecules from CGVAE.

Supplementary Fig. 3. Penalized LogP Task a-b Generated molecules with penalized LogP < 4 from the graph generative models.

	Generated molecules									
TRAIN	- tr	¢¢	Ř	مراجع	Ŷ	3	all a	in		<u>۲۰</u>
	Aron	مد	ર્ <sub>દ્</sub> રુંત	৵৾৻৵ঀ	9 <b>~</b> 9	art~a	94 <del>8</del>	रहंष	aya	getty
	88-8	બ્દુવૈઝ	<del></del> જ્યુ <sup>ક</sup>	ages	george .	Bras	6-03A	alas	fraction	8-acb
	Acardan .	- tool grade	topo and the second sec	సిచార్తారిచా	or the	de d	grage frage	egentine.	and the state of t	æ
CGVAE	¢¢	œ	æ	đ	<i>Yor</i>	Ø	p.c	α	<b>5</b> 00	д
	Ç¢.	Q.J	Q-	ф. С	-¢p	ago	ᡐᢁᢅᢓ	tQ\$	Jung	æ.
	$\mathcal{O}$	-àger	zofurz	<del>~</del> \$	केल	gundo	ণ্ট্ৰ স ম	Sr.St	Jan Contraction	₿ C>
	par Cafe	- A	Att Car	\$	States.	-Ardere	and the second	je to the second	Ģ	Ŷ
JTVAE	<b>%</b> (	-rt	*	40	x	q <del>r</del> (	alac	<i>h</i> ~	+	~
	28.	ક્ષ્ટ	Zefr	фан	alge	480	with	୍ଟ	rop do	5×0
	aga	870°	მადგ	య్రాం	870	ංංදේල	جمعهم	ბობ	చార్తిల	8 States
	Herperator	ar by	Jow Bridge	-G28-380	E total	*******	engene en	States	\$7888£	wyge
	yne	Yart	54	بلار	<del>Tol</del>	~?	ネ	~k	z	rt-
<b>SNN</b>	roug	ort	200	Correcto	ક્રેન્ટ્રે	mea	xrizo	ort	tre	٤ <u>.</u> ۶
SF-I	8gag	adars	age go	დგგ	9890	erada	ოზმ	ංදුර්ග	<u>६</u> ,फ़ि	ራሌያο
	nder frank	Broadt	ഷ്ക്ഷ്ക	2 to the second	apaabe	ڡڡڰٚڮۣڰٚؠڡ	af bad	at way a	ర్శంథంళ్ల	Jones .
SM-RNN	L	~~~D	æ	<i>~~</i>	~~~~t	-04	₽~~	10°	zhe	p2
	-0- <sup>9</sup> 3	-840	୯୪୦୮	rip.O	80-	مربع	or of the	ort	ogot	arta
	පරිත	<del>6</del> ይ <del>ኖ</del> ጵ	იძმყ	8 or an	ዮውያ	გ <sup>8</sup> თ	ంచ్లరిం	<u> </u>	ዮංක	ზფმ
	~~xxxx	<u>مَ</u> لْعِنْمُ	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	- proposed	* English		to and the second	t to the	S S S S S S S S S S S S S S S S S S S	యాచిర్తు

Supplementary Table IV. **Multi-distribution Task** Model generated molecules. Each sub-row is from a specific molecular mode.



ø





متهموهمي

A BELLER









-----









Supplementary Fig. 4. Large Scale Task Training molecules.



4 ton

600

Q

Supplementary Fig. 5. Large Scale Task Generated molecules from the SMILES RNN.





Supplementary Fig. 6. Large Scale Task Generated molecules from the SELFIES RNN.



Supplementary Fig. 7. Large Scale Task a-b Generated molecules with less than 100 heavy atoms from the RNN models.



 ${\bf b}~$  Molecules from the mode with higher LogP values.

Supplementary Fig. 8. Large Scale Task a-b Training molecules from each LogP mode.

- Wengong Jin, Regina Barzilay, and Tommi Jaakkola, "Hierarchical generation of molecular graphs using structural motifs," in International Conference on Machine Learning (PMLR, 2020) pp. 4839–4848.
- [2] Jiaxuan You, Bowen Liu, Zhitao Ying, Vijay Pande, and Jure Leskovec, "Graph convolutional policy network for goaldirected molecular graph generation," "Graph convolutional policy network for goal-directed molecular graph generation," Advances in Neural Information Processing Systems, , 6410–6421 (2018).
- [3] Chence Shi, Minkai Xu, Zhaocheng Zhu, Weinan Zhang, Ming Zhang, and Jian Tang, "Graphaf: a flow-based autoregressive model for molecular graph generation," "Graphaf: a flow-based autoregressive model for molecular graph generation," arXiv preprint arXiv:2001.09382 (2020).
- [4] Ari Seff, Wenda Zhou, Farhan Damani, Abigail Doyle, and Ryan P Adams, "Discrete object generation with reversible inductive construction," in Advances in Neural Information Processing Systems.
- [5] Phillip Lippe and Efstratios Gavves, "Categorical normalizing flows via continuous transformations," "Categorical normalizing flows via continuous transformations," arXiv preprint arXiv:2006.09790 (2020).
- [6] Mariya Popova, Mykhailo Shvets, Junier Oliva, and Olexandr Isayev, "Molecularrnn: Generating realistic molecular graphs with optimized properties," "Molecularrnn: Generating realistic molecular graphs with optimized properties," arXiv preprint arXiv:1905.13372 (2019).
- [7] Martin Simonovsky and Nikos Komodakis, "Graphvae: Towards generation of small graphs using variational autoencoders," in International Conference on Artificial Neural Networks (Springer, 2018) pp. 412–422.
- [8] Youngchun Kwon, Jiho Yoo, Youn-Suk Choi, Won-Joon Son, Dongseon Lee, and Seokho Kang, "Efficient learning of nonautoregressive graph variational autoencoders for molecular graph generation," "Efficient learning of non-autoregressive graph variational autoencoders for molecular graph generation," Journal of cheminformatics (2019).
- [9] Nicola De Cao and Thomas Kipf, "Molgan: An implicit generative model for small molecular graphs," "Molgan: An implicit generative model for small molecular graphs," arXiv preprint arXiv:1805.11973 (2018).
- [10] Kaushalya Madhawa, Katushiko Ishiguro, Kosuke Nakago, and Motoki Abe, "Graphnvp: An invertible flow model for generating molecular graphs," "Graphnvp: An invertible flow model for generating molecular graphs," arXiv preprint arXiv:1905.11600 (2019).
- [11] Yujia Li, Oriol Vinyals, Chris Dyer, Razvan Pascanu, and Peter Battaglia, "Learning deep generative models of graphs," "Learning deep generative models of graphs," arXiv preprint arXiv:1803.03324 (2018).
- [12] Yibo Li, Liangren Zhang, and Zhenming Liu, "Multi-objective de novo drug design with conditional graph generative model," "Multi-objective de novo drug design with conditional graph generative model," Journal of cheminformatics 10, 1–24 (2018).
- [13] Rocío Mercado, Tobias Rastemo, Edvard Lindelöf, Günter Klambauer, Ola Engkvist, Hongming Chen, and Esben Jannik Bjerrum, "Graph networks for molecular design," "Graph networks for molecular design," Machine Learning: Science and Technology 2, 025023 (2021).