LigGPT: Molecular Generation using a Transformer-Decoder Model

Viraj Bagal,[†,‡] Rishal Aggarwal,[†] P. K. Vinod,[†] and U. Deva Priyakumar[*,†]

[†]*International Institute of Information Technology, Hyderabad 500 032, India*

[‡]*Indian Institute of Science Education and Research, Pune 411 008, India*

E-mail: deva@iiit.ac.in

List of Tables

List of Figures

Table S1: Comparison of CharRNN and LigGPT on the generation of 10,000 molecules by training only on 10% of the MOSES train split. CharRNN and LigGPT have 11.9 M and 6.3 M trainable parameters respectively.

Model	Validity	Unique	Novelty	Temperature
CharRNN	0.961	1.0	0.888	0.9
LigGPT	0.983	1.0	0.903	0.9
CharRNN	0.581	1.0	0.987	1.6
LigGPT	0.707	1.0	0.985	1.6

Table S2: Scaffold + Single property (logP, TPSA) conditional training on MOSES dataset. Temperature 1.6 was used. Metric calculated only for molecules having tanimoto similarity of the scaffold of the generated molecule and the scaffold used for condition greater than 0.8. **(a)** O=C(Cc1ccccc1)NCc1ccccc1 **(b)** c1cnc2[nH]ccc2c1 **(c)** c1ccc(-c2ccnnc2)cc1 **(d)** c1ccc(-n2cnc3ccccc32)cc1 **(e)** O=C(c1cc[nH]c1)N1CCN(c2ccccc2)CC1

Cond	Validity	Unique	Novelty	MAD	Cond	Validity	Unique	Novelty	MAD
(a)+logP	0.893	0.812	1.0	0.145	(a)+TPSA	0.906	0.870	1.0	2.303
(b)+logP	0.712	0.975	1.0	0.151	(b)+TPSA	0.692	0.961	1.0	3.239
(c)+logP	0.826	0.922	1.0	0.146	(c)+TPSA	0.894	0.874	1.0	2.439
(d)+logP	0.891	0.858	1.0	0.160	(d)+TPSA	0.902	0.891	1.0	3.178
(e)+logP	0.898	0.461	1.0	0.125	(e)+TPSA	0.882	0.431	1.0	3.986
(a)+SAS	0.812	0.934	1.0	0.124	(a)+QED	0.872	0.951	1.0	0.05
(b)+SAS	0.726	0.775	1.0	0.174	(b)+QED	0.702	0.98	1.0	0.052
(c)+SAS	0.698	0.862	1.0	0.167	(c)+QED	0.849	0.947	1.0	0.045
(d)+SAS	0.823	0.910	1.0	0.173	(d)+QED	0.905	0.933	1.0	0.072
(e)+SAS	0.820	0.541	1.0	0.125	(e)+QED	0.824	0.571	1.0	0.081

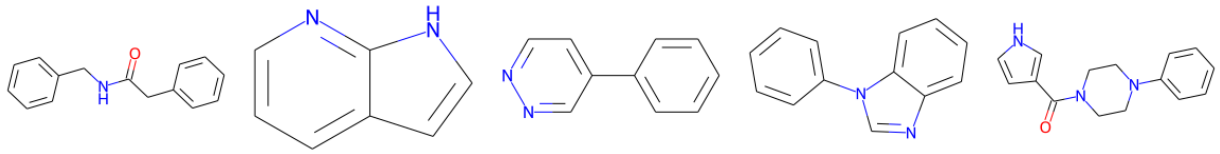
Table S3: Scaffold + Multi-property conditional training on MOSES dataset. Temperature 1.6 was used. Metric calculated only for molecules having tanimoto similarity of the scaffold of the generated molecule and the scaffold used for condition greater than 0.8. **(a)** O=C(Cc1ccccc1)NCc1ccccc1 **(b)** c1cnc2[nH]ccc2c1 **(c)** c1ccc(-c2ccnnc2)cc1 **(d)** c1ccc(-n2cnc3ccccc32)cc1 **(e)** O=C(c1cc[nH]c1)N1CCN(c2ccccc2)CC1

Cond	Validity	Unique	Novelty	MAD_TPSA	MAD_logP
(a)+TPSA+logP	0.812	0.737	1.0	3.667	0.249
(b)+TPSA+logP	0.693	0.931	1.0	4.117	0.199
(c)+TPSA+logP	0.830	0.852	1.0	3.903	0.152
(d)+TPSA+logP	0.773	0.818	1.0	4.617	0.204
(e)+TPSA+logP	0.776	0.511	0.999	4.046	0.242

Cond	Validity	Unique	Novelty	MAD_SAS	MAD_logP
(a)+SAS+logP	0.727	0.818	1.0	0.146	0.255
(b)+SAS+logP	0.591	0.649	1.0	0.193	0.191
(c)+SAS+logP	0.75	0.711	1.0	0.196	0.183
(d)+SAS+logP	0.748	0.731	1.0	0.171	0.246
(e)+SAS+logP	0.847	0.439	1.0	0.153	0.203

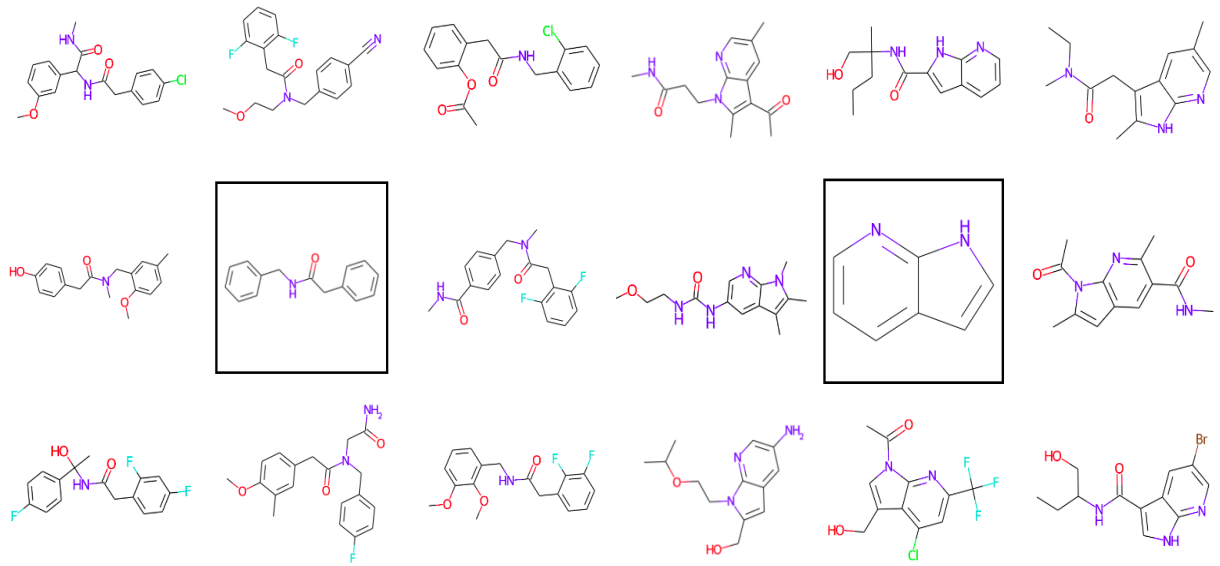
Cond	Validity	Unique	Novelty	MAD_TPSA	MAD_SAS
(a)+TPSA+SAS	0.751	0.901	1.0	3.947	0.192
(b)+TPSA+SAS	0.649	0.744	1.0	5.120	0.226
(c)+TPSA+SAS	0.683	0.803	1.0	4.074	0.210
(d)+TPSA+SAS	0.733	0.861	1.0	4.345	0.199
(e)+TPSA+SAS	0.838	0.482	1.0	3.827	0.162

Cond	Validity	Unique	Novelty	MAD_TPSA	MAD_logP	MAD_SAS
(a)+TPSA+logP+SAS	0.618	0.681	1.0	4.935	0.551	0.311
(b)+TPSA+logP+SAS	0.653	0.649	1.0	5.325	0.238	0.262
(c)+TPSA+logP+SAS	0.582	0.620	1.0	5.318	0.292	0.242
(d)+TPSA+logP+SAS	0.530	0.646	1.0	5.559	0.531	0.309
(e)+TPSA+logP+SAS	0.754	0.388	1.0	5.729	0.403	0.241



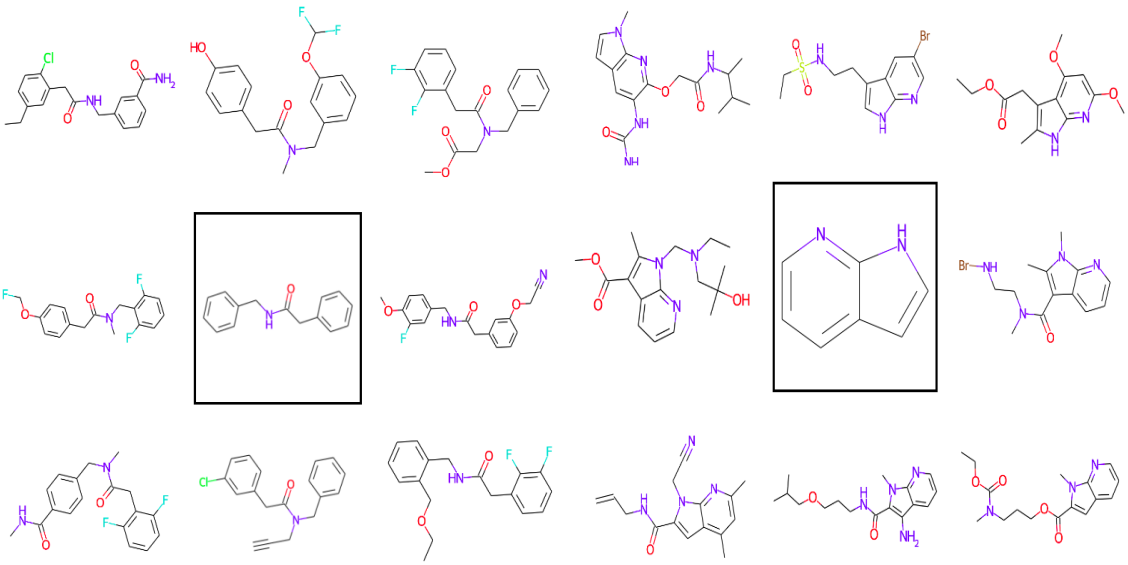
(a) Scaffold 1 (b) Scaffold 2 (c) Scaffold 3 (d) Scaffold 4 (e) Scaffold 5

Figure S1: Scaffolds from test set used for scaffold + property based conditioning results.



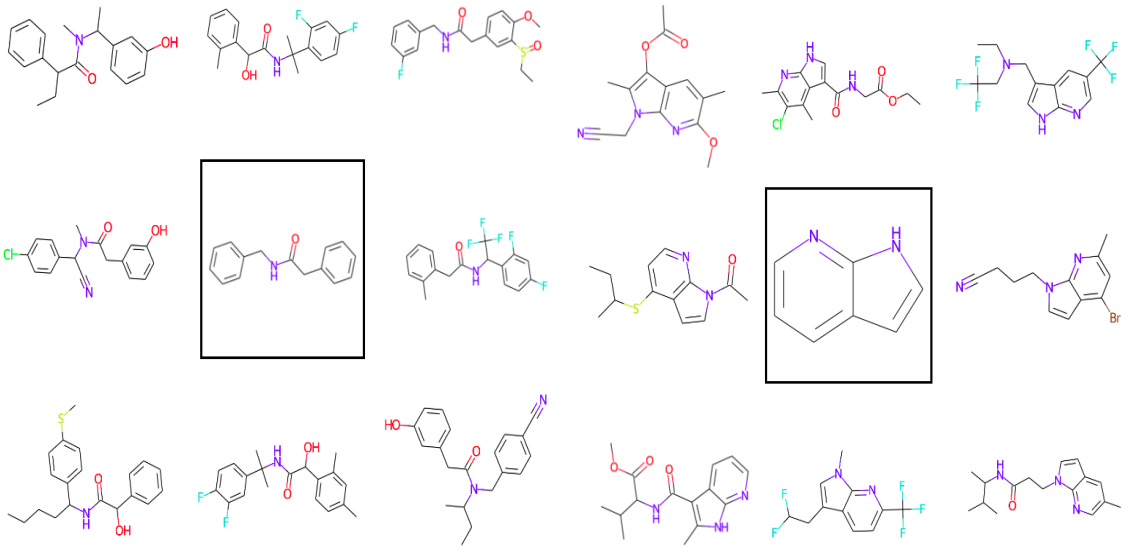
(a) Scaffold 1 (b) Scaffold 2

Figure S2: Scaffold 1: O=C(Cc1ccccc1)NCc1ccccc1. Scaffold 2: c1cnc2[nH]ccc2c1. In both the subfigures, the molecule in black box is the scaffold used for conditional generation. **(a, b)** 8 random generated molecules having the same scaffold as scaffold 1 and 2 respectively.



(a) Scaffold 1, logP

(b) Scaffold 2, logP



(c) Scaffold 1, SAS

(d) Scaffold 2, SAS

Figure S3: Scaffold 1: O=C(Cc1ccccc1)NCc1ccccc1. Scaffold 2: c1cnc2[nH]ccc2c1. In all the subfigures, the molecule in black box is the scaffold used for conditional generation. **(a, b)** Conditioned on scaffold as well as logP = 2. **(c, d)** Conditioned on scaffold as well as SAS = 2.75.