

Shetti-AI: Precise prediction of protein sequence mutational landscapes to accelerate functional assays

Haitham Sobhy

Supplementary Tables

Table S1. In Excel format, the 18-dimensional physicochemical feature table.

Table S2. In Excel format, Codon-aware substitution accessibility matrix, the codes used SNP for short.

Table S3. Output of architecture 1; Convex linear λ -baseline model. Lambda = 0.2 means that the model weighs physicochemical by 80% and the codon matrix is 20%.

motif	position	from	to	snp_distance	p_dna	p_phys	p_final
MKEL	2	R	K	3	0.0302	0.99235	0.8017
MRDL	3	E	D	3	0.01795	0.99408	0.79991
MREV	4	L	V	3	0.02871	0.5596	0.45438
MREI	4	L	I	3	0.02871	0.37357	0.30525
LREL	1	M	L	3	0.014	0.2866	0.23232
VREL	1	M	V	3	0.014	0.28385	0.23012
IREL	1	M	I	3	0.014	0.20618	0.16792
AREL	1	M	A	1	0.10342	0.12656	0.12205
CREL	1	M	C	1	0.10342	0.06344	0.07151
MREA	4	L	A	2	0.07804	0.06316	0.06628
MRCL	3	E	C	1	0.13262	0.00001	0.02657
MRWL	3	E	W	1	0.13262	0.00001	0.02656
MRFL	3	E	F	1	0.13262	0	0.02656
FREL	1	M	F	2	0.03805	0.02117	0.02457
PREL	1	M	P	1	0.10342	0.00054	0.02114
GREL	1	M	G	1	0.10342	0.00028	0.02093
DREL	1	M	D	1	0.10342	0.00014	0.02081
YREL	1	M	Y	1	0.10342	0.00013	0.02081
MNEL	2	R	N	2	0.08209	0.00217	0.01819
MEEL	2	R	E	2	0.08209	0.00002	0.01647
MDEL	2	R	D	2	0.08209	0.00001	0.01647
MYEL	2	R	Y	2	0.08209	0.00001	0.01646
MIEL	2	R	I	2	0.08209	0	0.01646
MVEL	2	R	V	2	0.08209	0	0.01646
MFEL	2	R	F	2	0.08209	0	0.01646

Table S4. Output of architecture 2; log-space Bayesian hybrid model

Mutant	Position	From	To	SNP Distance	Genetic Weight	Phys Similarity	Prior	Likelihood	Posterior
MRDL	3	E	D	1	3	0.4536	7.32%	6.1379	17.29%
MREI	4	L	I	1	3	0.5383	6.25%	8.6121	16.80%
MKEL	2	R	K	1	3	0.4806	6.12%	6.8364	16.16%
MREV	4	L	V	1	3	0.4436	6.25%	5.8979	11.50%
LREL	1	M	L	1	3	0.2931	8.11%	3.2299	10.67%
IREL	1	M	I	1	3	0.2862	8.11%	3.1419	10.38%
VREL	1	M	V	1	3	0.2768	8.11%	3.0256	9.99%
MRQL	3	E	Q	1	3	0.2782	7.32%	3.0434	8.57%
MQEL	2	R	Q	1	3	0.2951	6.12%	3.2551	7.69%
MREA	4	L	A	2	2	0.4344	4.17%	5.6835	7.39%
MREF	4	L	F	1	3	0.3204	6.25%	3.6022	7.03%
TREL	1	M	T	1	3	0.18	8.11%	2.0548	6.79%
MRAL	3	E	A	1	3	0.2117	7.32%	2.3319	6.57%
KREL	1	M	K	1	3	0.1718	8.11%	1.9885	6.57%
MRGL	3	E	G	1	3	0.2054	7.32%	2.2743	6.41%
MREM	4	L	M	1	3	0.2931	6.25%	3.2299	6.30%
MRKL	3	E	K	1	3	0.1941	7.32%	2.174	6.13%
MREW	4	L	W	1	3	0.2844	6.25%	3.1189	6.08%
FREL	1	M	F	2	2	0.2461	5.41%	2.6764	5.89%
MRVL	3	E	V	1	3	0.1828	7.32%	2.0772	5.85%
MTEL	2	R	T	1	3	0.2204	6.12%	2.4147	5.71%
WREL	1	M	W	2	2	0.2343	5.41%	2.5531	5.62%
MSEL	2	R	S	1	3	0.2148	6.12%	2.3609	5.58%
MHEL	2	R	H	1	3	0.2044	6.12%	2.2646	5.35%
MAEL	2	R	A	1	3	0.2006	6.12%	2.2311	5.27%

Table S5. Output of architecture 3; Generative GAN search expansion engine

Index	Motif	log-Bayes	MSE
1	LKDM	-8.8768	0.008568
2	LKRM	-18.5254	0.059822
3	LEDM	-19.0199	0.051706
4	LKMM	-19.8377	0.066674
5	LMDM	-21.5841	0.075302
6	LMRM	-31.2327	0.059985
7	LMMM	-32.5449	0.216999

Table S6. Output of architecture 3; Generative WGAN-GP search expansion engine

Index	Motif	log-Bayes	MSE
1	AKDV	-1.9757	0.035426
2	LKDV	-2.2572	0.028648
3	VKDV	-2.2668	0.034456
4	AKDI	-2.3799	0.033456
5	IKDV	-2.5864	0.034493
6	LKDI	-2.6613	0.028839
7	VKDI	-2.6709	0.034493
8	IKDI	-2.9905	0.035969
9	CKDI	-3.0701	0.055203
10	AKDA	-3.4641	0.049678
11	VKDA	-3.7551	0.035426
12	AKDM	-8.5954	0.016431
13	LKDM	-8.8768	0.008568
14	VKDM	-8.8864	0.012347
15	IKDM	-9.2061	0.011742
16	AMDV	-14.683	0.065012
17	LMDV	-14.9644	0.067804
18	LMDI	-15.3685	0.066875
19	CMDV	-15.3733	0.098609
20	AMDA	-16.1714	0.064896
21	AMDM	-21.3027	0.073595
22	LMDM	-21.5841	0.075302

Supplementary Figures

Fig S1. Binding between MREL motif in nsp1 of SARS-CoV-2 with human ribosome adopted from Sobhy COVID 2021, and Thoms et al. Science 2020. MREL motif was first reported in Sobhy preprint 2020 (<https://www.preprints.org/manuscript/202004.0171>).

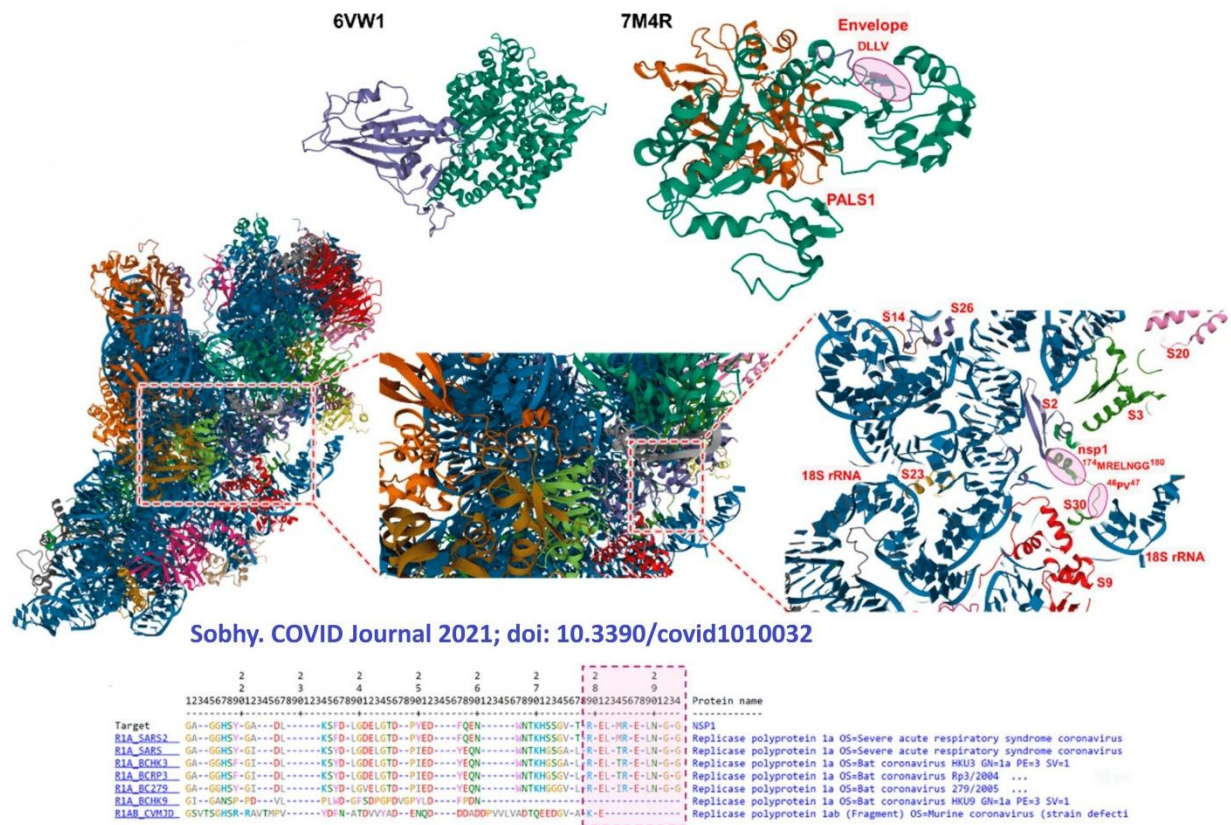


Fig S2. Heatmap of the 18-dimensional physicochemical properties, from Table S1.

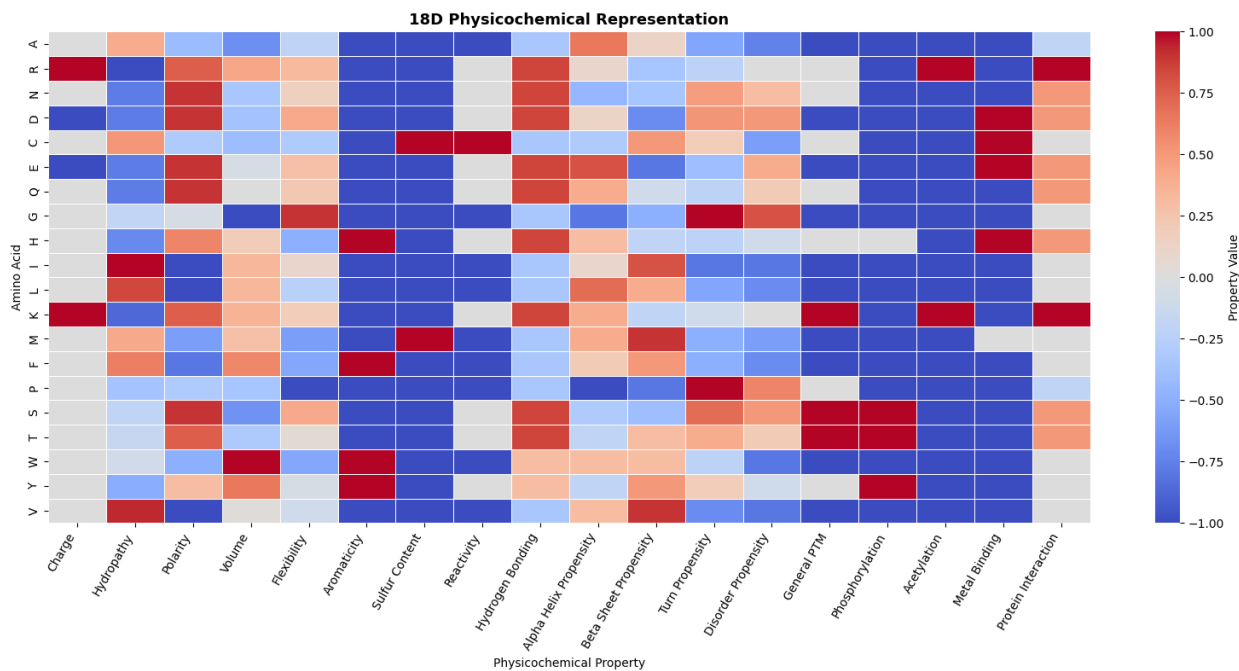


Fig S3. Heatmap of the 18-dimensional physicochemical properties, from Table S2.

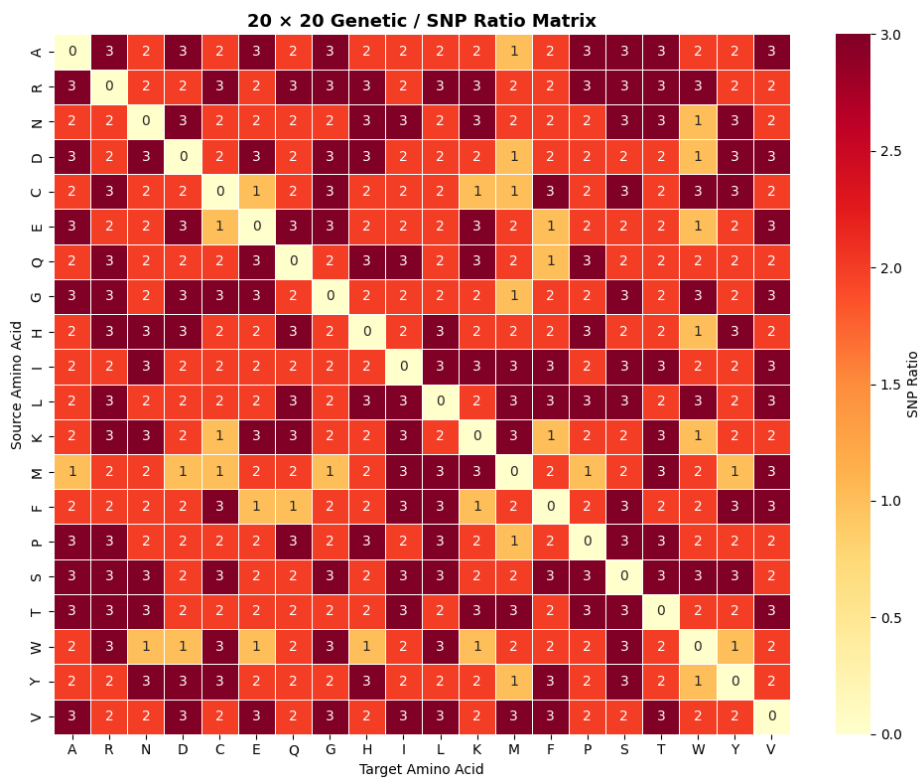


Fig S4. Lambda substitution with respect to mutations. The top MREL variants are: MKEL, MRDL, MREV, and MREI motifs.

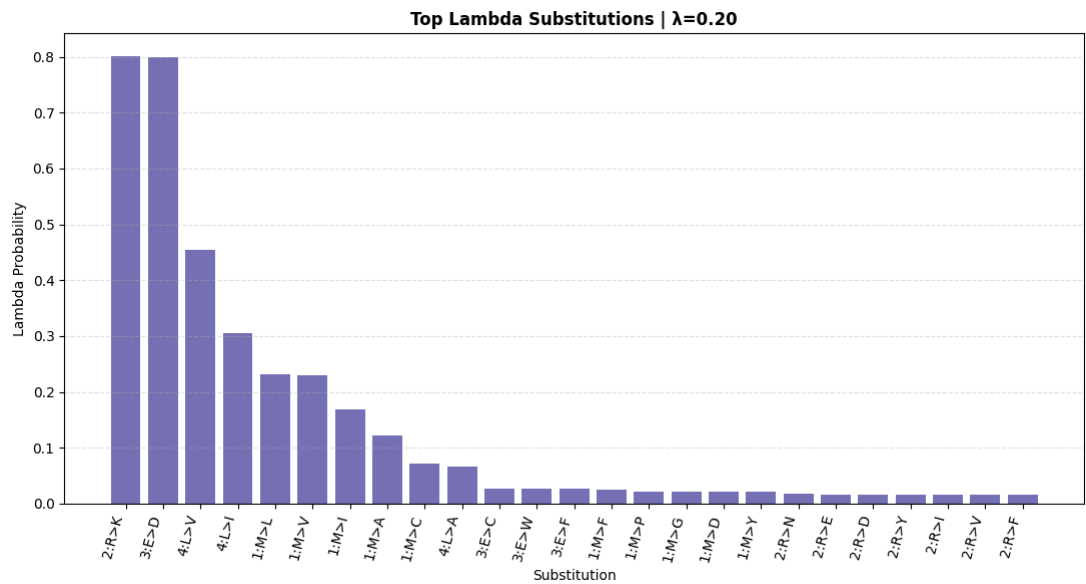


Fig S5. Contribution of genetics and codon substitution matrix (for short SNP distance) in Bayesian Posterior. The model favors the residues that are 1 SNP apart.

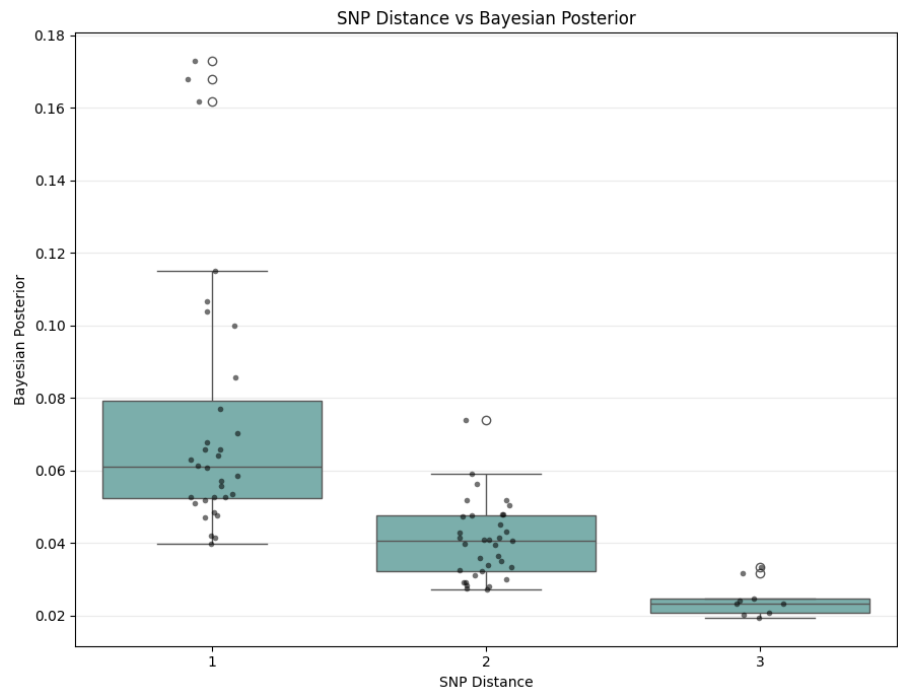


Fig S6. Top candidate resulted from Bayesian model. The top MREL variants are: MRDL, MREI, MKEL, and MREV motifs.

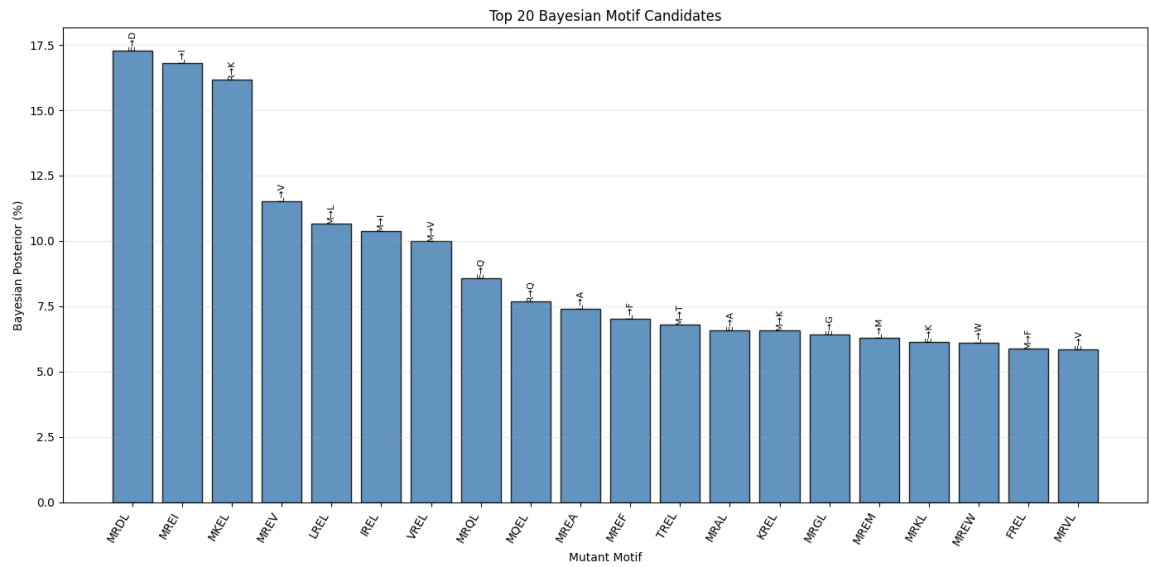


Fig S7. GAN and WGAN loss functions

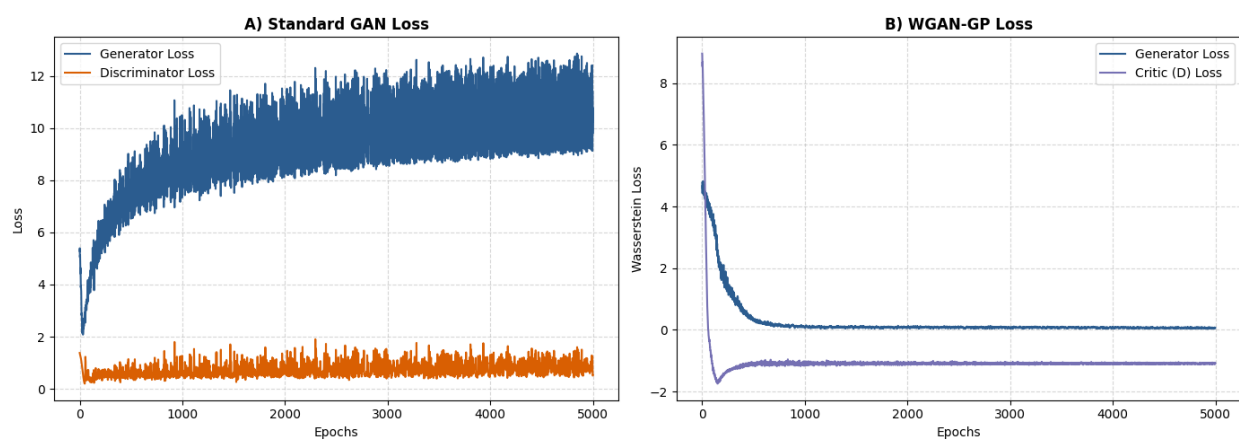


Fig S8. The MSE during the training.

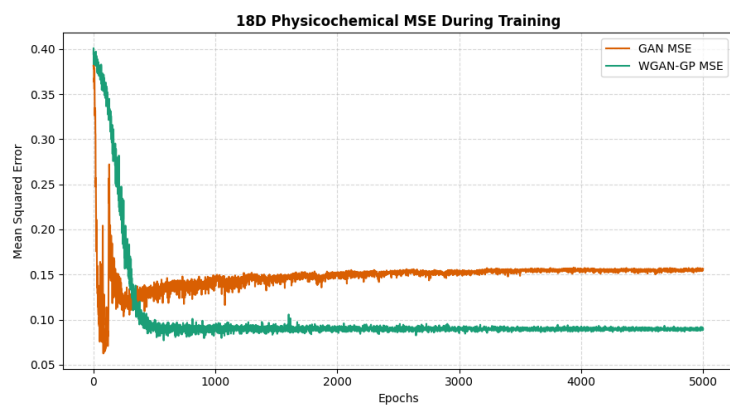


Fig S9. KL Divergence method measure the change in the Bayesian constraint around the generator.

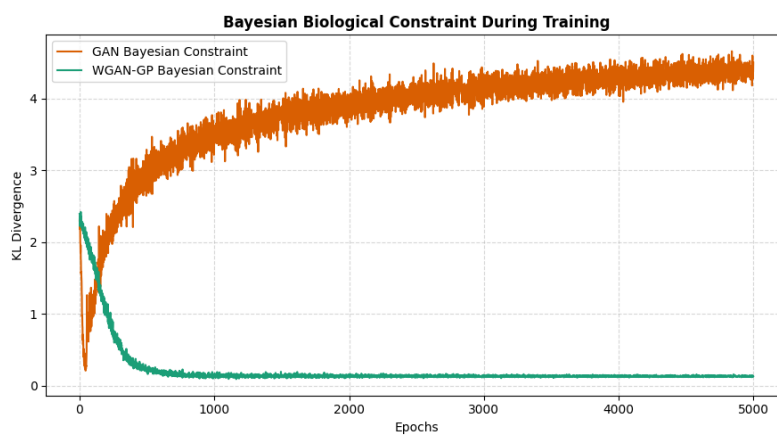


Fig S10. Heatmap of the MREL-like motifs generated by WGAN per position. TREL add as a control.

