Supplementary Material

A. Tree Decomposition

Algorithm 2 presents our tree decomposition of molecules. V_1 and V_2 contain non-ring bonds and simple rings respectively. Simple rings are extracted via RDKit's GetSymmSSSR function. We then merge rings that share three or more atoms as they form bridged compounds. We note that the junction tree of a molecule is not unique when its cluster graph contains cycles. This introduces additional uncertainty for our probabilistic modeling. To reduce such variation, for any of the three (or more) intersecting bonds, we add their intersecting atom as a cluster and remove the cycle connecting them in the cluster graph. Finally, we construct a junction tree as the maximum spanning tree of a cluster graph (\mathcal{V}, \mathcal{E}). Note that we assign an large weight over edges involving clusters in V_0 to ensure no edges in any cycles will be selected into the junction tree.

Algorithm 2 Tree decomposition of molecule G = (V, E)

 $V_1 \leftarrow$ the set of bonds $(u, v) \in E$ that do not belong to any rings.

 $V_2 \leftarrow$ the set of simple rings of G.

for r_1, r_2 in V_2 do

Merge rings r_1, r_2 into one ring if they share more than two atoms (bridged rings).

end for

 $V_0 \leftarrow$ atoms being the intersection of three or more clusters in $V_1 \cup V_2$.

 $\mathcal{V} \leftarrow V_0 \cup V_1 \cup V_2$

 $\mathcal{E} \leftarrow \{(i, j, c) \in \mathcal{V} \times \mathcal{V} \times \mathbb{R} \mid |i \cap j| > 0\}$. Set $c = \infty$ if $i \in V_0$ or $j \in V_0$, and c = 1 otherwise. **Return** The maximum spanning tree over cluster graph $(\mathcal{V}, \mathcal{E})$.





B. Stereochemistry

Though usually presented as two-dimensional graphs, molecules are three-dimensional objects, i.e. molecules are defined not only by its atom types and bond connections, but also the spatial configuration between atoms (chiral atoms and cis-trans isomerism). *Stereoisomers* are molecules that have the same 2D structure, but differ in the 3D orientations of their atoms in space. We note that stereochemical feasibility could not be simply encoded as context free or attribute grammars.

Empirically, we found it more efficient to predict the stereochemical configuration separately from the molecule generation. Specifically, the JT-VAE first generates the 2D structure of a molecule m, following the same procedure described in section 2. Then we generate all its stereoisomers S_m using RDKit's EnumerateStereoisomers function, which identifies atoms that could be chiral. For each isomer $m' \in S_m$, we encode its graph representation $\mathbf{h}_{m'}$ with the graph encoder and compute their cosine similarity $f^s(m') = \cos(\mathbf{h}_{m'}, \mathbf{z}_m)$ (note that \mathbf{z}_m is stochastic). We reconstruct the final 3D structure by picking the stereoisomer $\hat{m} = \arg \max_{m'} f^s(m')$. Since on average only few atoms could have stereochemical variations, this post ranking process is very efficient. Combining this with tree and graph generation, the molecule reconstruction loss \mathcal{L} becomes

$$\mathcal{L} = \mathcal{L}_c + \mathcal{L}_g + \mathcal{L}_s; \qquad \mathcal{L}_s = f^s(m) - \log \sum_{m' \in \mathcal{S}_m} \exp(f^s(m'))$$
(17)

C. Training Details

By applying tree decomposition over 240K molecules in ZINC dataset, we collected our vocabulary set \mathcal{X} of size $|\mathcal{X}| = 780$. The hidden state dimension is 450 for all modules in JT-VAE and the latent bottleneck dimension is 56. For the graph encoder, the initial atom features include its atom type, degree, its formal charge and its chiral configuration. Bond feature is a concatenation of its bond type, whether the bond is in a ring, and its cis-trans configuration. For our tree encoder, we represent each cluster with a neural embedding vector, similar to word embedding for words. The tree and graph decoder use the same feature setting as encoders. The graph encoder and decoder runs three iterations of neural message passing. For fair comparison to SMILES based method, we minimized feature engineering. We use PyTorch to implement all neural components and RDKit to process molecules.

D. More Experimental Results

Sampled Molecules Note that a degenerate model could also achieve 100% prior validity by keep generating simple structures like chains. To prove that our model does not converge to such trivial solutions, we randomly sample and plot 250 molecules from prior distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. As shown in Figure 10, our sampled molecules present rich variety and structural complexity. This demonstrates the soundness of the prior validity improvement of our model.

Neighborhood Visualization Given a molecule, we follow Kusner et al. (2017) to construct a grid visualization of its neighborhood. Specifically, we encode a molecule into the latent space and generate two random orthogonal unit vectors as two axis of a grid. Moving in combinations of these directions yields a set of latent vectors and we decode them into corresponding molecules. In Figure 11 and 12, we visualize the local neighborhood of two molecules presented in Dai et al. (2018). Figure 11 visualizes the same molecule in Figure 6, but with wider neighborhood ranges.

Bayesian Optimization We directly used open sourced implementation in Kusner et al. (2017) for Bayesian optimization (BO). Specifically, we train a sparse Gaussian process with 500 inducing points to predict properties of molecules. Five iterations of batch BO with expected improvement heuristic is used to propose new latent vectors. In each iteration, 50 latent vectors are proposed, from which molecules are decoded and added to the training set for next iteration. We perform 10 independent runs and aggregate results. In Figure 13, we present the top 50 molecules found among 10 runs using JT-VAE. Following Kusner et al.'s implementation, the scores reported are **normalized** to zero mean and unit variance by the mean and variance computed from training set.

Constrained Optimization For this task, a property predictor F is trained jointly with VAE to predict y(m) = logP(m) - SA(m) from the latent embedding of m. F is a feed-forward network with one hidden layer of dimension 450 followed by tanh activation. To optimize a molecule m, we start with its mean encoding $\mathbf{z}_m^0 = \boldsymbol{\mu}_m$ and apply 80 gradient ascent steps: $\mathbf{z}_m^t = \mathbf{z}_m^{t-1} + \alpha \frac{\partial y}{\partial z}$ with $\alpha = 2.0.$ 80 molecules are decoded from latent vectors $\{\mathbf{z}_m^i\}$ and their property is calculated. Molecular similarity sim(m, m') is calculated via Morgan fingerprint of radius 2 with Tanimoto similarity. For each molecule m, we report the best modified molecule m' with $sim(m, m') > \delta$ for some threshold δ . In Figure 14, we present three groups of modification examples with $\delta = 0.2, 0.4, 0.6$. For each group, we present top three pairs that leads to best improvement y(m') - y(m) as well as one pair decreased property (y(m') < y(m)). This is caused by inaccurate property prediction. From Figure 14, we can see that tighter similarity constraint forces the model to preserve the original structure.

~ <u>}</u> 0	2.0	~~~~	855	an	at a	plor	show	Ind	og
or is	Jog.	à.D	Serge	8,0	50°	x-8	200	0,20	540-0
000	20,0	~	9°-0	orto -	3.0	of the	5 mg	r Gt	dx0
8-20	-8 -0	200	na	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	piap	2-7	90270	Zp	Sport
9.00	Sia	ado	and the	0,5	dr yo	-8-	CH CI	264	50
ond	88+	930	pro	p. Po	040	ari	Roto	Próp	S.C.
, PF.	5 d	a B	rong	000	HAND	300	Ato	850	fro
5 K	frex.	20	and	~P-8	000	20-	ogt	300	300
<u>}</u> ?	~~~~	40	, etc	pmp	30	roa	Roi	~ gud	χ_{Σ}
0 ¹ ~0	poind	2018	50	Ja	2 ² 4+	g.	- pro	9,8	570
363	23	+04	alo	69-0	- Br	sia	borb	-St-S	~00
d Xq	2p	3.x	5p	dr-q	0077	Sc4	\$	33	aro
-otorto	Not	0900	13	-01-00-	or to	A4B	p-co	-2-8-	Dria
40	~ <u>{</u> _}O	200	040	540	Lib	Pro	2to	ang	9 3
and.	ron	904	0~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	43.2	quod	ordor	y a	and	, àp
and	en la constante) ga	之中	oza	3000	-079-	×-So	roz	08
Q7'	- qu	7S	× 000	3nP	690	200	070	-20-0	900
100	ϕ	-bron	2,0	20	groo	440	dis-	50	Era
ho	6-3	500	~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	C.P	A dra	905	do	$\gamma 0 \gamma_{2}^{2}$	2000
D'O'D	918	<u> </u>	ra	0 the	1-00	ortha	80	880	Jose .
nd	f.b	- pt	acco	1 Jac	2004	ada	aw	d, x	for
320	048	-800	d's	50	2,00	->	Sod	-010-	to the second se
pholo	2p	ANO	1040	ça	Sid D	5-8	or d	9. O	400
400	ors	at 2	d-	Sta .	into	10-81	ng	Cr X	d A
200	~~ \ }?	and	-	gro	0,00	quano	jos a	945	22

Figure 10. 250 molecules sampled from prior distribution $\mathcal{N}(\mathbf{0},\mathbf{I}).$

x.A	x fr	2-29	\$ °5	<u>४</u> %	909	909	RSR	R-82	909	409	bar	rong.
a fi	8-X-	800	3.A	\$ 3.	\$ 95	909	909	906	404	bar	bar	bard
8-X	&&	9×	8−2¢	\$ 95	bard	-994	-000	pop	909	dob	bard	407b.
\$-\$	d'as	dog	009	909	bad	ágro	مركسهم	д у в	909	poi	909	doð
9-6	8-8-	dog	409	bad	bad	bad	bad	9.C	J.S	L.S	909	ppa
9-60	\$-B	\$ ⁵ 8-	<u> </u>	bad	bad	bad	bad	bad	bad	L.S	d do	poq
gp-a	p-q	2-60	у ^р У-	bad	bad	bad	bad	bad	bad	bad	doa	pod
h.J	the	bra	bra	bad	sad	bad	bad	bad	bad	bad	trad	pad
h.f	p-o-	yp-a	y - a	& a	bad	bad	bad	bad	bad	roa	rad	803
h.f	à f	ha	हित्	使日	ha	bad	bad	dad	COP	aga	-and	-and
r f	r f	ha	ga	्रदेव	p flo	हिन्द	sad	dad	Roa	xoa	trad	trad
er er	r f	r f	- A	r f	fre-	fra	र्वन्द्	Rap	roa	trad	trad	tad
por	-mB	-orB	word	adora.	fra	fro	àqå	377	it of a	tad	trad	trad

Figure 11. Neighborhood visualization of molecule C[C@H]1CC(Nc2cncc(-c3nncn3C)c2)C[C@H](C)C1.

22	\$a	-og-	<u>क</u> र्वेच्	50	ord	and	and	app	app	para	para	àço
5.0	90-3 ⁰	and	and	tog	304	بمرم	مرم	app	app	dor	dora	.afra
'9°S	'9°0'	Pst	100	tog	jaort	مرم	,dora	30.0	,20.0	das	dora	asp
P34	304	tog	port	304	port	port	jort	30.0	, àcro	àcro	·द [°] र्द	ortan
por	Jor 1	304	port	304	port	port	jort	'goy	304	jao+	न्देर्घ	+020Q
por	304	port	jaort	'goy	port	port	por	port	404	jor	bot	poly
'goy	port	port	port	'goy	port	port	port	port	port	jort	port	bot
\$	port	port	port	304	port	port	jort	jaort	port	'go+	port	***
orde	304	port	jaort	'goy	port	port	por	port	404	jor	port	rdr
৾৵৶৾	993	port	port	jaon	port	port	port	304	port	port	port	1-95
stor.	केर	304	304	304	port	Jor	304	304	por	304	odor	1-95
stor.	for	gor.	zor	304	port	304	jaor	port	pox	adar	order	1-95
Story	62° 4	8CC 4	Bor	or y	'go+	por	yory	jocy	Joc 4	dory	292	1-95

Figure 12. Neighborhood visualization of molecule COc1cc(OC)cc([C@H]2CC[NH+](CCC(F)(F)F)C2)c1.



Figure 13. Top 50 molecules found by Bayesian optimization using JT-VAE.



Figure 14. Row 1-3: Molecule modification results with similarity constraint $sim(m, m') \ge 0.2, 0.4, 0.6$. For each group, we plot the top three pairs that leads to actual property improvement, and one pair with decreased property. We can see that tighter similarity constraint forces the model to preserve the original structure.