(Supplementary) Deterministic Decoding for Discrete Data in Variational Autoencoders

Daniil Polykovskiy
Insilico Medicine

Dmitry Vetrov
National Research University Higher School of Economics

A Proof of Theorem 1

We prove the theorem using five lemmas.

Lemma 1. *$\mathcal{L}_\tau$ convergences to $\mathcal{L}_*$ pointwise when τ converges to 0 from the right:*

$$\forall(\theta, \phi) \quad \lim_{\tau \to 0+} \mathcal{L}_\tau(\theta, \phi) = \mathcal{L}_*(\theta, \phi) \tag{1}$$

Proof. To prove Eq.1, we first show that our approximation in Eq.10 from the main paper converges pointwise to $\mathbb{I}\left[x > 0\right]$. $\forall x \in \mathbb{R}$:

$$\lim_{\tau \to 0+} \sigma_\tau(x) = \lim_{\tau \to 0+} \frac{1}{1 + e^{-x/\tau}\left[\frac{1}{\tau} - 1\right]} = \mathbb{I}\left[x > 0\right] \tag{2}$$

If x is negative, both $e^{-x/\tau}$ and $1/\tau$ converge to $+\infty$, hence $\sigma_\tau(x)$ converges to zero. If x is zero, then $\sigma_\tau(x) = \tau$ which also converges to zero. Finally, for positive x we apply L'Hôpital's rule to compute the limit:

$$\lim_{\tau \to 0+} \frac{e^{-x/\tau}}{\tau} = \lim_{\tau \to 0+} \frac{(1/\tau)'}{\left(e^{x/\tau}\right)'} = \lim_{\tau \to 0+} \frac{e^{-x/\tau}}{x} = 1 \tag{3}$$

To prove the theorem, we consider two cases. First, if $(\theta, \phi) \notin \Omega$, then for some x, i, and $x \neq s$,

$$\mathbb{E}_{z \sim q_\phi(z|x)} \mathbb{I}\left[\widetilde{\pi}^\theta_{x,i,x_i}(z) \leq \widetilde{\pi}^\theta_{x,i,s}(z)\right] > 0. \tag{4}$$

From the equation above follows that for given parameters the model violates indicators with positive probability. For those z, a smoothed indicator function takes values less than τ, so the expectation of its logarithm tends to $-\infty$ when $\tau \to 0+$.

The second case is $(\theta, \phi) \in \Omega$. Since $\mathcal{L}_*(\theta, \phi) > -\infty$, indicators are violated only with probability zero, which will not contribute to the loss neither in $\mathcal{L}_*$, nor in $\mathcal{L}_\tau$. For all x, i and s, consider a distribution of a random variable $\delta = \widetilde{\pi}^\theta_{x,i,x_i}(z) - \widetilde{\pi}^\theta_{x,i,s}(z)$ obtained from a distribution $q_\phi(z \mid x)$. Let $\delta_{\max} \leq 1$ be the maximal value of δ. We now need to prove that

$$\lim_{\tau \to 0+} \mathbb{E}_{\delta \sim p(\delta)} \log \sigma_\tau(\delta) = 0 \tag{5}$$

For any $\epsilon > 0$, we select $\delta_0 > 0$ such that $p(\delta < \delta_0) < \epsilon$. For the next step we will use the fact that $\sigma_\tau\left(\delta_{1/2}\right) = 0.5$, where $\delta_{1/2} = \tau \log\left(\frac{1}{\tau} - 1\right)$. By selecting τ small enough such that $\delta_{1/2} < \delta_0$, we split the integration limit for δ in expectation into three segments: $(0, \delta_{1/2}]$, $(\delta_{1/2}, \delta_0]$, $(\delta_0, \delta_{\max})$. A lower bound on $\log \sigma_\tau(\delta)$ in each segment is given by its value in the left end: $\log \tau$, $\log 1/2$, $\log \sigma_\tau(\delta_0)$. Also, since $p(\delta \leq 0) = 0$ and δ is continuous on compact support of $q_\phi(z \mid x)$, density $p(\delta)$ is bounded by some constant M. Such estimation gives us the final lower bound using pointwise convergence of $\sigma_\tau(\delta)$:

$$\begin{aligned}0 \geq \mathbb{E}_{\delta \sim p(\delta)} \log \sigma_\tau(\delta) \geq& \\ M \cdot \underbrace{\log \tau \cdot \delta_{1/2}}_{\lim_{\tau \to 0+} \cdots = 0} &+ \epsilon \cdot \log 1/2 \\ + M \cdot \underbrace{\log \sigma_\tau(\delta_0)}_{\lim_{\tau \to 0+} \cdots = 0} &\cdot (\delta_{\max} - \delta) \to_{\tau \to 0+} \epsilon \cdot \log 1/2.\end{aligned} \tag{6}$$

We used $\lim_{\tau \to 0+} \log \tau \cdot \delta_{1/2} = 0$ which can be proved by applying the L'Hôpital's rule twice. ☐

Proposition 1. *For our model, $\mathcal{L}_*$ is finite if and only if a sequence-wise reconstruction error rate is zero:*

$$(\theta, \phi) \in \Omega \Leftrightarrow \Delta(\widetilde{x}_\theta, \phi) = 0 \tag{7}$$

Lemma 2. *Sequence-wise reconstruction error rate $\Delta(\phi)$ is continuous.*

Proof. Following equicontinuity in total variation of $q_\phi(z \mid x)$ at ϕ for any x and finiteness of χ, for any $\epsilon > 0$ there exists $\delta > 0$ such that for any $x \in \chi$ and any ϕ' such that $\|\phi - \phi'\| < \delta$

$$\int |q_\phi(z|x) - q_{\phi'}(z|x)|\, dz < \epsilon. \tag{8}$$

For parameters ϕ and ϕ', we estimate the difference in

Δ function values

$$\begin{aligned}\Delta(\phi) - \Delta(\phi') \\ &= \underbrace{\Delta(\widetilde{x}^*_{\phi}, \phi) - \Delta(\widetilde{x}^*_{\phi'}, \phi)}_{\leq 0} + \Delta(\widetilde{x}^*_{\phi'}, \phi) - \Delta(\widetilde{x}^*_{\phi'}, \phi') \\ &\leq \mathbb{E}_{x\sim p(x)} \underbrace{\int \left(q_\phi(z|x) - q_{\phi'}(z|x)\right) \mathbb{I}\left[\widetilde{x}^*_{\phi'}(z) \neq x\right] dz}_{<\epsilon} \\ &\leq \epsilon \end{aligned} \tag{9}$$

Symmetrically, $\Delta(\phi') - \Delta(\phi) \leq \epsilon$, resulting in $\Delta(\phi)$ being continuous. □

Lemma 3. *Sequence-wise reconstruction error rate* $\Delta(\phi_n)$ *converges to zero:*

$$\lim_{n\to+\infty} \Delta(\phi_n) = \Delta(\widetilde{\phi}) = 0. \tag{10}$$

The convergence rate is $\mathcal{O}(\frac{1}{\log(1/\tau_n)})$.

Proof. Since Ω is not empty, there exists $(\widehat{\theta}, \widehat{\phi}) \in \Omega$. From pointwise convergence of $\mathcal{L}_\tau$ to $\mathcal{L}_*$ at point $(\widehat{\theta}, \widehat{\phi})$, for any $\epsilon > 0$ exists N such that for any $n > N$:

$$\underbrace{\mathcal{L}_{\tau_n}(\theta_n, \phi_n) \geq \mathcal{L}_{\tau_n}(\widehat{\theta}, \widehat{\phi})}_{\text{from the definition of } (\theta_n, \phi_n)} \geq \mathcal{L}_*(\widehat{\theta}, \widehat{\phi}) - \epsilon. \tag{11}$$

Next, we derive an upper bound on $\mathcal{L}_{\tau_n}(\theta_n, \phi_n)$ using the fact that $\log\sigma_\tau(x) < 0$ if $x > 0$, and $\log\sigma_\tau(x) \leq \log\tau_n$ if $x \leq 0$:

$$\begin{aligned}\mathcal{L}_{\tau_n}(\theta_n, \phi_n) &\leq \mathbb{E}_{x\sim p(x)}\Bigg[\mathbb{E}_{z\sim q_\phi(z|x)} \sum_{i=1}^{|x|} \sum_{s\neq x_i} \log\tau_n \cdot \\ &\quad \mathbb{I}\left[\pi_{x,i,x_i}(z) \leq \pi_{x,i,s}(z)\right] \underbrace{- \mathcal{KL} q_\phi(z \mid x) p(z)}_{\leq 0}\Bigg] \\ &\leq |V| L \cdot \log\tau_n \cdot \Delta(\widetilde{x}_{\theta_n}, \phi_n). \end{aligned} \tag{12}$$

Combining Eq. 11 and Eq. 12 together we get

$$|V|L \cdot \underbrace{\log\tau_n}_{<0} \cdot \Delta(\widetilde{x}_{\theta_n}, \phi_n) \geq \mathcal{L}_*(\theta^*, \phi^*) - \epsilon \tag{13}$$

Adding the defintion of $\Delta(\phi)$, we obtain

$$0 \leq \Delta(\phi_n) \leq \Delta(\widetilde{x}_{\theta_n}, \phi_n) \leq \frac{\epsilon - \mathcal{L}_*(\theta^*, \phi^*)}{|V|L \cdot \log(1/\tau_n)} \tag{14}$$

The right hand side goes to zero when n goes to infinity and hence $\lim_{n\to+\infty} \Delta(\widetilde{x}_{\theta_n}, \phi_n) = 0$ and $\lim_{n\to+\infty} \Delta(\phi_n) = 0$ with the convergence rate $\mathcal{O}(\frac{1}{\log(1/\tau_n)})$. Since $\Delta(\phi_n)$ is continuous, $\Delta(\widetilde{\phi}) = 0$. □

Lemma 4. $\mathcal{L}_*(\theta, \phi)$ *attains its supremum:*

$$\exists \theta^* \in \Theta, \phi^* \in \Phi : \mathcal{L}_*(\theta^*, \phi^*) = \sup_{\theta\in\Theta, \phi\in\Phi} \mathcal{L}_*(\theta, \phi). \tag{15}$$

Proof. From Lemma 3, $\Delta(\widetilde{\phi}) = 0$. Hence, for a choice of $\widetilde{\theta}$ from the theorem statement, $\Delta(\widetilde{\theta}, \widetilde{\phi}) = 0$. Equivalently, $(\widetilde{\theta}, \widetilde{\phi}) \in \Omega$.

Note that since $\Delta(\phi) \geq 0$ is continuous on a compact set, $\Phi_0 = \{\phi \mid \Delta(\phi) = 0\}$ is a compact set. Also, $\mathcal{L}_*(\theta, \phi)$ is constant with respect to θ on Ω. From the theorem statement, for any ϕ such that $\Delta(\phi) = 0$, there exists $\theta(\phi)$ such that $(\theta(\phi), \phi) \in \Omega$. Combining all statements together,

$$\sup_{\phi\in\Phi_0} \mathcal{L}_*(\theta(\phi), \phi) = \sup_{\theta\in\Theta, \phi\in\Phi} \mathcal{L}_*(\theta, \phi) \tag{16}$$

In Ω, $\mathcal{L}_*$ is a continuous function: $\forall(\theta, \phi) \in \Omega$,

$$\mathcal{L}_*(\theta, \phi) = -\mathcal{KL}(\phi) = -\mathbb{E}_{x\sim p(x)} \mathcal{KL}\left(q_\phi(z|x) \,\|\, p(z)\right) \tag{17}$$

Hence, continuous function $\mathcal{L}_*(\theta(\phi), \phi)$ attains its supremum on a compact set Φ at some point (θ^*, ϕ^*), where $\theta^* = \theta(\phi^*)$. □

Lemma 5. *Parameters* $(\widetilde{\theta}, \widetilde{\phi})$ *from theorem statement are optimal:*

$$\mathcal{L}_*(\widetilde{\theta}, \widetilde{\phi}) = \sup_{\theta\in\Theta, \phi\in\Phi} \mathcal{L}_*(\theta, \phi). \tag{18}$$

Proof. Assume that $\mathcal{L}_*(\widetilde{\theta}, \widetilde{\phi}) < \mathcal{L}_*(\theta^*, \phi^*)$. Since $(\widetilde{\theta}, \widetilde{\phi}) \in \Omega$ and $(\theta^*, \phi^*) \in \Omega$, $\mathcal{L}_*(\widetilde{\theta}, \widetilde{\phi}) = -\mathcal{KL}(\widetilde{\phi})$ and $\mathcal{L}_*(\theta^*, \phi^*) = -\mathcal{KL}(\phi^*)$. As a result, from our assumption, $\mathcal{KL}(\phi^*) < \mathcal{KL}(\widetilde{\phi})$.

From continuity of $\mathcal{KL}(\phi)$ divergence, for any $\epsilon > 0$, exists $\delta > 0$ such that if $\|\widetilde{\phi} - \phi\| < \delta$,

$$\mathcal{KL}(\phi) > \mathcal{KL}(\widetilde{\phi}) - \epsilon = \mathcal{L}_*(\widetilde{\theta}, \widetilde{\phi}) - \epsilon \tag{19}$$

From the convergence of ϕ_n to $\widetilde{\phi}$ and convergence of τ_n to zero, there exists N_1 such that for any $n > N_1$, $\|\widetilde{\phi} - \phi_n\| < \delta$.

From pointwise convergence of $\mathcal{L}_{\tau_n}$ at point (θ^*, ϕ^*) to $\mathcal{L}_*(\theta^*, \phi^*)$, for any $\epsilon > 0$, exists N_2 such that for all $n > N_2$, $\mathcal{L}_{\tau_n}(\theta^*, \phi^*) > \mathcal{L}_*(\theta^*, \phi^*) - \epsilon$. Also, $\mathcal{L}_{\tau_n}(\theta_n, \phi_n) \leq -\mathcal{KL}(\phi_n)$ from the definition of $\mathcal{L}_{\tau_n}$ as a negative $\mathcal{KL}$ divergence plus some non-positive penalty for reconstruction error.

Taking $n > \max(N_1, N_2)$, we get the final chain of inequalities:

$$\begin{aligned}\mathcal{L}_{\tau_n}(\theta_n, \phi_n) &\leq -\mathcal{KL}(\phi_n) < -\mathcal{KL}(\widetilde{\phi}) + \epsilon \\ &= \mathcal{L}_*(\widetilde{\theta}, \widetilde{\phi}) + \epsilon < \mathcal{L}_{\tau_n}(\theta^*, \phi^*) - \epsilon + \epsilon \\ &= \mathcal{L}_{\tau_n}(\theta^*, \phi^*) \end{aligned} \tag{20}$$

Hence, $\mathcal{L}_{\tau_n}(\theta_n, \phi_n) < \mathcal{L}_{\tau_n}(\theta^*, \phi^*)$, which contradicts $(\theta_n, \phi_n) \in \operatorname{Arg\,max}$ of $\mathcal{L}_{\tau_n}$. As a result, $\mathcal{L}_*(\widetilde{\theta}, \widetilde{\phi}) = \mathcal{L}_*(\theta^*, \phi^*)$. $\square$

B Implementation details

For all experiments, we provide configuration files in a human-readable format in the supplementary code. Here we provide the same information for convenience.

B.1 Synthetic data

Encoder and decoder were GRUs with 2 layers of 128 neurons. The latent size was 2; embedding dimension was 8. We trained the model for 100 epochs with Adam optimizer with an initial learning rate $5 \cdot 10^{-3}$, which halved every 20 epochs. The batch size was 512. We fine-tuned the model for 10 epochs after training by fixing the encoder and learning only the decoder. For a proposed model with a uniform prior and a uniform proposal, we increased $\mathcal{KL}$ weight β linearly from 0 to 0.1 during 100 epochs. For the Gaussian and tricube proposals, we increased $\mathcal{KL}$ weight β linearly from 0 to 1 during 100 epochs. For all three experiments, we pretrained the autoencoder for the first two epochs with $\beta = 0$. We annealed the temperature from 10^{-1} to 10^{-3} during 100 epochs of training in a log-linear scale. For a tricube proposal, we annealed the temperature to 10^{-2}.

B.2 Binary MNIST

We binarized the dataset by thresholding original MNIST pixels with a value of 0.3. We used a fully connected neural network with layer sizes $784 \to 256 \to 128 \to 32 \to 2$ with LeakyReLU activation functions. We trained the model for 150 epochs with a starting learning rate $5 \cdot 10^{-3}$ that halved every 20 epochs. We used a batch size 512 and clipped the gradient with value 10. We increased β from 10^{-5} to 0.005 for VAE and 0.05 for DD-VAE. We decreased the temperature in a log scale from 0.01 to 0.0001.

B.3 MOSES

We used a 2-layer GRU network with a hidden size of 512. Embedding size was 64, the latent space was 64-dimensional. We used a tricube proposal and a Gaussian prior. We pretrained a model with a fixed β for 20 epochs and then linearly increased β for 180 epochs. We halved the learning rate after pretraining. For DD-VAE models, we decreased the temperature in a log scale from 0.2 to 0.1. We linearly increased β divergence from 0.0005 to 0.01 for VAE models and from 0.0015 to 0.02.

B.4 ZINC

We used a 1-layer GRU network with a hidden size of 1024. Embedding size was 64, the latent space was 64-dimensional. We used a tricube proposal and a Gaussian prior. We trained a model for 200 epochs with a starting learning rate $5 \cdot 10^{-4}$ that halved every 50 epochs. We increased divergence weight β from 10^{-3} to 0.02 linearly during the first 50 epochs for DD-VAE models, from 10^{-4} to $5 \cdot 10^{-4}$ for VAE model, and from 10^{-4} to $8 \cdot 10^{-4}$ for VAE model with a tricube proposal. We decreased the temperature log-linearly from 10^{-3} to 10^{-4} during the first 100 epochs for DD-VAE models. With such parameters we achieved a comparable train sequence-wise reconstruction accuracy of 95%.

C MOSES distribution learning

In Figure 1, we report detailed results for the experiment from Section 4.3.

D Best molecules found for ZINC

In Figure 2, Figure 3, Figure 4, and Figure 5 we show the best molecules found with Bayesian optimization during 10-fold cross validation.

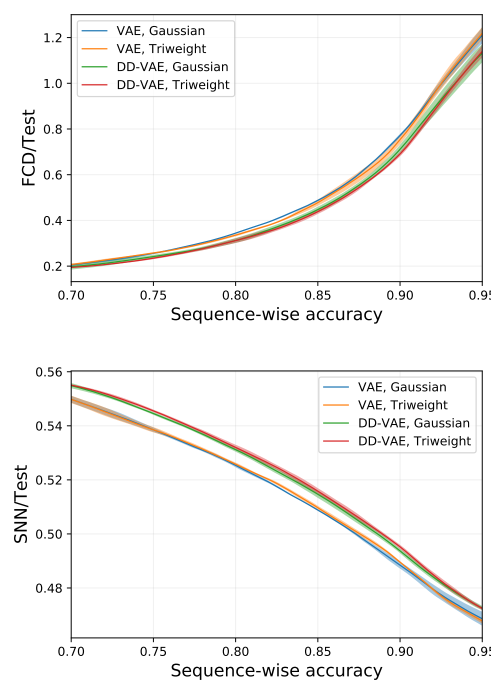

Figure 1: Distribution learning with deterministic decoding on MOSES dataset: FCD/Test (lower is better) and SNN/Test (higher is better). Solid line: mean, shades: std over multiple runs.

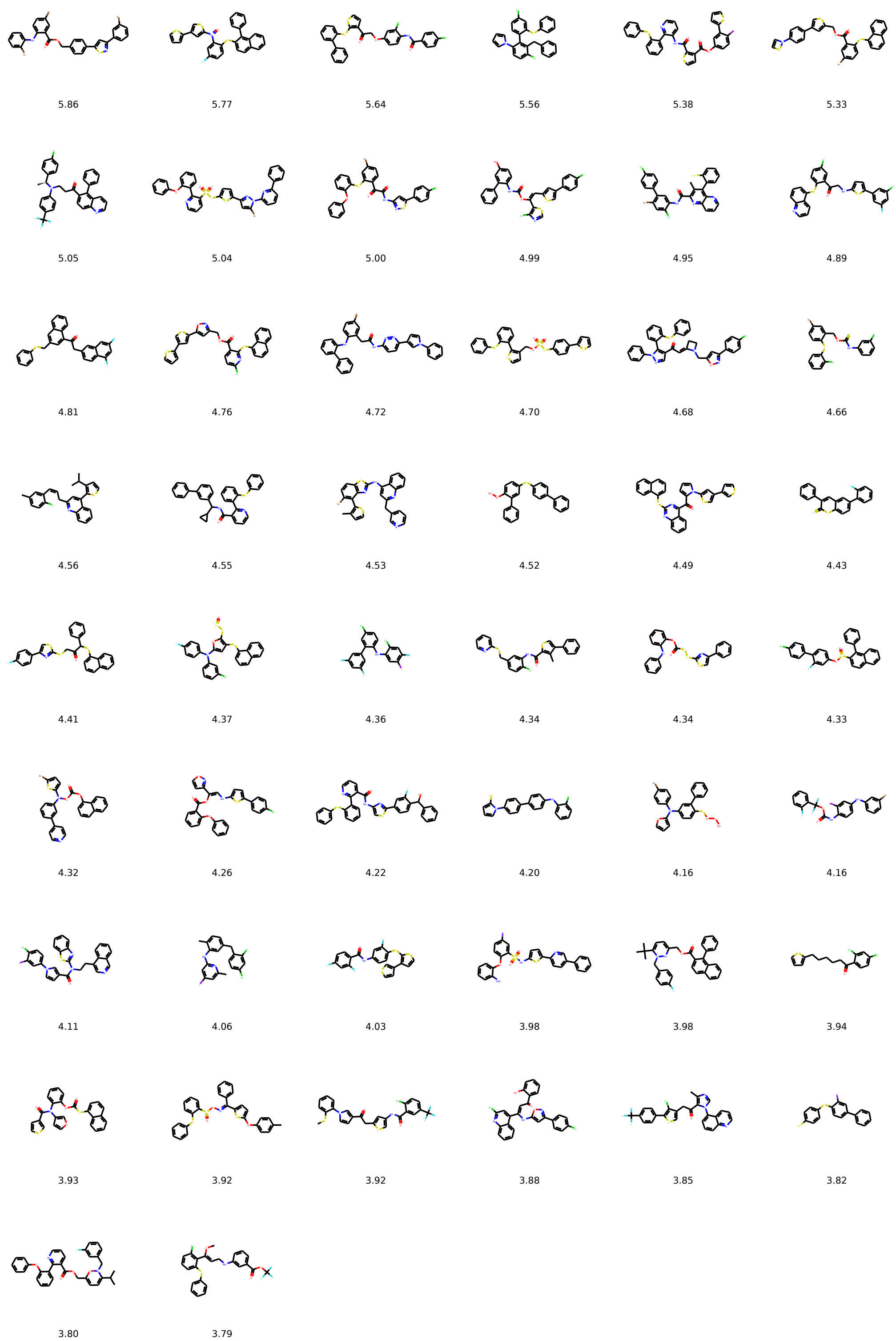

Figure 2: DD-VAE with Tricube proposal. The best molecules found with Bayesian optimization during 10-fold cross validation and their scores.

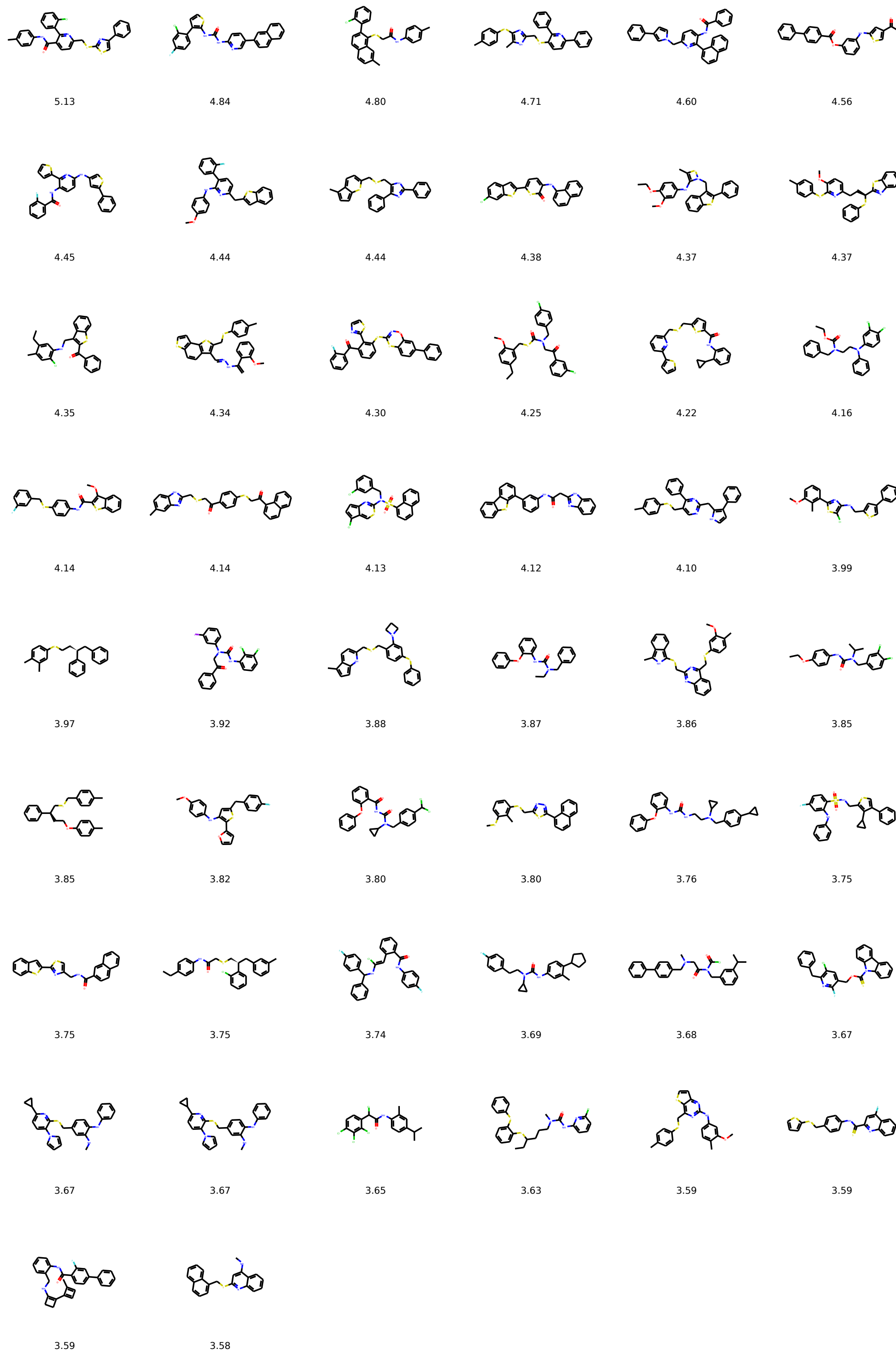

Figure 3: DD-VAE with Gaussian proposal. The best molecules found with Bayesian optimization during 10-fold cross validation and their scores.

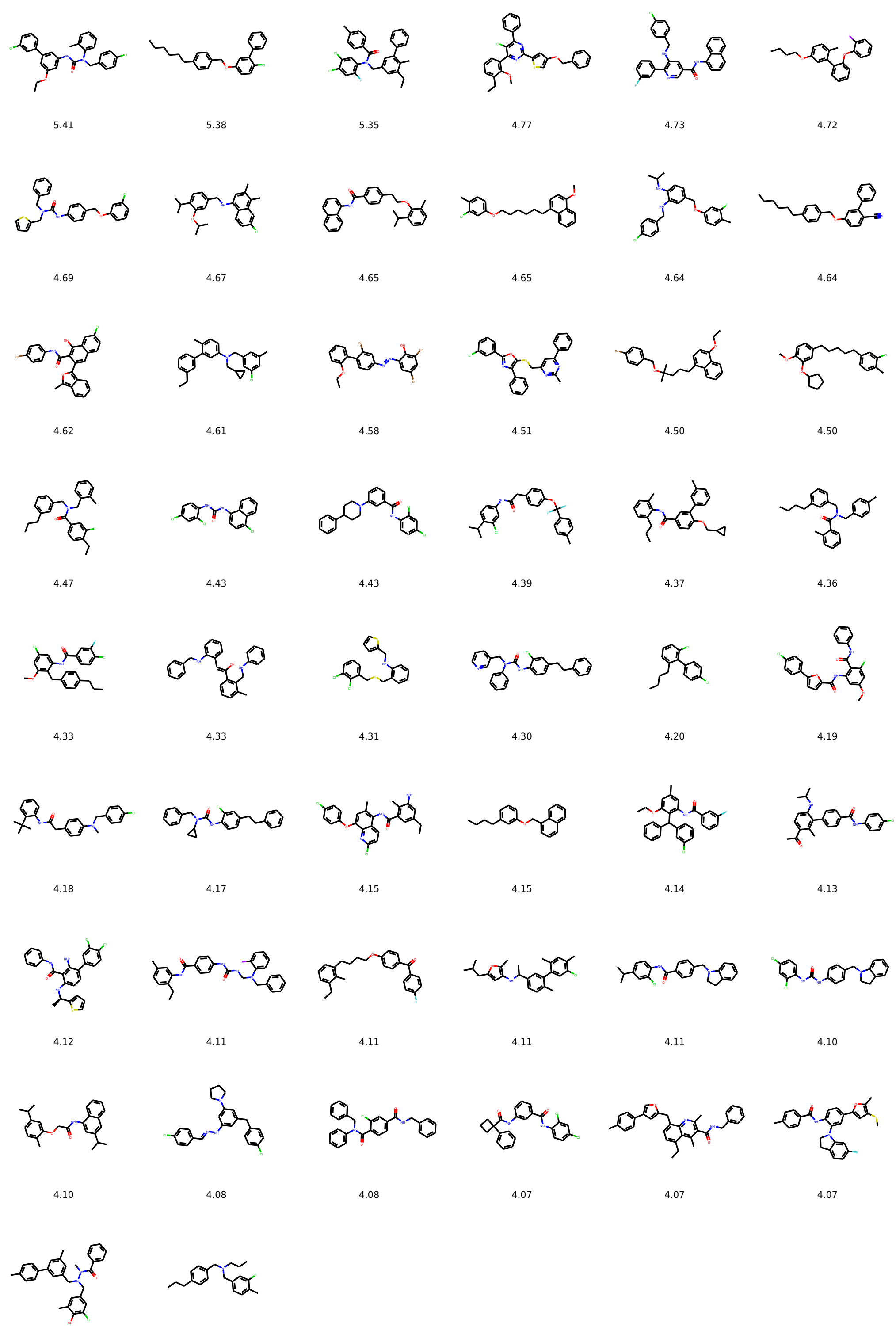

Figure 4: VAE with Tricube proposal. The best molecules found with Bayesian optimization during 10-fold cross validation and their scores.

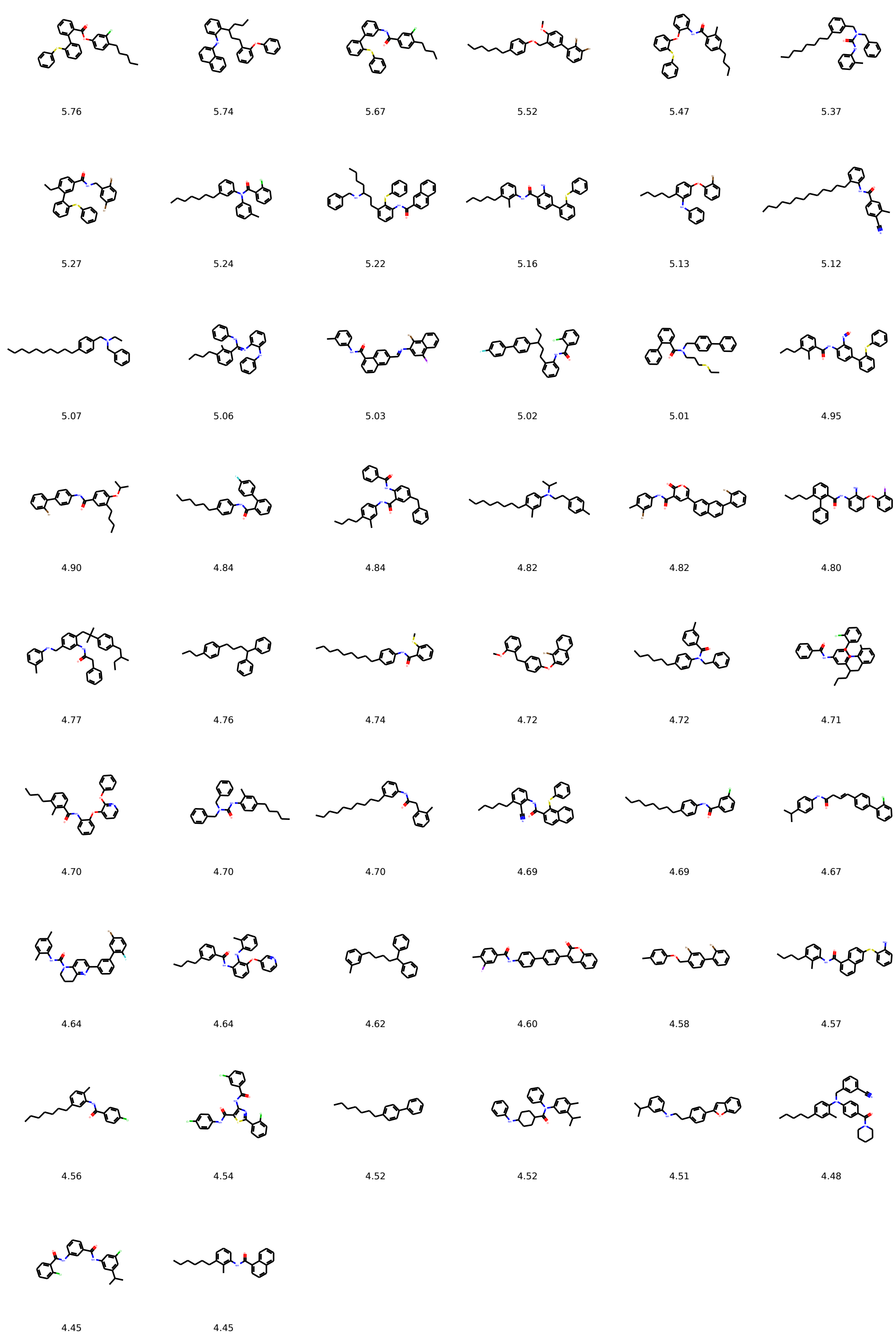

Figure 5: VAE with Gaussian proposal. The best molecules found with Bayesian optimization during 10-fold cross validation and their scores.