Supporting Information for Automatic Chemical Design Using a Data-Driven Continuous Representation of Molecules

Rafael Gómez-Bombarelli,[†,#] Jennifer N. Wei,[‡,#] David Duvenaud,[¶,#] José Miguel Hernández-Lobato,[§,#] Benjamín Sánchez-Lengeling,[‡] Dennis Sheberla,[‡] Jorge Aguilera-Iparraguirre,[†] Timothy D. Hirzel,[†] Ryan P. Adams,[‖] and Alán Aspuru-Guzik[*,‡,⊥]

Kyulux North America Inc., 10 Post Office Sq., Suite 800, Boston, MA 02109, USA, Department of Chemistry and Chemical Biology, Harvard University, Cambridge MA 02138, USA, Department of Computer Science, University of Toronto, Department of Engineering, University of Cambridge Trumpington Street, Cambridge CB2 1PZ, UK, Google Brain and Princeton University, and Canadian Institute for Advanced Research (CIFAR), Biologically-Inspired Solar Energy Program.

E-mail: aspuru@chemistry.harvard.edu

[h]

*To whom correspondence should be addressed
†Kyulux North America Inc., 10 Post Office Sq., Suite 800, Boston, MA 02109, USA
‡Department of Chemistry and Chemical Biology, Harvard University, Cambridge MA 02138, USA
¶Department of Computer Science, University of Toronto
§Department of Engineering, University of Cambridge Trumpington Street, Cambridge CB2 1PZ, UK
‖Google Brain and Princeton University
⊥Canadian Institute for Advanced Research (CIFAR), Biologically-Inspired Solar Energy Program.
#Equal contribution

Table 1: Percentage of successfully decoding of latent representation after 1000 attempts for 1000 molecules from the traning set, 1000 validation molecules randomly chosen from ZINC and a 1000 validation molecules randomly chosen from eMolecules. Both VAEs perform very well for training data, and they are well transferable within molecules of the same class outside the training data, as evidence by the good validation performance of the ZINC VAE and the underperformance of the QM9 VAE against real-life small molecules.

Dataset	ZINC	QM
Training set	92.1	99.6
Test set	90.7	99.4
ZINC	91.0	1.4
eMolecules	83.8	8.8

Table 2: Percentage of 5000 randomly-selected latent points that decode to valid molecules after 1000 attempts

Dataset	ZINC	QM
Decoding probability	73.9	79.3

Table 3: Variational autoencoder performance over different sizes of datasets. Training and tests were performed using randomly selected molecules from the ZINC dataset, the values reported here are the scores from the validation set. The categorical accuracy reflects the percentage of characters in the output SMILES that were accurately reconstructed. Mean Absolute Errors (MAE) are reported for QED and logP properties. Performance significantly decreases if only 10^5 molecules are used for training.

Training set size	Categorical Accuracy	logP MAE	QED MAE
225,000	99.3%	0.15	0.054
175,000	99.0%	0.18	0.076
125,000	98.5%	0.15	0.076
25,000	91.6%	0.23	0.079

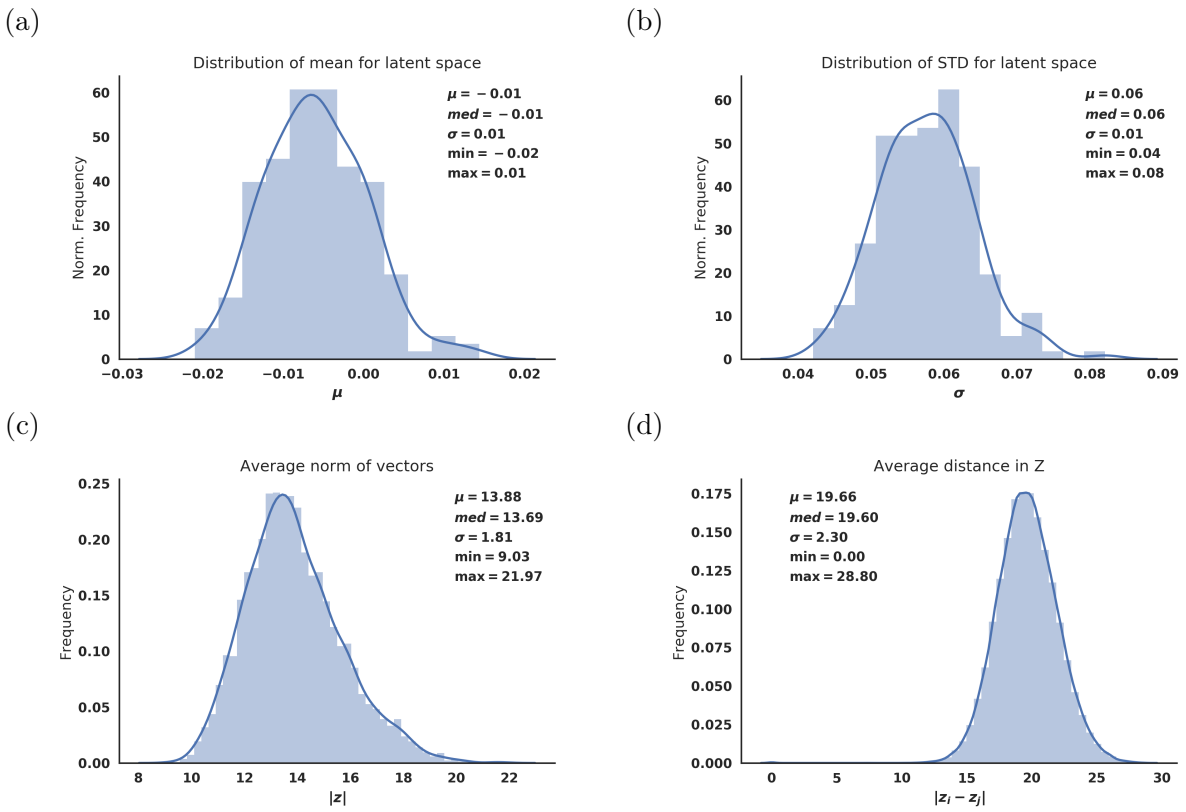

Figure 1: Distribution and statistics of (a) the mean of latent space coordinates (b) standard deviation of latent space coordinates (c) norm of latent space coordinates of the encoded representation of randomly selected molecules from the ZINC validation set. (d) Distribution of Euclidean distances between random pairs of validation molecules in the ZINC VAE

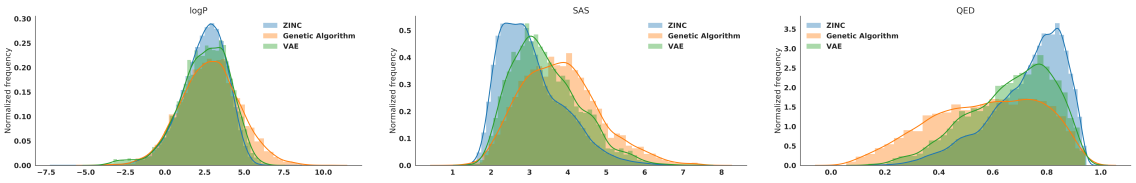

Figure 2: Histograms and KDE plots of the distribution of properties utilized in the jointly trained autoencoder (LogP, SAS, QED). Used to further showcase results from Table 2. For each property we compare the distribution of the source data (ZINC), a generatic algorithm and the VAE.

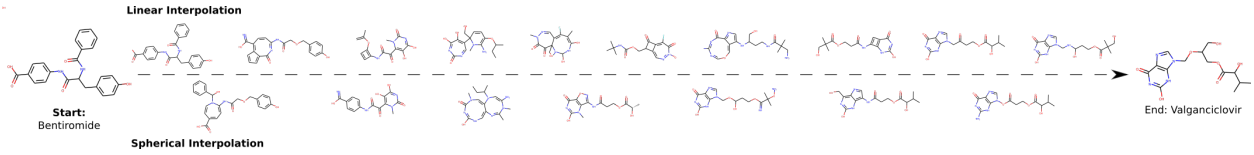

Figure 3: Comparison of between linear and spherical interpolation paths between two randomly selected FDA approved drugs. A constant step size was used.

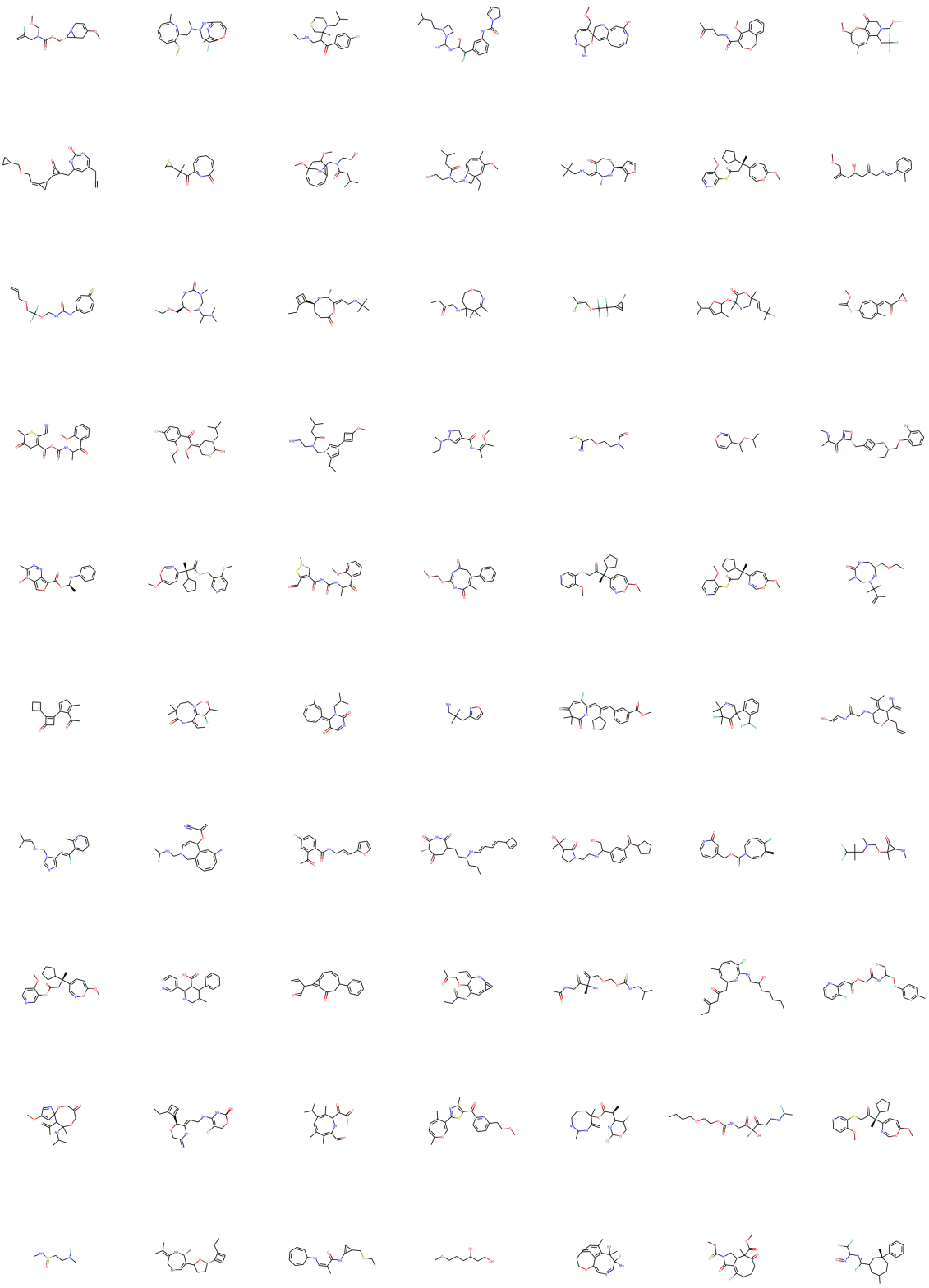

Figure 4: Molecules decoded from randomly-sampled points in the latent space of the ZINC VAE.