

## REFERENCES

- Guy W Bemis and Mark A Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of medicinal chemistry*, 39(15):2887–2893, 1996.
- G Richard Bickerton, Gaia V Paolini, Jérémy Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.
- Thomas Blaschke, Marcus Olivecrona, Ola Engkvist, Jürgen Bajorath, and Hongming Chen. Application of generative autoencoder in de novo molecular design. *Molecular informatics*, 37(1-2):1700123, 2018.
- Danail Bonchev. *Chemical graph theory: introduction and fundamentals*, volume 1. CRC Press, 1991.
- Nathan Brown, Marco Fiscato, Marwin HS Segler, and Alain C Vaucher. Guacamol: benchmarking models for de novo molecular design. *Journal of chemical information and modeling*, 59(3):1096–1108, 2019.
- Bin Dai and David Wipf. Diagnosing and enhancing vae models. *arXiv preprint arXiv:1903.05789*, 2019.
- Hanjun Dai, Yingtao Tian, Bo Dai, Steven Skiena, and Le Song. Syntax-directed variational autoencoder for structured data. *arXiv preprint arXiv:1802.08786*, 2018.
- Jörg Degen, Christof Wegscheid-Gerlach, Andrea Zaliani, and Matthias Rarey. On the art of compiling and using ‘drug-like’ chemical fragment spaces. *ChemMedChem: Chemistry Enabling Drug Discovery*, 3(10):1503–1507, 2008.
- David Duvenaud, Dougal Maclaurin, Jorge Aguilera-Iparraguirre, Rafael Gómez-Bombarelli, Timothy Hirzel, Alán Aspuru-Guzik, and Ryan P Adams. Convolutional networks on graphs for learning molecular fingerprints. *arXiv preprint arXiv:1509.09292*, 2015.
- Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):1–11, 2009.
- Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.
- Hao Fu, Chunyuan Li, Xiaodong Liu, Jianfeng Gao, Asli Celikyilmaz, and Lawrence Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing. *arXiv preprint arXiv:1903.10145*, 2019.
- Niklas Gebauer, Michael Gastegger, and Kristof Schütt. Symmetry-adapted generation of 3d point sets for the targeted discovery of molecules. *Advances in neural information processing systems*, 32, 2019.
- Niklas WA Gebauer, Michael Gastegger, Stefaan SP Hessmann, Klaus-Robert Müller, and Kristof T Schütt. Inverse design of 3d molecular structures with conditional generative neural networks. *Nature communications*, 13(1):1–11, 2022.
- 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. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.
- Gabriel Lima Guimaraes, Benjamin Sanchez-Lengeling, Carlos Outeiral, Pedro Luis Cunha Farias, and Alán Aspuru-Guzik. Objective-reinforced generative adversarial networks (organ) for sequence generation models. *arXiv preprint arXiv:1705.10843*, 2017.
- Emiel Hoogetboom, Victor Garcia Satorras, Clement Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International Conference on Machine Learning*, pages 8867–8887. PMLR, 2022.
- John J Irwin and Brian K Shoichet. Zinc- a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1):177–182, 2005.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical graph-to-graph translation for molecules. *arXiv preprint arXiv:1907.11223*, 2019.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In *International Conference on Machine Learning*, pages 4839–4848. PMLR, 2020a.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Multi-objective molecule generation using interpretable substructures. In *International conference on machine learning*, pages 4849–4859. PMLR, 2020b.
- Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov. drugan: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular pharmaceutics*, 14(9):3098–3104, 2017.
- Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- Frederic Koehler, Viraj Mehta, Andrej Risteski, and Chenghui Zhou. Variational autoencoders in the presence of low-dimensional data: landscape and implicit bias. *arXiv preprint arXiv:2112.06868*, 2021.
- Maria Korshunova, Niles Huang, Stephen Capuzzi, Dmytro S Radchenko, Olena Savych, Yuriy S Moroz, Carrow I Wells, Timothy M Willson, Alexander Tropsha, and Olexandr Isayev. Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Communications Chemistry*, 5(1):129, 2022.
- Matt J Kusner, Brooks Paige, and José Miguel Hernández-Lobato. Grammar variational autoencoder. In *International conference on machine learning*, pages 1945–1954. PMLR, 2017.
- Qi Liu, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L Gaunt. Constrained graph variational autoencoders for molecule design. *arXiv preprint arXiv:1805.09076*, 2018.
- Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. *Advances in Neural Information Processing Systems*, 34:6229–6239, 2021.
- Krzysztof Maziarz, Henry Jackson-Flux, Pashmina Cameron, Finton Sirockin, Nadine Schneider, Nikolaus Stiefl, Marwin Segler, and Marc Brockschmidt. Learning to extend molecular scaffolds with structural motifs. *arXiv preprint arXiv:2103.03864*, 2021.
- David Mendez, Anna Gaulton, A Patrícia Bento, Jon Chambers, Marleen De Veij, Eloy Félix, María Paula Magariños, Juan F Mosquera, Prudence Mutowo, Michał Nowotka, et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic acids research*, 47(D1):D930–D940, 2019.
- Daniil Polykovskiy, Alexander Zhebrak, Benjamin Sanchez-Lengeling, Sergey Golovanov, Oktai Tatanov, Stanislav Belyaev, Rauf Kurbanov, Aleksey Artamonov, Vladimir Aladinskiy, Mark Veselov, et al. Molecular sets (moses): a benchmarking platform for molecular generation models. *Frontiers in pharmacology*, 11:1931, 2020.
- Kristina Preuer, Philipp Renz, Thomas Unterthiner, Sepp Hochreiter, and Gunter Klambauer. Fréchet chemnet distance: a metric for generative models for molecules in drug discovery. *Journal of chemical information and modeling*, 58(9):1736–1741, 2018.
- Oleksii Prykhodko, Simon Viet Johansson, Panagiotis-Christos Kotsias, Josep Arús-Pous, Esben Jannik Bjerrum, Ola Engkvist, and Hongming Chen. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of Cheminformatics*, 11(1):1–13, 2019.
- Ali Razavi, Aaron van den Oord, Ben Poole, and Oriol Vinyals. Preventing posterior collapse with delta-vaes. *arXiv preprint arXiv:1901.03416*, 2019.
- David Rogers and Mathew Hahn. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50(5):742–754, 2010.
- Victor Garcia Satorras, Emiel Hoogeboom, Fabian B Fuchs, Ingmar Posner, and Max Welling. E (n) equivariant normalizing flows. *arXiv preprint arXiv:2105.09016*, 2021.
- Marwin HS Segler, Thierry Kogej, Christian Tyrchan, and Mark P Waller. Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131, 2018.

- Martin Simonovsky and Nikos Komodakis. Graphvae: Towards generation of small graphs using variational autoencoders. In *International conference on artificial neural networks*, pages 412–422. Springer, 2018.
- Peter C St. John, Caleb Phillips, Travis W Kemper, A Nolan Wilson, Yanfei Guan, Michael F Crowley, Mark R Nimlos, and Ross E Larsen. Message-passing neural networks for high-throughput polymer screening. *The Journal of chemical physics*, 150(23):234111, 2019.
- Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, and Pascal Frossard. Digress: Discrete denoising diffusion for graph generation. *arXiv preprint arXiv:2209.14734*, 2022.
- David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of chemical information and computer sciences*, 28(1):31–36, 1988.
- Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.
- Robin Winter, Floriane Montanari, Andreas Steffen, Hans Briem, Frank Noé, and Djork-Arné Clevert. Efficient multi-objective molecular optimization in a continuous latent space. *Chemical science*, 10(34):8016–8024, 2019.
- Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022.
- Kevin Yang, Kyle Swanson, Wengong Jin, Connor Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian Kelley, Miriam Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.
- Jiaxuan You, Bowen Liu, Rex Ying, Vijay Pande, and Jure Leskovec. Graph convolutional policy network for goal-directed molecular graph generation. *arXiv preprint arXiv:1806.02473*, 2018.

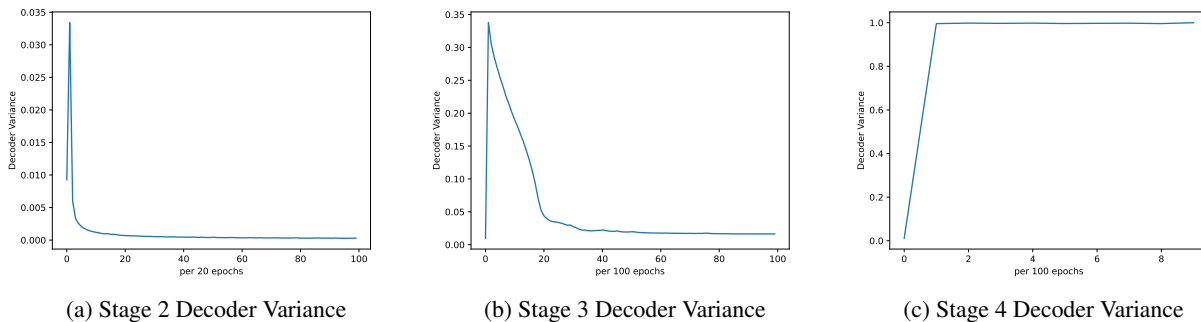


Figure 4: The decoder variance’s change over the course of training time.

## A BENCHMARK METRICS

Here we introduce the metrics used for the experiments in Section 4.1. Property Statistics include LogP (The Octanol-Water Partition Coefficient), SA (Synthetic Accessibility Score), QED (Quantitative Estimation of Drug-Likeness) and MW (Molecular Weight). These metrics determine the practicality of the generated molecules, for example, LogP measures the solubility of the molecules in water or an organic solvent (Wildman and Crippen, 1999), SA estimates how easily the molecules can be synthesized based on molecule structures (Ertl and Schuffenhauer, 2009), QED estimates how likely it can be a viable candidate of drugs (Bickerton et al., 2012). The values listed in the table for each metric are the Wasserstein distances between the distributions of the property statistics in the test set and the generate molecule set.

Structural statistics include SNN (Similarity to Nearest Neighbor), Frag (Fragment Similarity), and Scaf (Scaffold Similarity). These statistics calculate two molecular datasets’ structural similarity based on their extended-connectivity fingerprints (Rogers and Hahn, 2010), BRICS fragments (Degen et al., 2008) and Bemis–Murcko scaffolds (Bemis and Murcko, 1996).

The sample quality metrics are a lot more intuitive. Valid calculates the percentage of valid molecule outputs. Unique calculates the percentage of unique molecules in the first  $k$  molecules where  $k = 1000$  for the ChEMBL dataset. Novelty calculates the percentage of molecules generated that are not present in the training set. FCD is the Fréchet ChemNet Distance (Preuer et al., 2018).

## B CONVERGED DECODER VARIANCE IN DIFFERENT STAGES

For HGNN model, the stage #2 model’s decoder variance converges to 0.067 and the stage #3 decoder variance converges to 1.0. For the RNN-VAE model, the stage #2 model’s decoder variance converges to 1.

For the MoLeR model, the stage #2 model’s decoder variance converges to 0.00013, and the stage #3 model’s decoder variance converges to 0.016. We train a stage #4 VAE and it converges to 1.0. We include the plots of the decoder variance during training at each stage in Figure 4. For both stage #2 and #3, the decoder variance briefly goes up in the beginning of the training before converging to a much smaller value. In stage #4, the decoder variance reaches 1 very quickly and the value stays unchanged. We include the results generated by a 4-stage VAE in Table 5’s #4 row.

MoLeR + prop	Sample Quality				Structural Statistics			Property Statistics			
	Valid $\uparrow$	Unique $\uparrow$	Novelty $\uparrow$	FCD $\downarrow$	SNN $\uparrow$	Frag $\uparrow$	Scaf $\uparrow$	LogP $\downarrow$	SA $\downarrow$	QED $\downarrow$	MW $\downarrow$
#1	1.0	1.0	0.99	2.1	0.41	0.96	0.48	0.16 <sub>6.78e-3</sub>	<b>0.028</b> <sub>1.96e-3</sub>	0.047 <sub>2.78e-3</sub>	9.60.70
#2	1.0	1.0	0.99	2.2	0.42	0.96	0.53	0.12 <sub>6.13e-3</sub>	0.041 <sub>4.31e-3</sub>	0.036 <sub>9.25e-4</sub>	6.80.71
#3	1.0	1.0	0.99	1.8	0.42	0.96	0.49	0.087 <sub>6.16e-03</sub>	0.031 <sub>2.81e-03</sub>	<b>0.028</b> <sub>2.31e-03</sub>	8.24.0e-01
#4	1.0	1.0	0.99	1.9	0.42	0.96	0.48	<b>0.066</b> <sub>7.39e-03</sub>	0.030 <sub>3.27e-03</sub>	0.029 <sub>1.52e-03</sub>	10.6.44e-01

Table 5: Properties of the generated molecules from the 4th-stage VAE trained on the ChEMBL dataset using MoLeR without property matching as the first stage.

## C ADDITIONAL MULTI-STAGE VAE RESULTS

In this section, we include additional results on multi-stage VAE. We train a multi-stage VAE on a polymer dataset (St. John et al., 2019) with HGNN as the first-stage. The first-stage result is included in the HGNN paper (Jin et al., 2020a). We present the results from the second and third-stage VAE in relation to the first-stage in Table 6. We also trained two additional stages for MoLeR model with property matching included in the objective function in Table 7.

**The Polymer Dataset**(St. John et al., 2019) contains 86,353 polymers and it’s divided into training, test and validation set that contains 76,353, 5000 and 5000 molecules each. Polymers generally have heavier weight than the molecules in the ChEMBL dataset and the dataset size is smaller. Uniqueness is selected to be at top  $k = 500$  for the polymer dataset.

HGNN	Sample Quality				Structural Statistics			LogP ↓	Property Statistics		MW ↓
	Valid ↑	Unique ↑	Novelty ↑	FCD ↓	SNN ↑	Frag ↑	Scaf ↑		SA ↓	QED ↓	
#1	1.0	1.0	0.57	0.62	0.67	0.98	0.37	1.30.030	0.089 <sub>3.0e-3</sub>	0.020 <sub>1.2e-3</sub>	72.2 <sub>1.42</sub>
#2	1.0	1.0	0.51	0.27	0.69	0.99	0.37	<b>0.10</b> <sub>0.033</sub>	0.031 <sub>3.3e-3</sub>	0.004 <sub>19.5e-4</sub>	<b>7.7</b> <sub>1.1</sub>
#3	1.0	1.0	0.52	0.29	0.69	0.99	0.38	0.24 <sub>0.017</sub>	<b>0.024</b> <sub>4.1e-3</sub>	<b>0.0024</b> <sub>2.9e-4</sub>	9.4 <sub>2.3</sub>

Table 6: Properties of the generated molecules trained on the polymers dataset.

On the polymer dataset, the second stage VAE improves significantly across all metrics – from 72.2 to 7.7 on MW, 0.020 to 0.0024 on QED, 0.089 to 0.031 on SA and 1.3 to 0.1 on LogP. In the third stage, 2 of the metrics (SA and QED) improved while the other 2 degraded.

We train a multi-stage VAE on MoLeR with property matching as the first stage. Due to the modification to the objective function, none of the analysis described in the main paper necessarily apply here, but it is still interesting to see the results.

MoLeR + prop	Sample Quality				Structural Statistics			LogP ↓	Property Statistics		MW ↓
	Valid ↑	Unique ↑	Novelty ↑	FCD ↓	SNN ↑	Frag ↑	Scaf ↑		SA ↓	QED ↓	
#1	1.0	1.0	0.99	2.1	0.43	0.97	0.49	0.11 <sub>1.03e-02</sub>	0.13 <sub>2.51e-03</sub>	0.033 <sub>8.65e-04</sub>	6.6 <sub>4.80e-01</sub>
#2	1.0	1.0	0.99	2.2	0.42	0.97	0.48	<b>0.080</b> <sub>1.37-02</sub>	<b>0.090</b> <sub>6.03-03</sub>	0.030 <sub>1.80-03</sub>	<b>6.0</b> <sub>2.94-01</sub>
#3	1.0	1.0	0.99	2.0	0.43	0.97	0.41	0.096 <sub>1.21-02</sub>	0.13 <sub>5.17-03</sub>	<b>0.022</b> <sub>1.68-03</sub>	9.1 <sub>7.34-01</sub>

Table 7: Properties of the generated molecules from the multi-stage VAE trained on the ChEMBL dataset using MoLeR with property matching as the first stage.

In Table 7, we see that the second-stage VAE is able to improve upon the first-stage on all metrics while the third-stage improves only upon QED while the other 3 properties degraded.

## D ADDITIONAL ACTIVITY SCORE DISTRIBUTION FIGURES

In the main paper, we include the activity score distribution of the generated molecules trained on the EGFR dataset by Chemprop. We include the additional 3 figures that show the distribution of the activity scores generated by models trained on EGFR (Figure 6) and JAK2 dataset as predicted by Random Forest and JAK2 (Figure 7) dataset as predicted by Chemprop in Figure 5.

## E TRAINING DETAILS ON MULTI-STAGE VAE

Each stage of the multi-stage VAE with HGNN as the first stage has three fully-connected layers of size 512 for both encoders and decoders in addition to the input and output layer which are of size 20 (latent dimensions). The initial decoder variance is set at 0.05. Learning rate is set at 0.0001.

Each stage of the multi-stage VAE with MoLeR as the first stage has five fully-connected layers of size 1025 for both encoders and decoders in addition to the input and output layer which are of size 64 (latent dimensions). The initial decoder variance is set at 0.007. Learning rate is set at 0.0001. We trained our model for 10000 epochs but fewer epochs (e.g. 5000) can probably achieve similar results. For fine-tuning, the decoder variance is held as constant

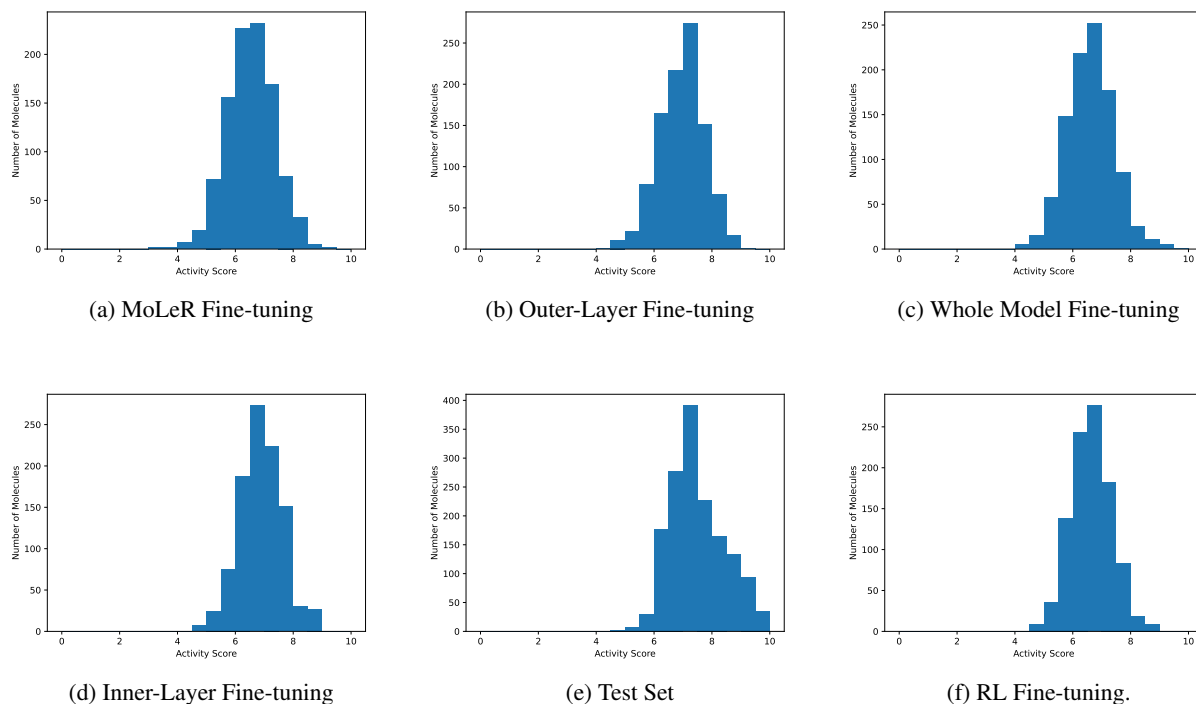


Figure 5: The distributions of the generated molecules’ activity scores by the Chemprop model on the JAK2 protein in six histograms.

during the process. Fine-tuning either inner-layer or extra-layer is done by loading the pre-trained model and add two extra layers either connecting to the latent layer or the output. The two extra layers are randomly initialized. The pre-trained part of the model is frozen while only the additional layers are being trained. Fine-tuning the whole model means to load the pre-trained model and only freeze the decoder variance while training the rest of it without additional layers. During fine-tuning, we use 0.00001 as the learning rate and train for 50000 epochs.

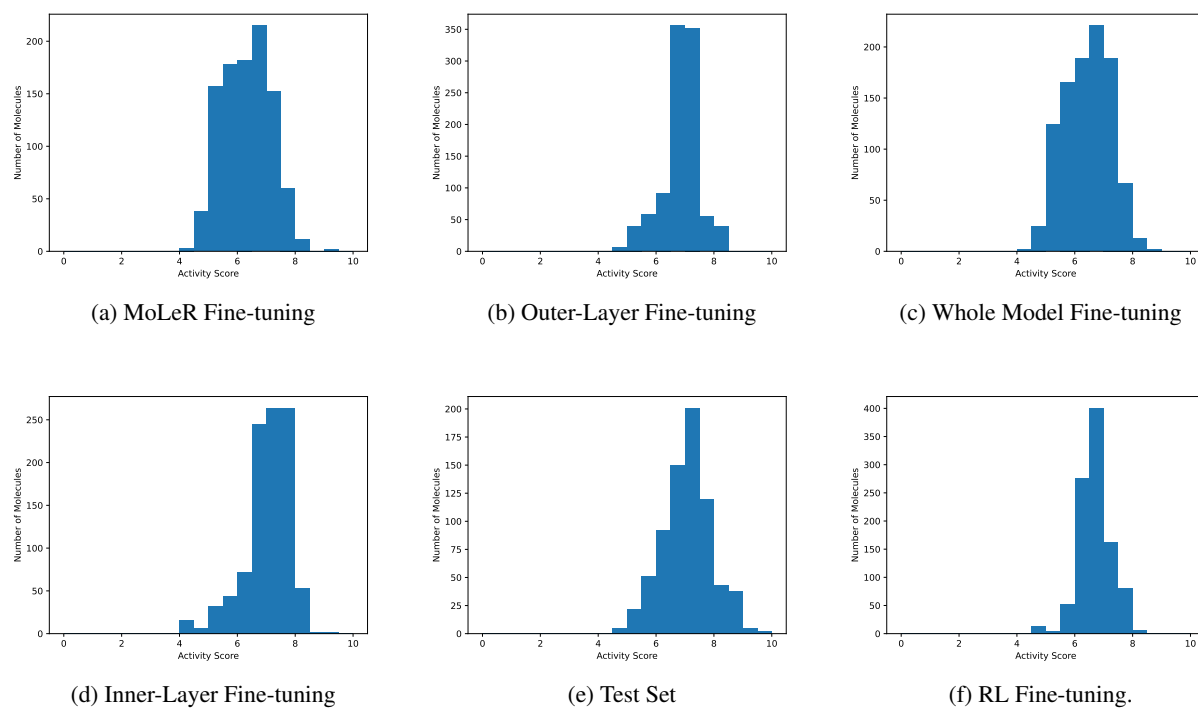


Figure 6: The distributions of the generated molecules' activity scores by the Random Forest model on the EGFR protein in six histograms.

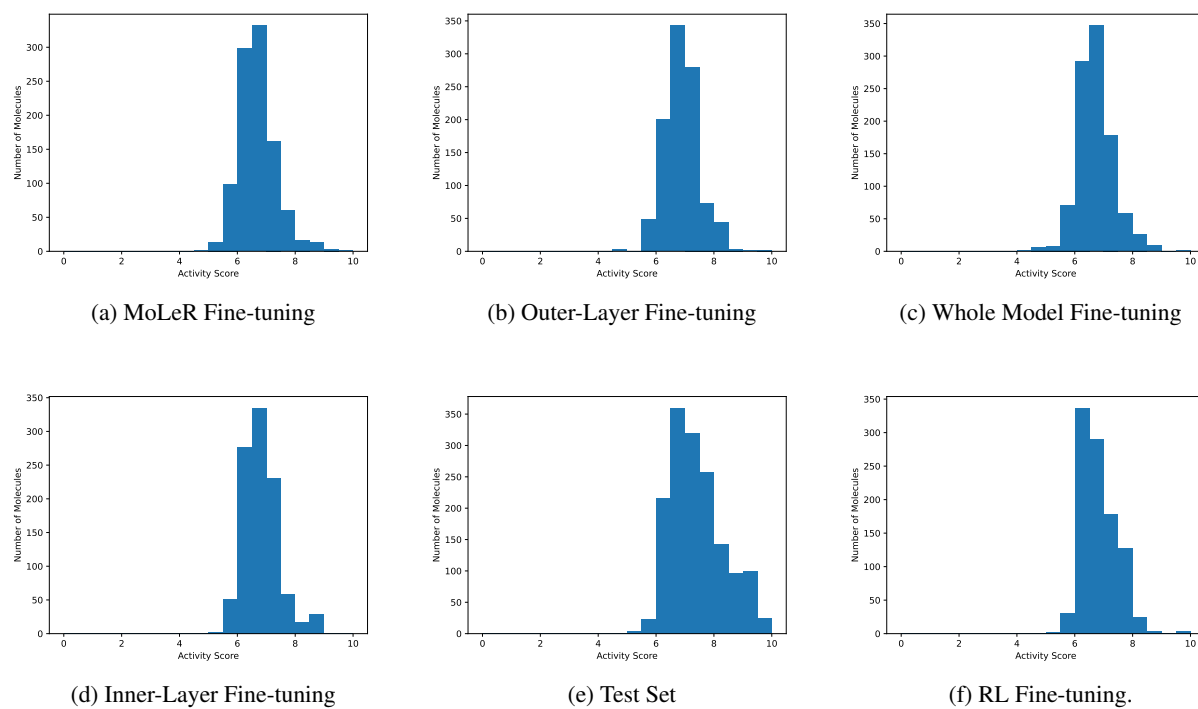


Figure 7: The distributions of the generated molecules' activity scores by the Random Forest model on the JAK2 protein in six histograms.