Molecular Fingerprint VAE

Andrea Karlova [*1,2] **Wim Dehaen** [*3] **Daniel Svozil** [3]

Abstract

We present low-dimensional latent representations learnt by the $\beta-$VAEs from the graph-topological structures encoding pharmacophoric features. The controlled information compression of these molecular fingerprints effectively removes the ambiguous redundancies and consequently results in encoding the chemically semantic latents. This latent molecular semantics allows for various tasks, from molecular similarity assessment to better-targeted search of the chemical space and drug discovery. We investigate the performance of the learnt latents of various dimensions on the ligand-based virtual screening task.

1. Introduction

Variational Autoencoder (VAE) (Kingma & Welling, 2014; Rezende et al., 2014) has been successfully used in generative modelling and representation learning due to efficient removal of the information redundancies in the original data manifold. Adding extra weight β to the KL-regularisor in VAE objective either improves the quality of reconstruction or results in learning disentangled representations (Higgins et al., 2017; Alemi et al., 2018).

The sparsity of molecular profiles and drug response high-dimensional data makes the downstream tasks, such as learning accurate prediction models, extremely difficult. VAEs have been applied to biological data to tackle the inconvenience of non-informative dimensionality. Recently, VAE latents were used to predict complex disease phenotypes from gene expression samples, (Dincer et al., 2018) or to improve drug response prediction accuracy (Rampášek L, 2019; Chiu et al., 2019)

In connection with molecular structure designs, VAEs were used to generate the novel chemical structures (Gómez-Bombarelli et al., 2018). To our best knowledge, these systems solely use the SMILEs[1] (Weininger, 1988) as an input. Despite its popularity, there are various pitfalls associated with this approach: the isomorphism between SMILES representing the same structure is not explicitly clear to most systems (Arús-Pous et al., 2019) as input sets will tend to include only canonized SMILES (for cross database consistency), further care is needed to infer the SMILES grammar which is only implicitly present in a set of grammatical compounds (Kusner et al., 2017).

Samanta et al. (2020); Shrivastava & Kell (2021) observed that the latent space of the SMILEs compressing VAEs exhibit certain chemical semantic similarity patterns and proposed to use VAE as a stand alone molecular similarity measure.

In contrast to SMILEs strings, the graph-topological representations of molecular structures, such as circular fingerprints (CFPs) (Rogers & Hahn, 2010), contain information about molecular features in an accessible way by encoding the topological environment of each atom. A useful variant of CFPs are functional-class fingerprints (FCFPs), which have feature based labels (corresponding to pharmacophore-like atom feature such as hydrogen bond acceptors), as opposed to the atom-based labels of standard extended CFPs (ECFPs) that describe the precise atomic environment.

Due to direct availability of the sub-structural information, CFPs become common tool in various ligand-based virtual screening (VS) tasks (Cereto-Massagué et al., 2015; Riniker & Landrum, 2013; Hu et al., 2012). VS is a widely used cost-effective alternative to the traditional high-throughput screening for the selection of initial hits in a search for drugs with a given biological activity (Walters et al., 1998; Bajorath, 2002; Gohlke & Klebe, 2002). VS is either structure-based (Cheng et al., 2012; Lionta et al., 2014) or ligand-based (Ripphausen et al., 2011; Geppert et al., 2010). While the former utilises the information about the macro-molecular target (protein or nucleic acid) via molecu-

[*]Equal contribution [1]Department of Computer Science, University College London, London, United Kingdom [2]April19 Discovery Inc., Delaware, USA [3]University of Chemistry and Technology, Prague, Czech Republic. Correspondence to: Andrea Karlova <akarlova@cs.ucl.ac.uk>, Wim Dehaen <dehaeni@vscht.cz>.

The 2021 ICML Workshop on Computational Biology.

[1]A SMILE string encodes a particular molecular graph using a vocabulary of atom and bond symbols and a grammar to parse these symbols into molecular connectivity. An isometric molecular graph (representing the same underlying molecule) does not have the same SMILE string. A canonical form that is identical for all isometric molecular graphs is generated using Morgan's algorithm.

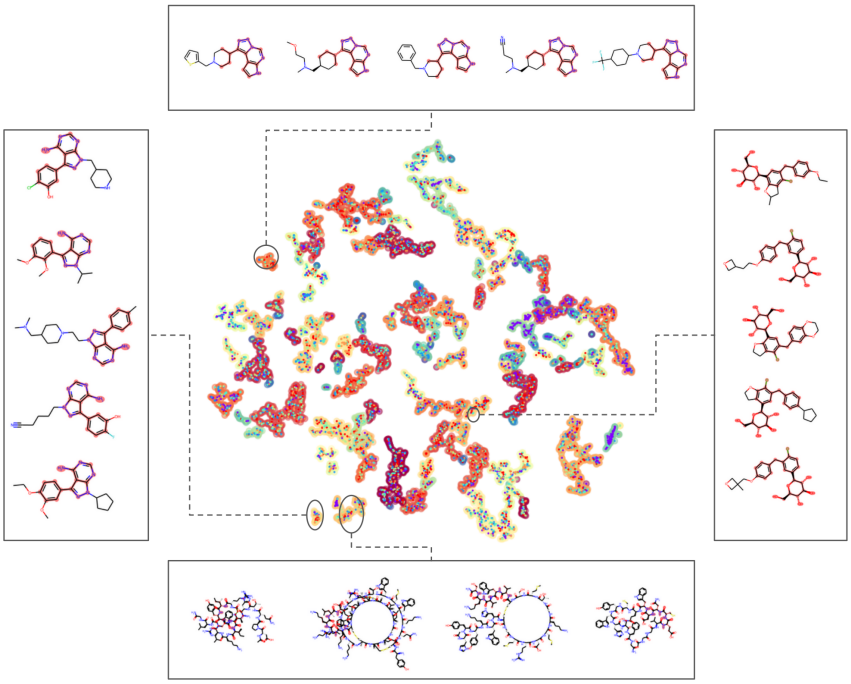

Figure 1. Chemical Semantics of Latent Space: The compression removes the redundancies of the 1024-dimensional chemically ambiguous binary fingerprint into the 16-dimensional chemically comparable latent code. Molecule classes in the top frame are Pyrrolo[3,2-*e*][1,2,3]triazolo[1,5-*c*]pyrimidines, right frame Aryl *C*-Glycosides, bottom frame Oligopeptide derivates, left frame 3-Aryl-pyrazolo[3,4-*d*]pyrimidin-4-amines. The color of each point was determined by the QED druglikeness score. Clusters (represented by background color) within the t-SNE visualization were assigned using DBSCAN. Common chemical substructures, highlighted in red, were automatically assigned using FCMS algorithm.

lar docking, this target structural information is rarely available. The ligand-based VS focuses on small-molecule ligand structure only. Via representations of these ligands (such as descriptors or structural fingerprints), meaningful relationships between the structure and the activity of these compounds are sought(Riniker & Landrum, 2013).

In this paper we investigate the latent space encoded by the FCFP molecular fingerprints. We demonstrate that the VAE leverages the graph-topological structure of the molecules encoded by the fingerprints and organises the molecules in the latent space in a chemically interpret-able way. Figure 1 depicts the representative molecular content of clusters identified by DBSCAN within VAE's latent space. The four molecular sets clearly demonstrate that the representation successfully capture high level features, such as specific drug-like scaffolds decorated with various substituents, natural product-like aryl C-glycosides and diverse oligopeptides apparently selected for large molecular size. These molecular patterns are challenging to characterize and interpret-ably compress.

Our main contribution is two-fold: 1) we demonstrate and investigate the chemical semantics of latent representations of the molecular fingerprints (LFPs); 2) we assess the impact of the compression of FCFP fingerprint into 8, 16, 32, 64 and 128 dimensional latent space with various compression weights β. We provide comparative evaluation of transferability of the latents on the target-assay pair prediction task.

2. Methods

The CFPs are generated by an iterative algorithm which assigns numeric identifiers to each atom of a molecule depending on the identifiers assigned to its neighboring atoms. The fingerprint algorithm generates a unique or near unique identifier which is independent of initial atom numbering while solving the molecular graph isomorphism problem. Importantly, the resulting identifiers generated by this scheme are comparable across molecules. These identifiers are then hashed into a smaller, fixed-length space to define a bit string of a predefined length. Typically, the length of 10 bits have an acceptable bit-collision rate.

VAE approximates the joint density $p(x, z)$ between ob-

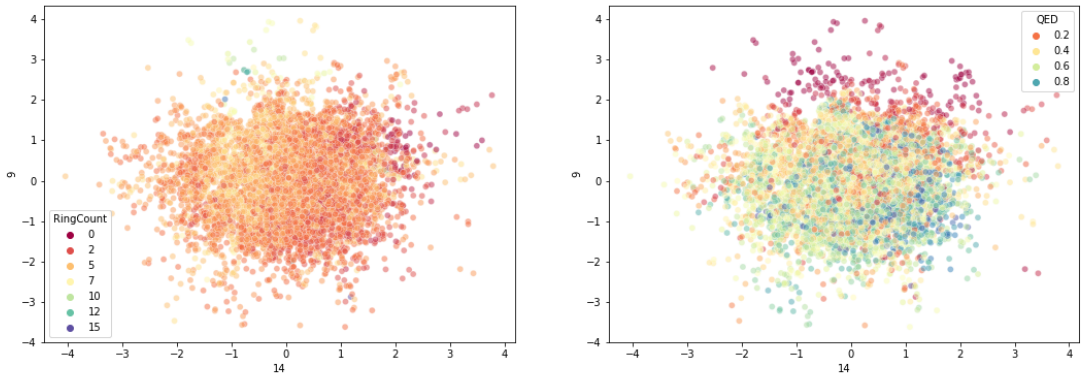

Figure 2. Latent Molecular Similarity: 2D slice of the latent space of two latent dimensions with the highest KL value. These are mainly used by the 16-dimensional latent VAE to compress the 1024-dimensional fingerprints. The depicted 9-th and 14-th coordinate of the locations of the variational posteriors are coloured according the QED values of the encoded molecule and counts of the carbon rings present in the structures. Neither is the number of the carbon rings explicitly encoded by the FCFP algorithm, nor is QED-score which quantitatively assesses the molecules according their drug-likeness based on the weighted average of Lipinski type descriptors(Bickerton et al., 2012).

served data $x \in \mathcal{X}$ and low-dimensional latents $z \in Z$ by first encoding the input data sample x via the probabilistic encoder $p(z|x)$ and then reconstructing the sample via latent decomposition: $p(x;\theta) = \int p(x|z;\theta)\pi(z)dz$. The intractable encoder $p(z|x)$ is approximated by using variational distribution $q(z|x;\phi)$. The representation learning is enhanced by squeezing the encoded latent representation to an uninformative prior $\pi(z)$ in the KL-divergence $KL[q(z|x;\phi)||\pi(z)]$. The training objective based on evidence lower bound is given by the constrained optimisation problem: $\max_{\theta,\phi} E_{x\sim\mathcal{D}}[E_{q(z|x;\phi)}[\ln p(x|z;\theta)]]$, with constrain $KL[q(z|x;\phi)||\pi(z)] < \varepsilon$. Higgins et al. (2017) introduced a multiplicative hyper-parameter β of the KL term, which improves the disentangling of the learnt latent representations. The objective rewrites as: $\mathrm{L}^{\beta} = E_{q(z|x;\phi)}[\ln p(x|z;\theta)] - \beta KL[q(z|x;\phi)||\pi(z)]$. The choice of $\beta = 1$ corresponds to vanilla VAE. For $\beta > 1$, the latent bottleneck becomes more constrained which encourages to learn more efficient representations of the data.

In our setting the fingerprint maps the SMILEs representation into one of the vertices of the 1024-dimensional cube. VAE then reduces ambiguous information of the high-dimensional cube structure into lower dimensional chemically informative variational posterior, which partially extracts the imprints of the fingerprint hyper-cube pressed in by Gaussian prior manifold via the KL-regulariser. As fingerprint takes binary values, we use Bernoulli decoder $p(x|z)$, which results in the cross-entropy objective. Due to the initial compression into the hashed binary code, it is sufficient to use MLP architecture to parameterise the inference and generative nets of VAE.

For evaluation of the chemical semantics via the distribution of various structural features of the latents we use basic physico-chemical molecular descriptors. These descriptors include: molecular weight, polarity of the molecule (log P), fraction of sp^3-hybridized carbon atoms, a measure of flexibility(count of NH and OH, number of rotatable bonds, ring count, generally), and specifically number of aromatic heterocycles, number of aromatic carbocycles, number of aliphatic heterocycles, number of aliphatic carbocycles, total polar surface area, quantitative estimation of drug-likeness (QED-score) (Bickerton et al., 2012), see Figure 2.

3. Experiments

To investigate the properties of the LFP representations[2] we trained VAE with latent dimensions 8, 16, 32, 64 and 128. Depending on the size of the latent space we used different values of β. For 8 and 16-dimensional latents, the latent space tend to collapse for $\beta > 2$, whereas for larger dimensions even $\beta = 3$ provided useful compression. We used fixed MLP architecture for all VAEs: input layer: 1024-long FCFP fingerprint; encoder: 256, 128, 128, 128, 64, 64, 64, 32, 32 with ReLu activation; decoder: 32, 64, 64, 64, 128, 256, 256, 1024 with ReLu activation. We explored βs: 1.0, 1.5, 1.75, 2.0, 2.5, 3.0.

We extracted 550k molecules from the ChEMBL database (v28) (Mendez et al., 2018), a public repository of biological data including assay measurements at various protein targets. We used cutoffs to extract the most reliable sets of molecules

[2]https://github.com/AndreaKarlova/molecular_fingerprint_latents

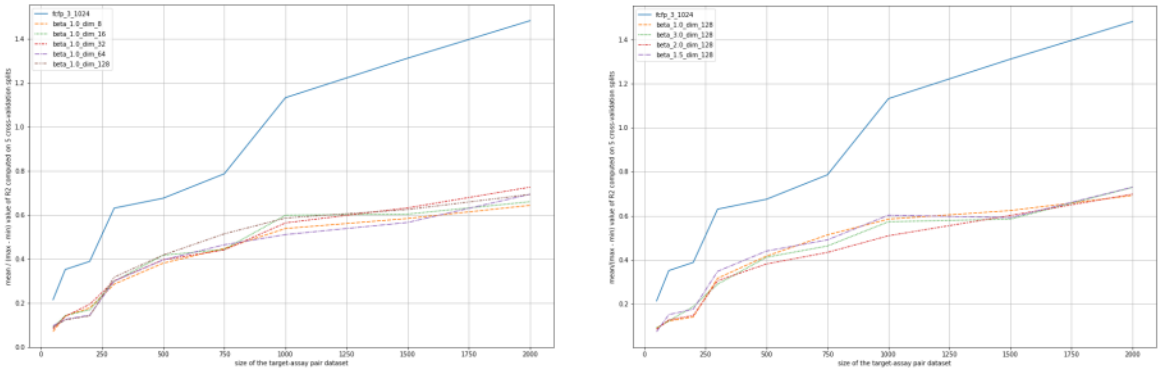

Figure 3. Performance of LFPs on the Quantitative Structure Activity Relationship (QSAR) task (Geppert et al., 2010): R^2 is computed via 5-fold cross-validation due to the presence of the small data-samples. We use mean of R^2 scores across all 5 folds scaled by the max-min spread. The spread is larger for the small size datasets in the left part of the plots. With the sufficiently large target assay set it shrinks to small bound comparable across all modelling approaches. For datasets around sizes between 250-500 data-points, the mean R^2 performance becomes nearly constant, so the larger slope of the curve in the right part of the plots indicates fastly shrinking max-min spread. The left figure depicts the performance of Molecular Fingerprint VAEs with various latent dimensions, while the right figure plots the performance for latent dimension 128 and various values of β. For the sufficiently large datasets, the improvements stemming from increasing the size latent space are rather minor. The larger value β for larger latent space also slightly helps with improving the performance on the task.

for which the measured activity at a specific target-assay pair was available. The cut-offs were: ChEMBL confidence score ≥ 7, ≥ 50 compounds per set, target organism must be Homo Sapiens, Mus Musculus or Rattus Norvegicus, assay type must be K_i, K_d, EC_{50} or IC_{50} (Škuta et al., 2020; Cortés-Ciriano et al., 2020). We were left with 2339 sets of compounds each corresponding to a specific target assay pair.

To train VAEs we sub-sampled 80k molecules, such that there was small portion of molecules available in 75% of all target-assay pairs sets. For the down-stream task evaluation, we use the trained VAE encoder and compressed each of 2339 target-assay set. On the compressed target-assay set we trained random forest regressor (RFs) (with 100 trees) to predict the assay activity, see Figure 3 for the performance summary. Despite the QSAR-RFs on raw 10-bit FCFPs demonstrate significantly better performance when measured by R^2-score, note, that, for easier reproducibility we do not compensate R^2 score for the order-of-magnitude larger number of predictors of raw FCFPs. It is a common practice to adjust the FCFP length to the length of the representation or fingerprint it is being compared to (Wu et al., 2018; Probst & Reymond, 2018). However, for ECFP and FCFP with typical radii, performance rolls off strongly as the amount of bit collisions and string density increases as the length of the fingerprint decreases (Gütlein & Kramer, 2016). In contrast, the LFPs performs in the robust manner: even 8, 16 and 32-dimensional LFPs have good transferability properties and well generalise to perform the task on the order of magnitude larger dataset than the subsample of it to which LFPs have been trained.

4. Conclusions and Discussion

We demonstrated that VAE with MLP architecture and Bernoulli decoder removes the redundancies in the fingerprint hash. Resulting LFP space is naturally organised according the similarity properties of physico-chemical descriptors allowing for investigating of novel latent molecular patterns.

The benefits of using fingerprints based VAEs over the SMILE-based VAEs are: 1) a simpler architecture (MLP is sufficient for the problem) comparing to SMILES which requires more complex architectures, such as: RNNs, ConvNets or Transformers; 2) using CFP encoding of the pharmacophoric features rather than specific element atom labels captured by SMILEs (and ECFP), which do not adequately capture bio-isosteric fragments (e.g. amide and ester functional groups).

The potential drawback of using directly CFPs is the lack of convertibility to a molecular graph, which limits de-novo structure generation. Le et al. (2020) proposed an approach to invert the ECFP fingerprint to SMILEs. Note, that the decoders of the SMILE-based VAEs suffer with generating parsable SMILEs that are grammatically correct but chemically meaningless (Kusner et al., 2017).

References

Alemi, A. A., Poole, B., Fischer, I., Dillon, J. V., Saurous, R. A., and Murphy, K. Fixing a broken ELBO. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 159–168. PMLR, 2018. URL http://proceedings.mlr.press/v80/alemi18a.html.

Arús-Pous, J., Johansson, S. V., Prykhodko, O., Bjerrum, E. J., Tyrchan, C., Reymond, J.-L., Chen, H., and Engkvist, O. Randomized SMILES strings improve the quality of molecular generative models. *Journal of Cheminformatics*, 11(1), nov 2019. doi: 10.1186/s13321-019-0393-0.

Bajorath, J. Integration of virtual and high-throughput screening. *Nature Reviews Drug Discovery*, 1(11):882–894, nov 2002. doi: 10.1038/nrd941.

Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S., and Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry*, 4(2):90–98, jan 2012. doi: 10.1038/nchem.1243.

Cereto-Massagué, A., Ojeda, M. J., Valls, C., Mulero, M., Garcia-Vallvé, S., and Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods*, 71:58–63, jan 2015. doi: 10.1016/j.ymeth.2014.08.005.

Cheng, T., Li, Q., Zhou, Z., Wang, Y., and Bryant, S. H. Structure-based virtual screening for drug discovery: a problem-centric review. *The AAPS Journal*, 14(1):133–141, jan 2012. doi: 10.1208/s12248-012-9322-0.

Chiu, Y.-C., Chen, H.-I. H., Zhang, T., Zhang, S., Gorthi, A., Wang, L.-J., Huang, Y., and Chen, Y. Predicting drug response of tumors from integrated genomic profiles by deep neural networks. *BMC Medical Genomics*, 12, 2019. doi: 10.1186/s12920-018-0460-9. URL https://doi.org/10.1186/s12920-018-0460-9.

Cortés-Ciriano, I., Škuta, C., Bender, A., and Svozil, D. QSAR-derived affinity fingerprints (part 2): modeling performance for potency prediction. *Journal of Cheminformatics*, 12(1), jun 2020. doi: 10.1186/s13321-020-00444-5.

Dincer, A. B., Celik, S., Hiranuma, N., and Lee, S.-I. Deepprofile: Deep learning of cancer molecular profiles for precision medicine. In *Join ICML AND ICJCAI Workshop on Computational Biology*, 2018. doi: 10.1101/278739. URL https://doi.org/10.1101/278739.

Geppert, H., Vogt, M., and Bajorath, J. Current trends in ligand-based virtual screening: Molecular representations, data mining methods, new application areas, and performance evaluation. *Journal of Chemical Information and Modeling*, 50(2):205–216, jan 2010. doi: 10.1021/ci900419k.

Gohlke, H. and Klebe, G. Approaches to the description and prediction of the binding affinity of small-molecule ligands to macromolecular receptors. *Angewandte Chemie International Edition*, 41(15):2644–2676, aug 2002. doi: 10.1002/1521-3773(20020802)41:15⟨2644::aid-anie2644⟩3.0.co;2-o.

Gómez-Bombarelli, R., Wei, J., Duvenaud, D., JM, H.-L., and Sánchez-Lengeling B, Sheberla D, A.-I. J. H. T. A. R. A.-G. A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Cent Sci.*, 2:268–276, Feb 2018. doi: doi:10.1021/acscentsci.7b00572.

Gütlein, M. and Kramer, S. Filtered circular fingerprints improve either prediction or runtime performance while retaining interpretability. *Journal of Cheminformatics*, 8, 2016. doi: 10.1186/s13321-016-0173.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M. M., Mohamed, S., and Lerchner, A. β-vae: Learning basic visual concepts with a constrained variational framework. In *ICLR*, 2017.

Hu, G., Kuang, G., Xiao, W., Li, W., Liu, G., and Tang, Y. Performance evaluation of 2d fingerprint and 3d shape similarity methods in virtual screening. *Journal of Chemical Information and Modeling*, 52(5):1103–1113, may 2012. doi: 10.1021/ci300030u.

Kingma, D. P. and Welling, M. Auto-encoding variational bayes. In Bengio, Y. and LeCun, Y. (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

Kusner, M. J., Paige, B., and Hernández-Lobato, J. M. Grammar variational autoencoder. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1945–1954. PMLR, 06–11 Aug 2017. URL http://proceedings.mlr.press/v70/kusner17a.html.

Le, T., Winter, R., Noé, F., and Clevert, D.-A. Neuraldecipher – reverse-engineering extended-connectivity fingerprints (ECFPs) to their molecular structures. *Chemical Science*, 11(38):10378–10389, 2020. doi: 10.1039/d0sc03115a.

Lionta, E., Spyrou, G., Vassilatis, D., and Cournia, Z. Structure-based virtual screening for drug discovery: Principles, applications and recent advances. *Current Topics in Medicinal Chemistry*, 14(16):1923–1938, oct 2014. doi: 10.2174/1568026614666140929124445.

Mendez, D., Gaulton, A., Bento, A. P., Chambers, J., Veij, M. D., Félix, E., Magariños, M. P., Mosquera, J. F., Mutowo, P., Nowotka, M., Gordillo-Marañón, M., Hunter, F., Junco, L., Mugumbate, G., Rodriguez-Lopez, M., Atkinson, F., Bosc, N., Radoux, C. J., Segura-Cabrera, A., Hersey, A., and Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research*, 47 (D1):D930–D940, nov 2018. doi: 10.1093/nar/gky1075.

Probst, D. and Reymond, J.-L. A probabilistic molecular fingerprint for big data settings. *Journal of Cheminformatics*, 10, 2018. doi: 10.1186/s13321-018-0321-8.

Rampášek L, Hidru D, S. P.-H.-K. B. G. A. Dr.vae: improving drug response prediction via modeling of drug perturbation effects. *Bioinformatics*, pp. 3743–3751, Oct 2019. doi: 10.1093/bioinformatics/btz158.

Rezende, D. J., Mohamed, S., and Wierstra, D. Stochastic backpropagation and approximate inference in deep generative models. In Xing, E. P. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 1278–1286, Bejing, China, 22–24 Jun 2014. PMLR. URL http://proceedings.mlr.press/v32/rezende14.html.

Riniker, S. and Landrum, G. A. Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1), may 2013. doi: 10.1186/1758-2946-5-26.

Ripphausen, P., Nisius, B., and Bajorath, J. State-of-the-art in ligand-based virtual screening. *Drug Discovery Today*, 16(9-10):372–376, may 2011. doi: 10.1016/j.drudis.2011.02.011.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, apr 2010. doi: 10.1021/ci100050t.

Samanta, S., O'Hagan, S., Swainston, N., Roberts, T. J., and Kell, D. Vae-sim: A novel molecular similarity measure based on a variational autoencoder. *Molecules*, 25, 2020.

Shrivastava, A. D. and Kell, D. B. Fragnet, a contrastive learning-based transformer model for clustering, interpreting, visualizing, and navigating chemical space. *Molecules*, 26(7), 2021. doi: 10.3390/molecules26072065. URL https://www.mdpi.com/1420-3049/26/7/2065.

Škuta, C., Cortés-Ciriano, I., Dehaen, W., Kříž, P., van Westen, G. J. P., Tetko, I. V., Bender, A., and Svozil, D. QSAR-derived affinity fingerprints (part 1): fingerprint construction and modeling performance for similarity searching, bioactivity classification and scaffold hopping. *Journal of Cheminformatics*, 12(1), may 2020. doi: 10.1186/s13321-020-00443-6.

Walters, W., Stahl, M. T., and Murcko, M. A. Virtual screening—an overview. *Drug Discovery Today*, 3(4):160–178, apr 1998. doi: 10.1016/s1359-6446(97)01163-x.

Weininger, D. SMILES, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1):31–36, feb 1988. doi: 10.1021/ci00057a005.

Wu, J., Zhang, Q., Wu, W., Pang, T., Hu, H., Chan, W. K. B., Ke, X., and Zhang, Y. WDL-RF: predicting bioactivities of ligand molecules acting with G protein-coupled receptors by combining weighted deep learning and random forest. *Bioinformatics*, 34(13):2271–2282, 02 2018. doi: 10.1093/bioinformatics/bty070.