A generative recommender system with GMM prior for cancer drug generation and sensitivity prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Recent emergence of high-throughput drug screening assays sparkled an intensive development of machine learning methods, including models for prediction of sensitivity of cancer cell lines to anti-cancer drugs, as well as methods for generation of potential drug candidates. However, the concept of generation of compounds with specific properties and simultaneous modeling of their efficacy against cancer cell lines has not been comprehensively explored. To address this need, we present VADEERS, a Variational Autoencoder-based Drug Efficacy Estimation Recommender System. The generation of compounds is performed by a novel variational autoencoder with a semi-supervised Gaussian Mixture Model (GMM) prior. The prior defines a clustering in the latent space, where the clusters are associated with specific drug properties. In addition, VADEERS is equipped with a cell line autoencoder and a sensitivity prediction network. The model combines data for SMILES string representations of anti-cancer drugs, their inhibition profiles against a panel of protein kinases, cell lines' biological features and measurements of the sensitivity of the cell lines to the drugs. The evaluated variants of VADEERS achieve a high $r = 0.87$ Pearson correlation between true and predicted drug sensitivity estimates. We show that the learned latent representations and new generated data points accurately reflect the given clustering. In summary, VADEERS offers a comprehensive model of drugs' and cell lines' properties and relationships between them, as well as a guided generation of novel compounds.

1 Introduction

Kinase inhibitors are a class of anticancer drugs that target specific mutated kinases and disregulated biological processes in tumor cells [1]. As such, they constitute flagship examples of personalized cancer treatments [2, 3]. Their chemical structure is typically represented as strings termed SMILES [4]. In addition, the set of kinase inhibitors is deeply investigated experimentally. They are commonly characterized by their *inhibition profiles*, measuring their strength of inhibition of a panel of kinases [5, 6]. In addition, the *sensitivity* of cancer cell lines to kinase inhibitors was measured by large-scale experiments [7, 8, 9]. The *molecular features* of these cancer cell lines, such as gene mutations and gene expression were also profiled [7, 8, 10]. Despite their limitations, cancer cell lines commonly act as laboratory proxies for patients' tumors and it is known that their molecular features are key determinants of their response to anticancer drugs [8, 11]. While a number of kinase inhibitor drugs is already successfully applied in the clinic, the mechanism of resistance to treatment and a large number of cancer mutations that could be additionally targeted to circumvent this resistance creates a pressing need for novel drug discovery [12, 13, 14]. Unfortunately, the current pre-clinical attempts of proposing novel compounds proves inefficient, as the drug candidates fail further stages of clinical trials, yielding the process of novel drug discovery a daunting, time and money consuming task [15, 16, 17].

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

Machine learning, in particular deep generative models, transform the field of molecule discovery, providing promising drug candidates with with desired chemical properties [18, 19, 20, 21, 22, 23, 24].

Key problems. This work addresses several important research problems.

- **Existing generative molecule models are not directly applicable to kinase inhibitors.** They require large amounts of compounds for training, while the number of known kinase inhibitors is scarce. Moreover, they do not account for the molecular features of the drugs and of the tumors that the drugs are supposed to act on. Drug sensitivity is a function of both compound's and tumor's features, and it is the relationship between these two feature sets that determines the treatment outcome.
- **Existing machine learning models for drug sensitivity prediction are not generative.** Combining the functionalities of prediction and generation in a single model has the potential to mutually strengthen the performance of the model in both tasks, while regularizing the model and preventing overfitting to a single task.
- **The data available for the drugs and the cell lines pose a difficult integration problem with missing data:** for some drugs, only the sensitivity of the cell lines to these drugs was measured and not their inhibition profiles, and vice versa. Finally, there exist compounds for which only the SMILES strings are known.

Proposed solutions. In this work, we propose a novel generative framework for simultaneous kinase inhibitor discovery and sensitivity prediction. The framework restricts the vast space of potential generative model hypotheses by accounting for a large variety of experimental data (Fig. 1a). Specifically, we cluster the drugs by their inhibitory profiles, and provide the clustering of the drugs together with the drugs' SMILES representations, cell line molecular features, the inhibitory profiles and the sensitivity values as input to the model for training. Due to the fact that for some drugs the inhibition profiles are not available, the clustering provides only partial cluster labels for the drugs, posing a semi-supervised clustering problem. The generative **drug module** of the framework is implemented using SS GMM VAE, a new semi-supervised variational autoencoder (VAE) model with a Gaussian mixture model (GMM) prior (Fig. 1b). SS GMM VAE infers representations of the drugs' SMILES and enables generation of specific types of kinase inhibitors, guided by the clustering of their inhibitory profiles within the GMM prior. In addition, the framework includes also a **cancer cell line module** for identification of representations of cancer cell lines and a **sensitivity prediction module** that performs the prediction of the sensitivity of the cell lines to the drugs (Fig. 1c, d).

On the most general level, the proposed framework can be thought of as an extension of a recommender system with side information [25, 26, 27, 28, 29] with a generative model. In our particular application, in the generative recommender system the objects correspond to drugs from the family of kinase inhibitors, users to cancer cell lines, while the scores correspond to the sensitivity of the cell lines to the drugs. Hence the name of the framework, i.e. Variational Autoencoder-based Drug Efficacy Estimation Recommender System (VADEERS).

Key contributions. This work offers the following key novel contributions:

- VADEERS, an integrative framework that combines i) generation of kinase inhibitor drugs with ii) finding their representations, iii) modeling of cancer cell lines and their representations, and iv) prediction of cancer cell line sensitivity to drugs (Fig. 1e).
- SS GMM VAE, which is trainable with partial cluster labels. We introduce a novel formulation of the prior, which, in contrast to previous GMM VAEs, enables semi-supervised cluster inference without an additional inference model. Thanks to SS GMM VAE, VADEERS is able to generate novel drugs having specific types of inhibitory profiles and readily predict their their sensitivity profile on cancer cell lines.

2 Results

We evaluated three versions of the proposed model, differing by the way the drug module was implemented: i) a classical VAE with the standard normal prior ("Vanilla VAE"), ii), the SS GMM VAE, however, only weights π_k's and components' means μ_k's were the trainable parameters of the GMM prior, while components' covariance matrices Σ_k's were fixed as identity matrices, iii) the SS

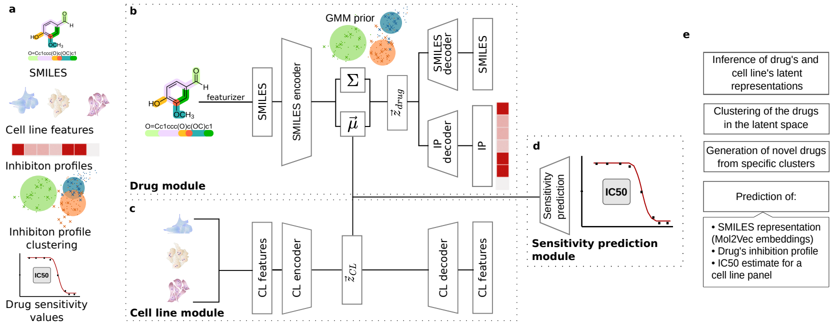
a
SMILES
Cell line features
Inhibiton profiles
Inhibiton profile clustering
IC50
Drug sensitivity values
b
GMM prior
featurizer
SMILES
SMILES encoder
Σ
$\vec{\mu}$
$\vec{z}_{drug}$
SMILES decoder
SMILES
IP decoder
IP
Drug module
c
CL features
CL encoder
$\vec{z}_{CL}$
CL decoder
CL features
Cell line module
d
Sensitivity prediction
IC50
Sensitivity prediction module
e
Inference of drug's and cell line's latent representations
Clustering of the drugs in the latent space
Generation of novel drugs from specific clusters
Prediction of:
• SMILES representation (Mol2Vec embeddings)
• Drug's inhibition profile
• IC50 estimate for a cell line panel

Figure 1: Framework's overview. (**a**) Data types used for training. (**b**) Drug module. (**c**) Cancer cell line module. (**d**) Sensitivity prediction module. Sensitivity prediction module takes concatenation of drug module's encoder output, i.e. mean vector, and cancer cell line module's latent vector as input. (**e**) Key framework's functionalities.

Table 1: IC50 and IP prediction performance for VADEERS with different versions of the drug module (top three rows), and two other models as reported in the corresponding works (bottom two rows). The models of Liu *et al.* and Koras *et al.* lack the generative ability and do not perform inference of inhibition profiles, hence the lack of corresponding metrics.

Model	IC50 RMSE	IC50 Pearson	IP RMSE
VADEERS w. Vanilla VAE	1.33 ± 0.022	0.87 ± 0.006	1.13 ± 0.109
VADEERS w. SS GMM VAE constrained	1.33 ± 0.023	0.87 ± 0.006	1.09 ± 0.062
VADEERS w. SS GMM VAE unconstrained	1.34 ± 0.012	0.87 ± 0.004	1.04 ± 0.030
Liu *et al.* [30]	–	0.89	–
Koras *et al.* [31]	–	0.82	–

GMM VAE, in its least constrained version, where all parameters of the GMM, including Σ_k's, were trainable ("SS GMM VAE unconstrained").

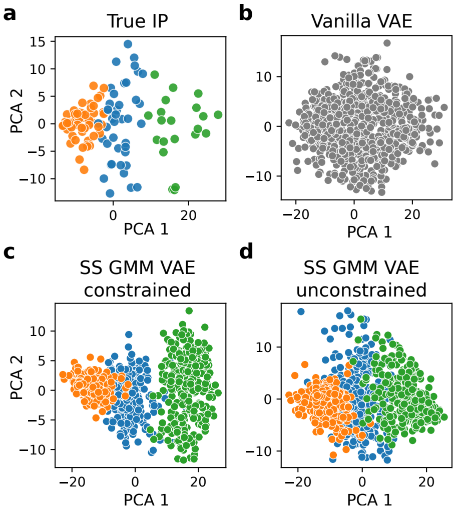

Figure 2: True and generated inhibition profiles visualized in 2D. (**a**) The true IPs for the 117 available drugs. (**b**) 900 IPs generated from the Vanilla VAE. (**c**) IPs generated from the SS GMM VAE constrained model. 300 samples are drawn. (**d**) IPs generated from the SS GMM VAE unconstrained model. Again, 300 samples are drawn per-component. Colors correspond to guiding label or a corresponding GMM component.

3 Conclusions

In this work, we propose VADEERS, a multi-task framework for generation of novel drugs with specific types of inhibition profiles and simultaneous drug sensitivity prediction. The framework exploits a novel SS GGM VAE model that enables semi-supervised clustering of the drugs' representaions in the latent space. We showed that the framework achieves state-of-the-art sensitivity prediction performance, and preserves a given clustering structure of the drugs both in the latent space and in the space of the predicted inhibitory profiles.

One of the limitations of the proposed model is its inability to generate data points with totally arbitrary features. Namely, the model allows to generate new data points with properties that strictly reflect the clustering observed in the training data. In principle, this could be bypassed by performing various operations on multiple generated data points, however, testing this hypothesis was not in the scope of this analysis. Another important limitation corresponds to the analyzed data; a different choice of data for drugs' representations (e.g. representing SMILES strings as graphs) and guiding data might be more suitable for generating molecule candidates, which, at least in theory, could be synthesized. Both above aspects are directions of future work regarding this study.

This work introduces several general concepts important for drug sensitivity modeling and compound generation. The proposed SS GMM VAE model is generic and not limited only to modeling compounds. The notion of optimizing latent space with guiding labels can potentially be beneficial and improve the performance of generative models also in other applications. Moreover, the proposed model offers additional functionality not exploited in this study. For example, setting the number of Gaussian components K greater than number of unique labels G might lead to identification of novel subgroups of samples, not limited to the original choice of guiding labels. In summary, VADEERS opens new avenues in integrative modeling of cancer data and generation of anticancer compounds.

References

[1] Radhamani Kannaiyan and Daruka Mahadevan. A comprehensive review of protein kinase inhibitors for cancer therapy. *Expert Review of Anticancer Therapy*, 18(12):1249–1270, 2018.

[2] Jianming Zhang, Priscilla Yang, and Nathanael Gray. Zhang j, yang pl, gray nstargeting cancer with small molecule kinase inhibitors. nat rev cancer 9: 28-39. *Nature reviews. Cancer*, 9:28–39, 02 2009.

[3] Robert Roskoski. Properties of fda-approved small molecule protein kinase inhibitors: A 2020 update. *Pharmacological Research*, 152:104609, 2020.

[4] David Weininger. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.*, 28:31–36, 1988.

[5] Krisna C. Duong-Ly, Karthik Devarajan, Shuguang Liang, Kurumi Y. Horiuchi, Yuren Wang, Haiching Ma, and Jeffrey R. Peterson. Kinase inhibitor profiling reveals unexpected opportunities to inhibit disease-associated mutant kinases. *Cell Reports*, 14(4):772–781, 2016.

[6] Chandrasekhar V. Miduturu, Xianming Deng, Nicholas Kwiatkowski, Wannian Yang, Laurent Brault, Panagis Filippakopoulos, Eunah Chung, Qingkai Yang, Juerg Schwaller, Stefan Knapp, Randall W. King, Jiing-Dwan Lee, Sanna Herrgard, Patrick Zarrinkar, and Nathanael S. Gray. High-throughput kinase profiling: A more efficient approach toward the discovery of new kinase inhibitors. *Chemistry & Biology*, 18(7):868–879, 2011.

[7] Cyril Benes, Daniel A. Haber, Dave Beare, Elena J. Edelman, Howard Lightfoot, I. Richard Thompson, James A. Smith, Jorge Soares, Michael R. Stratton, Nidhi Bindal, P. Andrew Futreal, Patricia Greninger, Simon Forbes, Sridhar Ramaswamy, Wanjuan Yang, Ultan McDermott, and Mathew J. Garnett. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Research*, 41(D1):D955–D961, 11 2012.

[8] J Barretina, Giordano Caponigro, N Stransky, Kavitha Venkatesan, Adam Margolin, Sunghyok Kim, C.J. Wilson, Joseph Lehar, G.V. Kryukov, D Sonkin, A Reddy, M Liu, L Murray, M.F. Berger, J.E. Monahan, Paula Keskula, J Meltzer, A Korejwa, J Jane-Valbuena, and M de Silva. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity (vol 483, pg 603, 2012). *Nature*, 492:290–290, 01 2012.

[9] Brinton Seashore-Ludlow, Matthew G. Rees, Jaime H. Cheah, Murat Cokol, Edmund V. Price, Matthew E. Coletti, Victor Jones, Nicole E. Bodycombe, Christian K. Soule, Joshua Gould, Benjamin Alexander, Ava Li, Philip Montgomery, Mathias J. Wawer, Nurdan Kuru, Joanne D. Kotz, C. Suk-Yee Hon, Benito Munoz, Ted Liefeld, Vlado Dančík, Joshua A. Bittker, Michelle Palmer, James E. Bradner, Alykhan F. Shamji, Paul A. Clemons, and Stuart L. Schreiber. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery*, 5(11):1210–1223, 2015.

[10] Francesco Iorio, Theo A. Knijnenburg, Daniel J. Vis, Graham R. Bignell, Michael P. Menden, Michael Schubert, Nanne Aben, Emanuel Gonçalves, Syd Barthorpe, Howard Lightfoot, Thomas Cokelaer, Patricia Greninger, Ewald van Dyk, Han Chang, Heshani de Silva, Holger Heyn, Xianming Deng, Regina K. Egan, Qingsong Liu, Tatiana Mironenko, Xeni Mitropoulos, Laura Richardson, Jinhua Wang, Tinghu Zhang, Sebastian Moran, Sergi Sayols, Maryam Soleimani, David Tamborero, Nuria Lopez-Bigas, Petra Ross-Macdonald, Manel Esteller, Nathanael S. Gray, Daniel A. Haber, Michael R. Stratton, Cyril H. Benes, Lodewyk F.A. Wessels, Julio Saez-Rodriguez, Ultan McDermott, and Mathew J. Garnett. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016.

[11] Jean-Pierre Gillet, Sudhir Varma, and Michael M. Gottesman. The Clinical Relevance of Cancer Cell Lines. *JNCI: Journal of the National Cancer Institute*, 105(7):452–458, 02 2013.

[12] Michael Gottesman. Mechanisms of cancer drug resistance. *Annual review of medicine*, 53:615–27, 02 2002.

[13] Behzad Mansoori, Ali Mohammadi, Sadaf Davudian, Solmaz Shirjang, and Behzad Baradaran. The different mechanisms of cancer drug resistance: A brief review. *Adv Pharm Bull*, 7(3):339–348, 2017.

[14] K. B., Naiara Orrego-Lagarón, Eileen Mcgowan, Indu Parmar, Amitabh Jha, Basil Hubbard, and H P Vasantha Rupasinghe. Kinase-targeted cancer therapies: Progress, challenges and future directions. *Molecular Cancer*, 17, 02 2018.

[15] H.C. Stephen Chan, Hanbin Shan, Thamani Dahoun, Horst Vogel, and Shuguang Yuan. Advancing drug discovery via artificial intelligence. *Trends in Pharmacological Sciences*, 40(8):592–604, 2019. Special Issue: Rise of Machines in Medicine.

[16] Joseph A. DiMasi, Henry G. Grabowski, and Ronald W. Hansen. Innovation in the pharmaceutical industry: New estimates of r&d costs. *Journal of Health Economics*, 47:20–33, 2016.

[17] Steven M Paul, Daniel S Mytelka, Christopher T. Dunwiddie, Charles C. Persinger, Bernard H. Munos, Stacy R Lindborg, and Aaron Leigh Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9:203–214, 2010.

[18] Joe Greener, Lewis Moffat, and David Jones. Design of metalloproteins and novel protein folds using variational autoencoders. *Scientific Reports*, 8, 11 2018.

[19] Brian L. Hie and Kevin K. Yang. Adaptive machine learning for protein engineering. *Current Opinion in Structural Biology*, 72:145–152, 2022.

[20] Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2323–2332. PMLR, 10–15 Jul 2018.

[21] Wengong Jin, Dr.Regina Barzilay, and Tommi Jaakkola. Hierarchical generation of molecular graphs using structural motifs. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4839–4848. PMLR, 13–18 Jul 2020.

[22] Donatas Repecka, Vykintas Jauniskis, Laurynas Karpus, Elzbieta Rembeza, Irmantas Rokaitis, Jan Zrimec, Simona Poviloniene, Audrius Laurynenas, Sandra Viknander, Wissam Abuajwa, Otto Savolainen, Rolandas Meškys, Martin Engqvist, and Aleksej Zelezniak. Expanding

functional protein sequence spaces using generative adversarial networks. *Nature Machine Intelligence*, 3:1–10, 04 2021.

[23] Paulina Szymczak, Marcin Możejko, Tomasz Grzegorzek, Marta Bauer, Damian Neubauer, Michał Michalski, Jacek Sroka, Piotr Setny, Wojciech Kamysz, and Ewa Szczurek. Hydramp: a deep generative model for antimicrobial peptide discovery. 2022.

[24] Joshua Meyers, Benedek Fabian, and Nathan Brown. De novo molecular design and generative models. *Drug Discovery Today*, 26(11):2707–2715, 2021.

[25] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[26] Mi Yang, Jaak Simm, Chi Chung Lam, Pooya Zakeri, Gerard J. P. van Westen, Yves Moreau, and Julio Saez-Rodriguez. Linking drug target and pathway activation for effective therapy using multi-task learning. *Scientific Reports*, 8, 12 2018.

[27] J. Simm, A. Arany, P. Zakeri, T. Haber, J. K. Wegner, V. Chupakhin, H. Ceulemans, and Y. Moreau. Macau: Scalable bayesian factorization with high-dimensional side information using mcmc. In *2017 IEEE 27th International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2017.

[28] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. Neural collaborative filtering, 2017.

[29] Shuai Zhang, Lina Yao, Aixin Sun, and Yi Tay. Deep learning based recommender system: A survey and new perspectives. 52(1), February 2019.

[30] Qiao Liu, Zhiqiang Hu, Rui Jiang, and Mu Zhou. DeepCDR: a hybrid graph convolutional network for predicting cancer drug response. *Bioinformatics*, 36(Supplement_2):i911–i918, 12 2020.

[31] Krzysztof Koras, Ewa Kizling, Dilafruz Juraeva, Eike Staub, and Ewa Szczurek. Interpretable deep recommender system model for prediction of kinase inhibitor efficacy across cancer cell lines. *Scientific reports*, 11:15993, 2021.