Zero-Shot Prediction of Drug Combination Activity for High-Throughput Screens

Karl Pichotta [1] **Wesley Tansey** [1]

Abstract

Biologists use high-throughput drug screens to evaluate how pairs of drugs will interact. The paired drug search space grows quadratically with the number of drugs considered. Predictive models can help trim this large search space by allowing biologists to prioritize the most promising candidates. To this end, we develop and evaluate four different deep learning approaches to predicting drug synergy. We find structural molecular descriptors lead to the best performing synergy prediction models, both for settings with bioactivity observations of all molecules and for zero-shot settings with unseen molecules. The results illustrate the potential of deep learning models to help design more cost- and time-effective combination drug screens.

1. Introduction

Combination therapies–combining multiple drugs at different doses–represent some of the most effective anti-cancer treatments (Jia et al., 2009; Al-Lazikani et al., 2012). One key roadblock to discovering novel effective drug combinations is the exponential growth of the space of possible multi-compound combinations. Exhaustive experimental measurements of the combination space, even in modern high-throughput assay setups, are often impossible (Sun et al., 2013). To keep the experiment design feasible, often the scientist must choose to focus on a small library of drugs up-front. This presents an opportunity for predictive models to help guide the library construction by suggesting novel drug combinations which may have high efficacy (Weiss et al., 2015; Nowak-Sliwinska et al., 2016).

To maximally prune the combination search space, predictive models need to generalize even to molecules without any measurements. Such models typically leverage structural information of molecules, such as a graphical representation of the atomic structures. To-date, most work has focused only on single-drug predictions or on predicting novel combinations of previously-tested drugs. Here, we consider a number of neural models trained to predict the effects on cell viability of unseen drugs in combination in a high-throughput preclinical assay. In particular, we train and evaluate on a corpus of 108,259 drug response measurements of 2,025 distinct drug concentrations applied to 125 distinct cancer cell lines Jaaks et al. (2022).

We consider a number of representations derived from the graph topology of the molecular structure of compounds. We probe the extent to which these different molecular representation schemes enable modeling and prediction of drug synergies across cell lines. We find that simple structural molecular fingerprints improve predictive performance of neural models of drug synergy compared to the neural featurization schemes tested in settings where molecular features are fixed, combination-agnostic, and end-task-agnostic. These results hold both for compounds with some measurements present in the train set (Section 4.1) and in the zero-shot setting of predicting drug synergy measurements for unseen molecules (Section 4.2). These initial investigations suggest neural models leveraging simple molecular fingerprints have the potential to help scientists design better combination drug screens.

2. Related Work

Several methods apply neural networks to *in vitro* drug efficacy prediction (e.g. Mayr et al., 2018; Yang et al., 2019; Tansey et al., 2021). Neural models which take SMILES strings as input have demonstrated strong performance on multiple related tasks including single-molecule bioactivity predictions (Xue et al., 2021; Chithrananda et al., 2020; Gómez-Bombarelli et al., 2018; Bjerrum & Sattarov, 2018; Honda et al., 2019; Ross et al., 2021) and atom-mapping prediction in chemical reactions (Schwaller et al., 2021) .

Techniques for predicting multi-drug interactions in preclinical assays have also been studied, many using deep learning models. Li et al. (2017) find drug structure information is empirically beneficial to synergy prediction in a random

[1]Memorial Sloan Kettering Cancer Center, New York, USA. Correspondence to: Karl Pichotta <pichottk@mskcc.org>.

The 2022 ICML Workshop on Computational Biology.

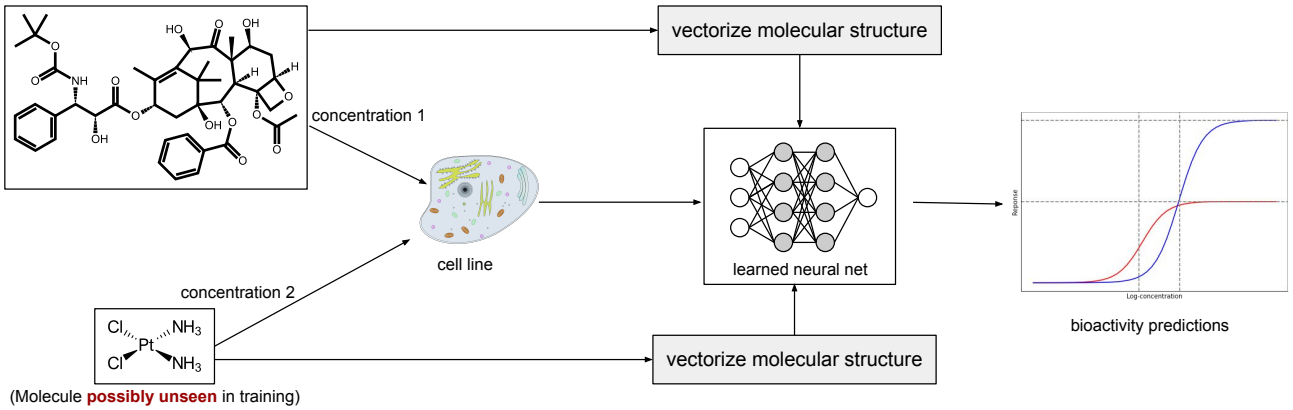

Figure 1. System setup. Different cell lines are exposed to pairs of small molecules at varying doses; classifiers are trained to predict cell survival in vitro. At inference time, we may probe the behavior of unseen molecules, possibly conditioning only on molecular structure (that is, without any bioactivity measurements for the unseen molecules). Different system setups vary in how molecules are vectorized.

forest model. Xia et al. (2018) present a model of the NCI-ALMANAC dataset (Holbeck et al., 2017) which fruitfully incorporates chemical descriptors but focuses on in-sample drugs. Preuer et al. (2018) present a method which incorporates both chemical descriptors and genomic information for predicting Loewe synergy values of combinations; however, their model is shown to perform poorly on unseen drugs. We defer to Adam et al. (2020) for a more complete survey on the broader field of drug response modeling, including in the multi-drug setting.

3. Methods

3.1. Data

We train and test on the drug combination high-throughput-assay dataset of Jaaks et al. (2022), comprising measurements from cell viability screens for 2,025 distinct drug combinations across differing concentrations for 65 different drugs applied to 125 distinct cancer cell lines (51 breast, 45 colorectal, and 29 pancreatic). Measurements are taken in a 2×7 "anchored" approach, with each drug combination having an "anchor" compound tested at 2 concentrations and a "library" compound tested across 7 concentrations. This dataset contains 296,707 combination measurements for 108,259 combination-cell-line pairs.

3.2. Prediction Targets

We follow Jaaks et al. (2022) in analyzing two distinct, complementary notions of pairwise synergy: differences in potency and efficacy (Meyer et al., 2019). Differences in *potency*, ΔIC_{50}, represent the discrepancy between measurement of the concentration for a 50% reduction in cell viability, compared to baseline expectations with independent behavior (in log-concentration space). Differences in maximum *efficacy*, ΔE_{max}, represent the discrepancy between a combination's maximum effect on cell-line viability compared to a baseline independent-activity model. (in cell-percentage viability scale, between -1 and 1). The baseline independent-activity model is a Bliss potency model (Bliss, 1939) fitted to the data. We use the fitted synergy values published by Jaaks et al. (2022) as target variables.

3.3. Training and Evaluation Setups

We evaluate in two settings. First (Sec. 4.1), we evaluate on the task of predicting synergy where all compounds are present in the training set, holding out some combinations and evaluating on the ability to predict measured synergy. Second (Sec. 4.2), we evaluate on the more challenging task of zero-shot prediction for unseen compounds, holding out all of a test compound's measurements from the train set.

3.4. Compound Representations

We consider four molecular representation schemes:

- **Learned**: randomly-initialized continuous embeddings, one per compound. This setup ignores all structural information.
- **ECFP**: extended-connectivity fingerprints (ECFP, Rogers & Hahn, 2010), which hash local neighborhood substructures of a molecule's graph structure and represent molecules as fixed-length bit vectors representing the set of all such local graph substructures in a molecule. These are sometimes called *Morgan fingerprints* or *circular fingerprints*.
- **mol2vec**: the mol2vec representation of Jaeger et al.

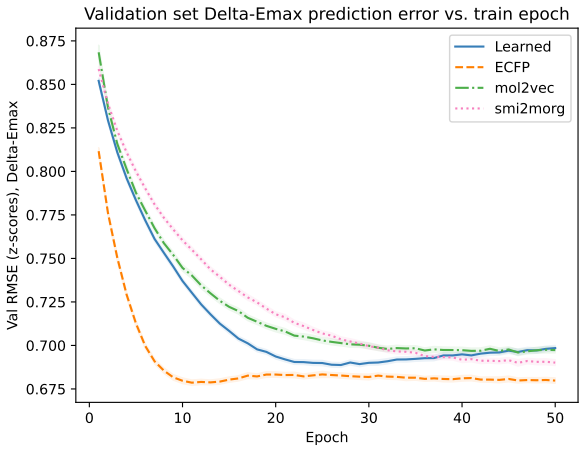

Figure 2. Validation-set efficacy loss: root-mean-square error (RMSE) on ΔE_{max} prediction versus training time for differing molecular representation schemes (lower is better).

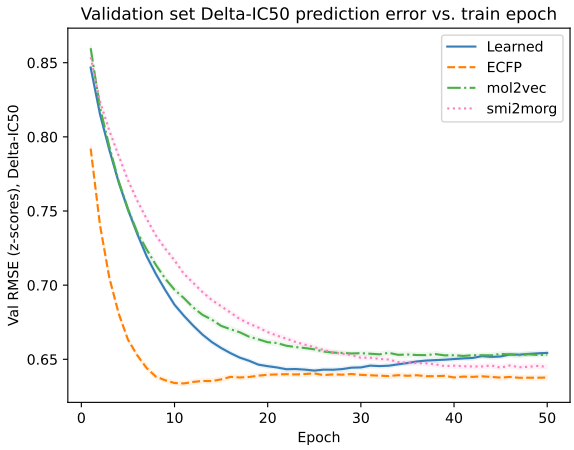

Figure 3. Validation-set potency loss: root-mean-square error (RMSE) on ΔIC_{50} prediction versus training time for differing molecular representation schemes (lower is better).

(2018), in which a word2vec model (Mikolov et al., 2013) is trained on ECFP features of a corpus, and a molecule is represented by the sum of its vectors.

- **smi2morg**: the pooled top hidden embedding layer of a transformer (Vaswani et al., 2017) pre-trained to encode a molecule's raw SMILES string (Weininger, 1988) and decode a multi-hot vector representation of the multiset of its ECFP/Morgan features.

Details on feature preprocessing and smi2morg training may be found in Appendices A and B. The smi2morg system has a similarly motivated setup to multiple published systems with unsupervised pre-training of molecular representations from large corpora of the unlabeled chemical structures of molecular species (Goh et al., 2018; Bjerrum & Sattarov, 2018; Winter et al., 2019).

3.5. Models and Training

All systems use 3-layer feedforward neural networks with 1028-dimension hidden units and ReLU activations. The two drugs are each vectorized using one of the four methods described in Sec. 3.4; these are concatenated to learned embeddings of cell-lines (one embedding per cell line) and anchor drug concentration (one per concentration). Predictors of ΔIC_{50} and ΔE_{max} are trained separately to minimize ℓ_2 regression loss. We optimize weights with Adam (Kingma & Ba, 2014) with a learning rate of 1×10^{-3} and batch size of 256. Hyperparameters were selected from preliminary held-out validation experiments with learned molecule embeddings and fixed for all settings.

4. Results

4.1. Bioactivity Modeling of Known Compounds

We first hold out a random 10% of train data to evaluate the models in the setting in which all compounds are observed in the train set. Figures 2 and 3 give the root-mean-square error curves for predictions of the two synergy targets, ΔE_{max} and ΔIC_{50}, respectively, on held-out validation sets. ECFP generally outperforms the other setups, including the pre-trained transformer model. The learned embedding models overfit before the other representations, reflecting in part the greater number of parameters in the setup.

4.2. Bioactivity Modeling of Unseen Compounds

We next evaluate by holding out each compound entirely from the train set and evaluating against a system trained only on measurements for other drugs. Figures 4 and 5 give recall of high-ranked predictions of synergies for efficacy and potency, respectively. Predictions are sorted by magnitude and truncated to a smaller set (we truncate to $k = 100$). Recall is then calculated against the set of the 100 experiments with highest measured synergy. We do not give results for the *learned* system as we do not incorporate a method for representing unseen compounds Curves are averages across the 65 hold-one-out experiments.

The simpler ECFP features consistently outperform the other molecular representations on the tasks, including the large pre-trained neural smi2morg encoder. The baseline recall-at-100 of ordering concentration-cell-line setups uniformly at random is 0.020 (not pictured). All trained systems evaluated offer considerably performance improvements to

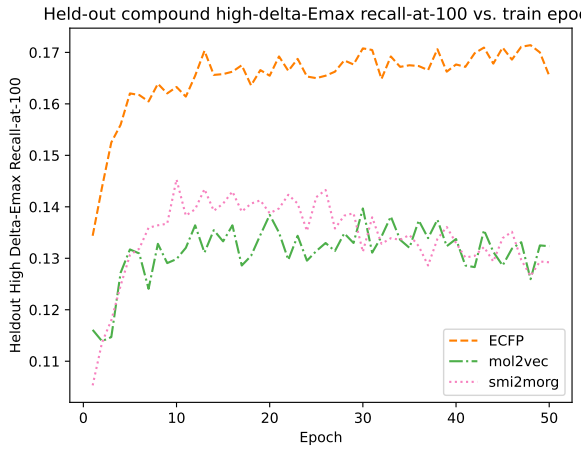

Figure 4. Zero-shot efficacy synergy prediction: macro-average recall-at-100 of top-ranked ΔE_{max} system predictions (higher is better).

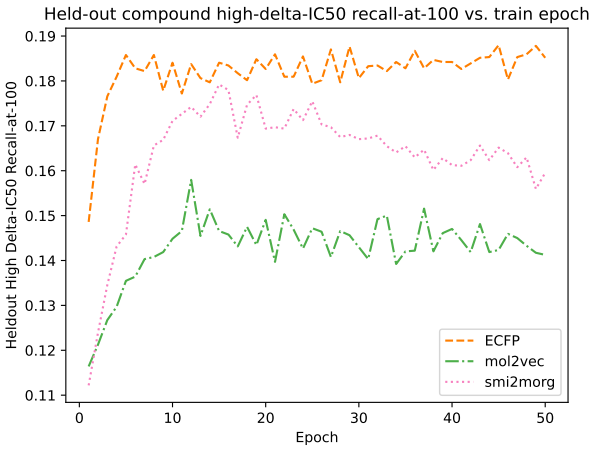

Figure 5. Zero-shot potency synergy prediction: macro-average recall-at-100 of top ΔIC_{50} system predictions for differing molecular representation schemes (higher is better).

this uninformed zero-shot baseline.

4.3. Stratifying Held-out Performance by Drug Target

To determine which classes of compounds differ most in zero-shot performance, we quantify the mean difference in squared-error predictions between smi2morg and ECFP systems. That is, for a model trained on a given held-out compound, for every held-out evaluation setting x for that compound (comprising two compounds, an anchor concentration, and a cell-line), we calculate $SE_s(x)$, the squared error for the smimorg-aware predictor, $SE_e(x)$, the squared error of the ECFP-aware predictor, and average across all differences $SE_s(x) - SE_e(x)$. For each held-out compound we thus get a measure of how much more accurate zero-shot predictions are using ECFP compared to smi2morg. We then stratify by drug target to see if different classes of drugs perform differently.

Across both ΔIC_{50} and ΔE_{max}, we observe drugs targeting RTK signaling perform better with ECFP features (mean difference-in-squared-error 0.393 and 0.450 on efficacy and potency, respectively), particularly the two FGFR inhibitors PD173074 and AZD4547, both of which appear in the top 5 compounds ranked by prediction-error difference. All four EGFR inhibitors in the dataset also see improvement (mean difference of 0.7186 and 0.312 on efficacy and potency). In contrast, the two WNT-targeting drugs see similar performance between the two systems (mean differences 0.0411 and -0.0751 on efficacy and potency). We suspect the former two groups may have pharmacophores strongly associated with synergy predictability which are well-characterized by the discrete ECFP features while the latter do not, but more analysis is required for clear conclusions about which pharmacophores are not being learned from the distributed representations of the neural-representation systems.

5. Conclusions and Future Work

We have presented results indicating that structural molecular information can be useful for synergy prediction both for compounds with observed activity data and for unseen compounds. Pre-trained transformers typically improve from being jointly fine-tuned on downstream tasks of practitioner interest (Devlin et al., 2019). Though we have focused on comparing different fixed molecular featurization schemes, the smi2morg system may benefit from being adapted to a fine-tuning setup. It is possible that a multitask setup (Ramsundar et al., 2015) could be used to leverage heterogeneous supervision signals to improve performance. Directly incorporating genetic and epigenetic features for cell lines would likely increase model performance. The extent to which the above results transfer to neural models of raw measurements (as opposed to predicting pre-fitted synergy models) is an open question. Though the present work compares the utility of different fixed vector-valued representations of molecules, comparing to different means of integrating fully differentiable graph-convolution based systems (Kearnes et al., 2016; Gilmer et al., 2017) is an important area of future work. Finally, more work is required to develop a robust platform for designing combination drug screens using predictive models.

References

Adam, G., Rampášek, L., Safikhani, Z., Smirnov, P., Haibe-Kains, B., and Goldenberg, A. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, 4(1):1–10, 2020.

Al-Lazikani, B., Banerji, U., and Workman, P. Combinatorial drug therapy for cancer in the post-genomic era. *Nature biotechnology*, 30(7):679–692, 2012.

Bjerrum, E. J. and Sattarov, B. Improving chemical autoencoder latent space and molecular de novo generation diversity with heteroencoders. *Biomolecules*, 8(4):131, 2018.

Bliss, C. I. The toxicity of poisons applied jointly. *Annals of Applied Biology*, 26:585–615, 1939.

Chithrananda, S., Grand, G., and Ramsundar, B. ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *North American Association for Computational Linguistics (NAACL)*, 2019.

Gaulton, A., Bellis, L. J., Bento, A. P., Chambers, J., Davies, M., Hersey, A., Light, Y., McGlinchey, S., Michalovich, D., Al-Lazikani, B., et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research*, 40:D1100–D1107, 2012.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 2017.

Goh, G. B., Siegel, C., Vishnu, A., and Hodas, N. Using rule-based labels for weak supervised learning: A ChemNet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 302–310, 2018.

Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. Automatic chemical design using a data-driven continuous representation of molecules. *ACS Central Science*, 4(2):268–276, 2018.

Holbeck, S. L., Camalier, R., Crowell, J. A., Govindharajulu, J. P., Hollingshead, M. G., Anderson, L. W., Polley, E. C., Rubinstein, L. V., Srivastava, A. K., Wilsker, D. F., Collins, J. M., and Doroshow, J. H. The National Cancer Institute ALMANAC: A comprehensive screening resource for the detection of anticancer drug pairs with enhanced therapeutic activity. *Cancer research*, pp. 3564–3576, 2017.

Honda, S., Shi, S., and Ueda, H. R. SMILES transformer: Pre-trained molecular fingerprint for low data drug discovery. *arXiv preprint arXiv:1911.04738*, 2019.

Irwin, J. J. and Shoichet, B. K. ZINC – a free database of commercially available compounds for virtual screening. *Journal of chemical information and modeling*, 45(1): 177–182, 2005.

Jaaks, P., Coker, E. A., Vis, D. J., Edwards, O., Carpenter, E. F., Leto, S. M., Dwane, L., Sassi, F., Lightfoot, H., Barthorpe, S., van der Meer, D., Yang, W., Beck, A., Mironenko, T., Hall, C., Hall, J., Mali, I., Richardson, L., Tolley, C., Morris, J., Thomas, F., Lleshi, E., Aben, N., Benes, C. H., Bertotti, A., Trusolino, L., Wessels, L. F. A., and Garnett, M. J. Effective drug combinations in breast, colon and pancreatic cancer cells. *Nature*, 603: 166–173, 2022.

Jaeger, S., Fulle, S., and Turk, S. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35, 2018.

Jia, J., Zhu, F., Ma, X., Cao, Z. W., Li, Y. X., and Chen, Y. Z. Mechanisms of drug combinations: interaction and network perspectives. *Nature reviews Drug discovery*, 8 (2):111–128, 2009.

Kearnes, S., McCloskey, K., Berndl, M., Pande, V., and Riley, P. Molecular graph convolutions: Moving beyond fingerprints. *Journal of computer-aided molecular design*, 30(8):595–608, 2016.

Kim, S., Thiessen, P. A., Bolton, E. E., and Bryant, S. H. PUG-SOAP and PUG-REST: Web services for programmatic access to chemical information in PubChem. *Nucleic acids research*, 43(W1):W605–W611, 2015.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Landrum, G. RDKit: Open-source cheminformatics, 2016. URL http://www.rdkit.org.

Li, X. and Fourches, D. SMILES pair encoding: a data-driven substructure tokenization algorithm for deep learning. *Journal of Chemical Information and Modeling*, 61 (4):1560–1569, 2021.

Li, X., Xu, Y., Cui, H., Huang, T., Wang, D., Lian, B., Li, W., Qin, G., Chen, L., and Xie, L. Prediction of

synergistic anti-cancer drug combinations based on drug target network and drug induced gene expression profiles. *Artificial Intelligence in Medicine*, 83:35–43, 2017.

Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chemical science*, 9(24):5441–5451, 2018.

Meyer, C. T., Wooten, D. J., Paudel, B. B., Bauer, J., Hardeman, K. N., Westover, D., Lovly, C. M., Harris, L. A., Tyson, D. R., and Quaranta, V. Quantifying drug combination synergy along potency and efficacy axes. *Cell systems*, 8(2):97–108, 2019.

Mikolov, T., Chen, K., Corrado, G. S., and Dean, J. Efficient estimation of word representations in vector space. In *ICLR*, 2013.

Nowak-Sliwinska, P., Weiss, A., Ding, X., Dyson, P. J., Van Den Bergh, H., Griffioen, A. W., and Ho, C.-M. Optimization of drug combinations using feedback system control. *Nature protocols*, 11(2):302–315, 2016.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, pp. 8024–8035. 2019.

Preuer, K., Lewis, R. P., Hochreiter, S., Bender, A., Bulusu, K. C., and Klambauer, G. DeepSynergy: Predicting anti-cancer drug synergy with deep learning. *Bioinformatics*, 34(9):1538–1546, 2018.

Ramsundar, B., Kearnes, S., Riley, P., Webster, D., Konerding, D., and Pande, V. Massively multitask networks for drug discovery. *arXiv preprint arXiv:1502.02072*, 2015.

Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., and Wu, Z. *Deep Learning for the Life Sciences*. O'Reilly Media, 2019.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of chemical information and modeling*, 50 (5):742–754, 2010.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Do large scale molecular language representations capture important structural information? *arXiv preprint arXiv:2106.09553*, 2021.

Schwaller, P., Hoover, B., Reymond, J.-L., Strobelt, H., and Laino, T. Extraction of organic chemistry grammar from unsupervised learning of chemical reactions. *Science Advances*, 7(15):eabe4166, 2021.

Sun, X., Vilar, S., and Tatonetti, N. P. High-throughput methods for combinatorial drug discovery. *Science translational medicine*, 5(205):205rv1–205rv1, 2013.

Tansey, W., Li, K., Zhang, H., Linderman, S. W., Rabadan, R., Blei, D. M., and Wiggins, C. H. Dose-response modeling in high-throughput cancer drug screenings: A case study with recommendations for practitioners. *Biostatistics*, 2021. PMC Journal - In Process.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Weininger, D. 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.

Weiss, A., Ding, X., van Beijnum, J. R., Wong, I., Wong, T. J., Berndsen, R. H., Dormond, O., Dallinga, M. G., Shen, L., Schlingemann, R. O., Pili, R., Ho, C.-M., Dyson, P. J., van den Bergh, H., Griffioen, A. W., and Nowak-Sliwinska, P. Rapid optimization of drug combinations for the optimal angiostatic treatment of cancer. *Angiogenesis*, 18(3):233–244, 2015.

Winter, R., Montanari, F., Noé, F., and Clevert, D.-A. Learning continuous and data-driven molecular descriptors by translating equivalent chemical representations. *Chemical science*, 10(6):1692–1701, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. HuggingFace's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

Xia, F., Shukla, M., Brettin, T. S., Garcia-Cardona, C., Cohn, J. D., Allen, J. E., Maslov, S., Holbeck, S. L., Doroshow, J. H., Evrard, Y. A., Stahlberg, E. A., and Stevens, R. L. Predicting tumor cell line response to drug pairs with deep learning. *BMC Bioinformatics*, 19, 2018.

Xue, D., Zhang, H., Xiao, D., Gong, Y., Chuai, G., Sun, Y., Tian, H., Wu, H., Li, Y., and Liu, Q. X-MOL: large-scale pre-training for molecular understanding and diverse molecular analysis. *bioRxiv*, 2021.

Yang, K., Swanson, K., Jin, W., Coley, C. W., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K. F., and Barzilay, R. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59:3370–3388, 2019.

A. System Details

The smi2morg model is a 16-layer multiheaded attention model with 8 attention heads and a width of 1024, trained using Adam (Kingma & Ba, 2014) with learning rate 3×10^{-1} and a batch size of 512 (using gradient accumulation). The learning rate is linearly warmed up for 1000 epochs. The system is implemented using the HuggingFace framework (Wolf et al., 2019) and written in PyTorch (Paszke et al., 2019). Input SMILES strings (encoded during pre-training) are tokenized using the atom-level tokenizer from the SmilesPE library (Li & Fourches, 2021). The multi-set of Morgan features (decoded during pre-training) is a 500k-dimensional binary vector, with a dimension representing a particular Morgan feature occurring a particular number of times in a molecule (that is, a dimension has the semantics that, for some Morgan substructure m and some non-negative integer k, that dimension is 1 iff m occurs k distinct times in the molecule); the top 500k such (substructure, count) pairs are calculated from the pre-training corpus during preprocessing. Embeddings are pooled by representing a molecule by the top hidden activation values at a reserved `[CLS]` meta-token prepended to each sequence, a standard pooling technique for BERT (Devlin et al., 2019) and related transformer networks.

ECFP bit-vectors are of length 1024 (that is, feature hash values are vectorized mod 1024). Learned molecule embeddings are also length 1024. Concentration embeddings and cell-line embeddings are of dimension 4 and 32, respectively.

B. Data Preprocessing Details

ECFP features for all systems are calculated by RDKit (Landrum, 2016) with radius 2. We use the DeepChem library (Ramsundar et al., 2019) to calculate mol2vec features. SMILES Strings are canonicalized using RDKit. We pre-train the smi2morg system on 20M unlabeled compounds from ChEMBL (Gaulton et al., 2012) and ZINC (Irwin & Shoichet, 2005).

Of the 65 compounds trialed in the data set of Jaaks et al. (2022), one appears to be a mixture of two separate compounds (Afatinib and Trametinib, keyed in the data as ID `1032|1372`). We discard all observations for this mixture, as it requires 3 molecules rather than 2 as input. SMILES strings for molecules were collected using the PUG-REST API for PubChem data (Kim et al., 2015).

All regression targets are standardized before training. Log-concentrations of anchor compounds are represented categorically (that is, distinct concentrations correspond to distinct one-hot vectors parameterizing lookups in the concentration-embeddings matrix).