# Causal-Chemprop: Causal Machine Learning for Molecular Property Prediction and Design

Christian Natajaya[1], Lucas Attia[2], Jackson Burns[2], and Patrick S. Doyle[2]

[1]Neopoly Ltd, London, United Kingdom
[2]Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA

## Abstract

A priori estimation of molecular properties has long been of immense interest to the pharmaceutical sciences for molecular generation and optimization. While neural network-based models have achieved high predictive accuracy, they still find limited utility in molecular design. High-dimensional molecular representations are difficult to optimize especially in small data regimes, and neural network-based models lack mechanisms to explicitly incorporate domain knowledge from experts and prior knowledge from existing data. Herein, we introduce a causal machine learning framework built on the Chemprop and DAGMA architectures for molecular property prediction called Causal-Chemprop. To our knowledge, this is the first application of causal machine learning to molecular property prediction and optimization. Via intervention-based inference, Causal-Chemprop demonstrates strong predictive performances on $IC_{50}$ from the Kinase Knowledgebase and $logS$ of aqueous solutions from BigSolDB and SolProp. Counterfactual-based inference offers support for human-in-the-loop optimization of molecular structure, which we demonstrate by accurately extrapolating the $IC_{50}$ of out-of-distribution kinase inhibitors and $logS$ of a quinolinyltriazole series of MIF inhibitors given a seed structure. Finally, we integrate Causal-Chemprop with the molecular optimization algorithm EvoMol to perform inverse molecular design, yielding soluble analogs of the quinolinyltriazole MIF inhibitor.

## 1 Introduction

In pharmaceutical development, accurate prediction of physicochemical and biological properties is essential for hit generation and optimization. Graph neural networks (GNN) models like Chemprop have achieved state-of-the-art performance on a variety of property prediction tasks including solubility [1], absorption, distribution, metabolism, and excretion [5]. Despite their success on benchmark datasets, GNN models find limited utility in molecular optimization.

High-dimensional GNN molecular representations tend to overfit and learn spurious correlations on small datasets, limiting their extrapolative accuracy on out-of-distribution (OOD) molecules [17]. Further, GNN models lack explicit mechanisms to incorporate expert domain knowledge, which could otherwise help uncover relationships buried in noise or model the noise itself. GNN models also lack explicit mechanisms to incorporate prior knowledge from existing experimental evidence, which could otherwise help make more accurate predictions.

Here, we propose a causal machine learning framework built on the Chemprop [7] and DAGMA [3] architectures for molecular property prediction called Causal-Chemprop. *Figure 1 (a)* illustrates the Causal-Chemprop model architecture. Instead of passing learned molecular representations to a feed-forward network for property prediction, we pass it into a structural causal model (SCM) on which we perform intervention- and counterfactual- based inferences. To our knowledge, this is the first application of causal machine learning to molecular property prediction and optimization, and we believe this approach can be generalizable to other representation-based models.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

The SCM facilitates targeted do-interventions on features of the molecular representation, as illustrated in *Figure 1 (b)*. We propose that estimating the target property under these interventional effects reduces the impact of spurious correlations found in high-dimensional representations. Further, the SCM offers a pathway beyond naive black-box property predictions through counterfactual reasoning, as illustrated in *Figure 1 (c)*. By observing a molecule or a molecular cluster and its properties, we can isolate noise variables associated with it, then use this prior knowledge to estimate counterfactual outcomes of the target property under interventional effects. This enables more reliable property prediction around the observed chemical space.

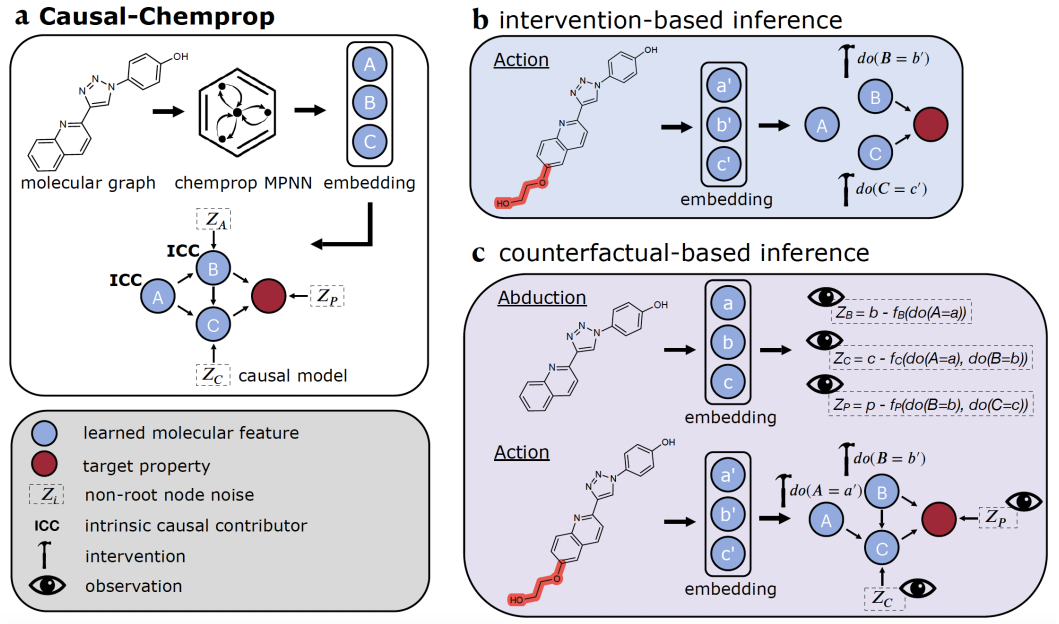



Figure 1: (a) Causal-Chemprop is a structural causal model built on top of Chemprop; (b) do-interventions gives accurate predictions of target properties; (c) counterfactual reasoning makes use of prior evidences to get better predictions, enabling human-in-the-loop molecular optimization.

We compare the performances of Chemprop and Causal-Chemprop on 3 tasks:

1. **Single-component tasks**: We train Causal-Chemprop on an AURKA inhibitor dataset (containing 810 datapoints) and an ABL1 inhibitor dataset (containing 615 datapoints) taken from the Kinase Knowledgebase [15], then apply intervention-based inference for model testing. To test OOD extrapolation, we then perform counterfactual-based inference on the pretrained model to predict the $IC_{50}$ values of AURKA inhibitors (*Figure 11*) conceptualized by Bavetsias, et al. [2] and ABL1 inhibitors (*Figure 12*) conceptualized by Huang, et al. [9]
2. **Multi-component tasks**: We train Causal-Chemprop on an aqueous solution dataset curated from BigSolDB [11] and SolProp [16], containing 3630 datapoints and 423 solutes, then apply intervention-based inference for model testing. To test OOD extrapolation, we then perform counterfactual-based inference on the pretrained model to predict the $logS$ of the molecular derivatives of a quinolinyltriazole MIF inhibitor (*Figure 10*) conceptualized by Cisneros, et al. [4].
3. **Molecular optimization**: We use an evolutionary algorithm, EvoMol [12], to optimize the quinolinyltriazole MIF inhibitor seed structure and maximize its aqueous solubility. As the scoring function for EvoMol, we compared both Chemprop and counterfactual-based inference with the Causal-Chemprop model pretrained for the $logS$ task.

## 2 Results and Discussion

### 2.1 Single-component tasks

*Figure 2* shows the parity plots comparing Chemprop and Causal-Chemprop on ABL1 and AURKA inhibitors from the Kinase Knowledgebase. While Chemprop was unable to correctly rank order the

$IC_{50}$ values of both kinase inhibitors, Causal-Chemprop demonstrated substantial improvements in predictive accuracy using the same molecular embedding via intervention-based inference, particularly for the Causal-Chemprop model with histogram gradient boosting regressor trees. By isolating the feature space to the parent nodes of $IC_{50}$, we obtain improved separation of molecular clusters as corroborated by the t-SNE plots (*Figure 8*).

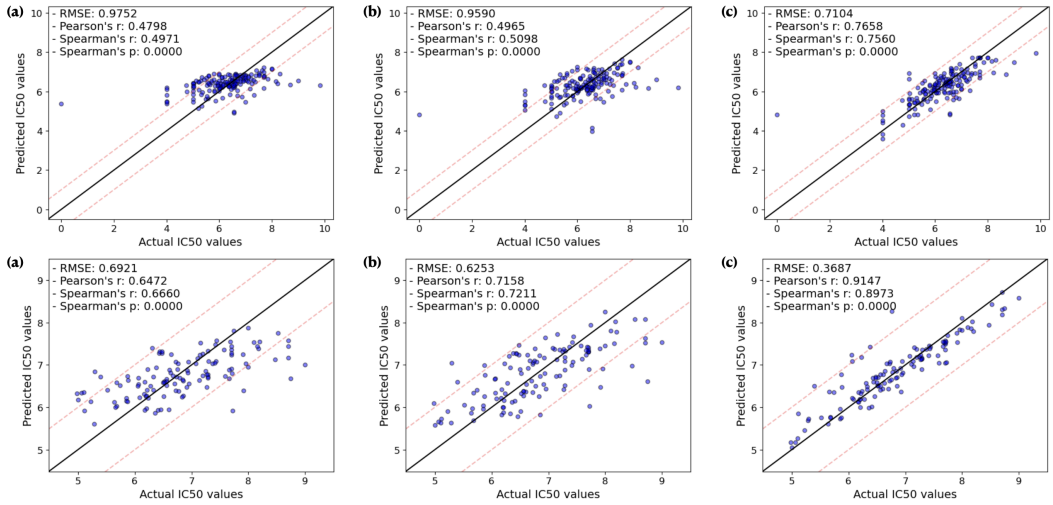



Figure 2: $IC_{50}$ predictions for AURKA (top) and ABL1 (bottom) inhibitors using (a) Chemprop; (b) intervention-based inference on Causal-Chemprop with linear regressors; (c) intervention-based inference on Causal-Chemprop with histogram gradient boosting regressor trees

We use the pretrained models to extrapolate the $IC_{50}$ values of AURKA inhibitors conceptualized by Bavetsias, et al., and ABL1 inhibitors conceptualized by Huang, et al. *Figure 3* shows the parity plots comparing the performances of Chemprop and Causal-Chemprop. Counterfactual-based inference with both Causal-Chemprop models outperformed Chemprop for OOD predictions on both AURKA and ABL1 tasks, as indicated by the improvements in RMSE, Pearson, and Spearman correlation scores. We suspect that intervening only features of the molecular representation that are intrinsic causal contributors to $IC_{50}$ eliminates the effect of spurious correlations.

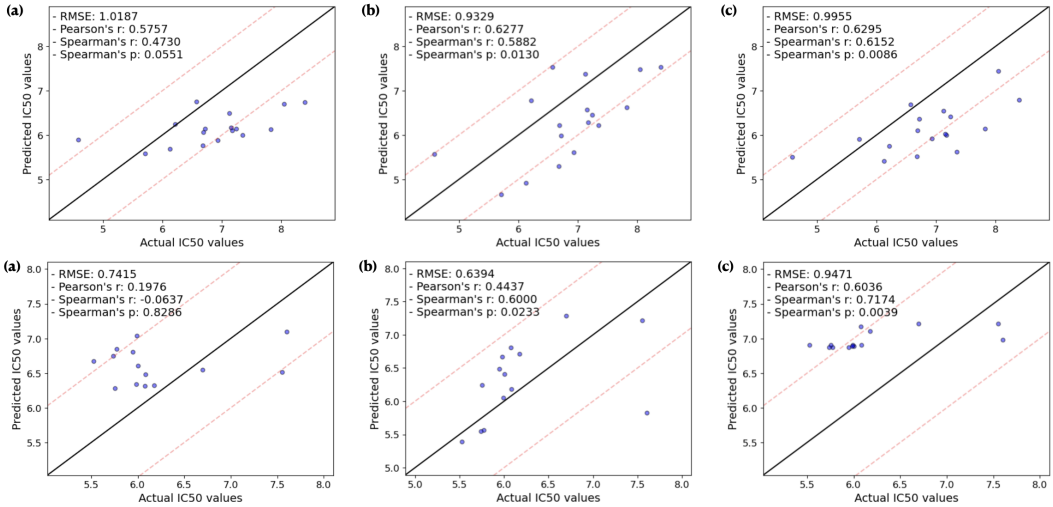



Figure 3: $IC_{50}$ predictions for AURKA (top) and ABL1 (bottom) inhibitors using (a) Chemprop; (b) counterfactual-based inference on Causal-Chemprop with linear regressors; (c) counterfactual-based inference on Causal-Chemprop with histogram gradient boosting regressor trees

## 2.2 Multi-component tasks

The aqueous solubility dataset provides a challenge due to experimental noise. *Figure 4* compares the parity plots of Chemprop and Causal-Chemprop. Unlike Chemprop, both Causal-Chemprop models successfully captured the temperature dependence of solubility via a causal edge, and intervention-based inference with the Causal-Chemprop model showed a particularly significant improvement in rank-ordering of the solutes, particularly with histogram gradient boosting regressor trees. These improvements are corroborated by t-SNE plots across the aqueous solubility dataset (*Figure 9*); utilizing only the features of the molecular representation that are intrinsic causal contributors to $logS$ leads to better separation of chemotypes.

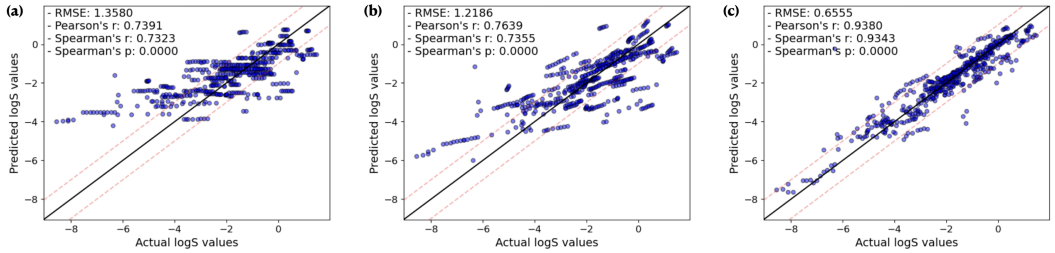



Figure 4: $logS$ predictions on a combined aqueous dataset comprised of BigSolDB and SolProp using (a) Chemprop; (b) intervention-based inference on Causal-Chemprop with linear regressors; (c) intervention-based inference on Causal-Chemprop with histogram gradient boosting regressor trees

The sparsity of the aqueous solubility dataset and differences in solvation mechanisms makes it a challenging task to extrapolate the aqueous solubility of OOD small molecules. Here, we predict the aqueous solubility of the derivatives of a quinolinyltriazole MIF inhibitor seed structure. *Figure 5* shows the parity plots comparing the performances of Chemprop and Causal-Chemprop. By observing a quinolinyltriazole seed structure, counterfactual reasoning allows both Causal-Chemprop models to more accurately predict the aqeuous solubility of the quinolinyltriazole molecular derivatives. We suspect that abducting exogenous noise associated with the quinolinyltriazole seed structure allowed Causal-Chemprop to bias its predictions towards the observed chemical space.

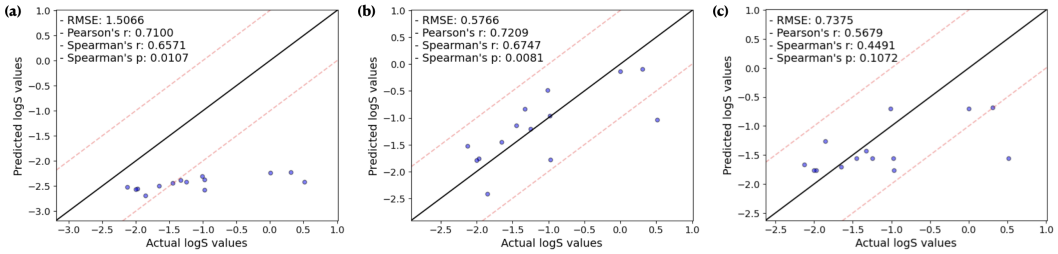



Figure 5: $logS$ predictions on the molecular derivative of a quinolinyltriazole MIF inhibitor seed structure using (a) Chemprop; (b) counterfactual-based inference on Causal-Chemprop with linear regressors; (c) counterfactual-based inference on Causal-Chemprop with histogram gradient boosting regressor trees

## 2.3 Molecular optimization

This advancement beyond naive black-box prediction with counterfactual reasoning enables rational molecular optimization, which we show with an evolutionary algorithm. When Chemprop is deployed as a scoring function for EvoMol, the algorithm frequently perturbs the seed structure, driving unstable optimization and yielding unrealistic molecular structures. In contrast, integrating Causal-Chemprop's counterfactual-based inference as the scoring function yields stable evolutionary trajectories, resulting in soluble analogs of the quinolinyltriazole seed structure that parallel those reported by Cisneros et al. The evolution trajectories are shown in *Figure 6*.

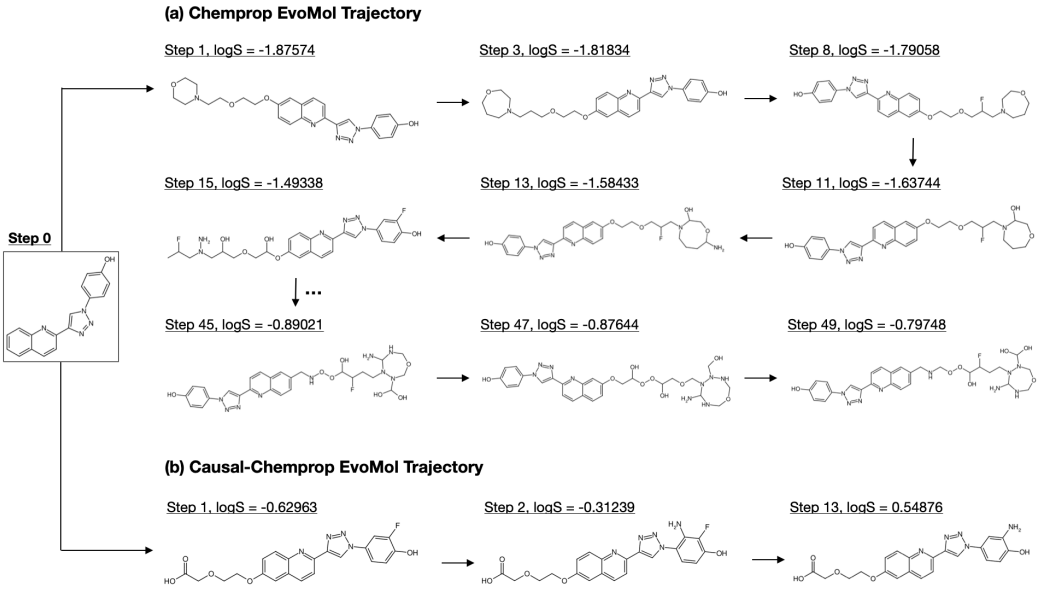



Figure 6: (a) EvoMol trajectory optimizing a quinolinyltriazole MIF inhibitor seed structure over 50 steps using Chemprop as a scoring function and; (b) using counterfactual-based inference on a linear Causal-Chemprop model as a scoring function

# 3 Causal-Chemprop

## 3.1 Chemprop

Chemprop consists of a local features encoding function, a directed message passing neural network to learn atomic embeddings from the local features, an aggregation function to join atomic embeddings into molecular embeddings, and a standard feed-forward neural network for the transformation of molecular embeddings to target properties. Here, the sole purpose of Chemprop is to generate molecular representations, which we use to learn a structural causal model and perform causal inferences on to predict molecular properties.

## 3.2 Structural causal models

Following the framework by Pearl [13], a structural causal model, $M$, consists of two sets of variables, $X = (X_1, \ldots, X_d)$ and $Z = (Z_1, \ldots, Z_d)$, and a set of non-parameterized structural equations, $F = (f_1, \ldots, f_d)$, that assigns the value $x_i$ to each variable $X_i \in X$ in response to the current values of $X$ and $Z$:

$$x_i = f_i(x, z), \quad \forall i \in [d] \tag{1}$$

Here, the variables in X correspond to the Chemprop embedding, molecular properties, and experimental conditions. The variables in $Z$ are considered "exogenous" and correspond to unobserved noise for which no mechanism is encoded in $M$. The family of structural equations $F$ induces a causal graph $G(F)$ that defines a joint distribution $\mathbb{P}(X)$ over the observed data. Here, we assume a set of linear structural equations, which we solve for using DAGMA.

### 3.2.1 Learning the causal graph

The DAGMA algorithm frames the combinatorial problem of learning causal graphs from observed data as a continuous optimization problem. It solves for the set of structural equations $F$ in equation (1) that minimizes a loss function $Q(F; X)$, which measures the quality of a candidate causal graph $G(F)$ against the observed data $X = [x_1, \ldots, x_d]$:

$$\min_{f \in \mathcal{F}} Q(F; X) = \min_{f \in \mathcal{F}} \sum_{i=1}^{d} \text{loss}(x_i, f_i(x)) \quad \text{s.t.} \quad G(F) \in \text{DAGs} \tag{2}$$

The causal graph $G(F)$ is represented as a weighted adjacency matrix $W(F) \in \mathbb{R}^{d\times d}$ with elements $W_{i,j} = \|\partial_j f_i\|_{L^2}$, where $\partial_j f_i$ is the partial derivative of $f_i$ w.r.t. $x_j$. DAGMA defines the log-determinant characterization for acyclicity $h^s_{\text{ldet}}$ as follows:

$$h^s_{\text{ldet}} = -\log\det(sI - W \circ W) + d\log s = 0 \iff W \in \text{DAGs} \tag{3}$$

$$W(\theta^{(0)}) \in \mathbb{W}^s = \left\{W \in \mathbb{R}^{d\times d} \,\middle|\, s > \rho(W \circ W)\right\} \tag{4}$$

DAGMA reformulates the objective function in equation (2) as an unconstrained problem in which $h^s_{\text{ldet}}$ acts as a non-negative regularizer that we seek to minimize along with the loss function.

#### 3.2.2 Learning the causal mechanisms

We then assign and fit causal mechanisms for each node in the causal graph $G(F)$, replacing the set of non-parametrized structural equations $f_i$ in equation (1). Here, root nodes are assigned an empirical distribution that allows us to randomly sample from the provided data. Non-root nodes are assigned an Additive Noise Model [8], which assigns the value $x_i$ to the node $X_i$ following:

$$x_i = f_i(pa_i) + z_i \tag{5}$$

Here, $f_i$ is either a linear regressor or a histogram-based gradient boosting regressor tree, which takes as input the values $pa_i$ for the parents $PA_i$ of $X_i$. We assume the exogenous noise variables $Z_i$ are random variables independent of $PA_i$ and are thus root nodes that we also model with an empirical distribution.

### 3.3 Causal inference

We can now perform causal inferences on the SCM following methods proposed by Pearl [14]. Here, we use intervention-based inference to predict the $IC_{50}$ of AURKA and ABL1 kinase inhibitors, and the aqueous $logS$ of small molecules. We then use counterfactual-based inference to predict the solubility of the molecular derivatives of a quinolinyltriazole MIF inhibitor seed structure, and apply it as the scoring function of an evolutionary algorithm to optimize the seed structure.

#### 3.3.1 Intervention-based inference

An intervention tells us that: "$Y$ would be $y$ if $X$ is $x$," denoted by $Y_{do(x)} = y$. We can estimate the target property under interventional effects as follows:

1. **Action**: Generate the Chemprop embedding then identify the parents $PA_i$ of $IC_{50}$ or $logS$. Perform the atomic intervention $PA_i = do(x_i)$ for each parent, where $x_i$ is its corresponding value in the Chemprop embedding. By performing the do-intervention, we set the values $pa_i = x_i$ in equation (5) and remove any causal influences on $PA_i$, resulting in the submodel $M_x$;
2. **Prediction**: The solution for $IC_{50}$ or $logS$ in the submodel $M_x$ gives us $Y_{do(x)}$; assume $z_i = 0$ for the noise variables then predict the value $x_i$ for $IC_{50}$ or $logS$.

#### 3.3.2 Counterfactual-based inference

A counterfactual sentence tells us that: "$Y$ would be $y$ had $X$ been $x$ in the event $U = u$," denoted by $Y_x(u) = y$. Given any propositional evidence $e$ observed from the event $u$, we can compute the target property in a counterfactual scenario using a three-step process as follows:

1. **Abduction**: Generate and observe the Chemprop embedding, temperature, and $logS$ of the MIF seed structure. Then update the event $U = u$ in light of the propositional evidence $e$; for each node in the SCM, assign the observed value of its parent nodes $pa_i = x_e$, the observed value of itself $x_i = y_e$, then retrieve the event noise of $z_i$ for the node following:

$$z(u) = y_e - f(x_e); \tag{6}$$

2. **Action**: Generate the Chemprop embedding of the MIF molecular derivative. Identify which features of the Chemprop embedding are intrinsic causal contributors of $logS$, then perform the intervention $X = do(x)$, where $X$ are the causal contributors and $x$ is their corresponding value in the Chemprop embedding. This again replaces any causal influences on $X$ with the hypothetical antecedent $X = x$, yielding a modified SCM $M_x$;
3. **Prediction**: Estimate $logS$ of the MIF molecular derivative from the submodel $M_x$, which gives us the solution $Y_x(u)$ based on our updated understanding of event $u$ and the do-interventions:

$$Y_x(u) = f(do(x)) + z(u) \tag{7}$$

The accuracy of counterfactual-based inference with Causal-Chemprop depends on the similarity between the chosen seed structure and the evaluated target structures. Here, we investigate this relationship for the AURKA $IC_{50}$, ABL1 $IC_{50}$, and aqueous $logS$ extrapolative tasks by plotting the prediction RMSE against the Tanimoto similarity between the seed and target structures. This relationship is illustrated in *Figure 7*.

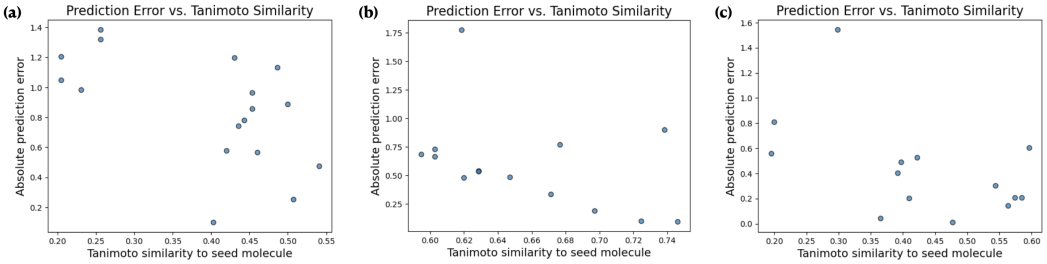



Figure 7: Error in counterfactual-based inference using Causal-Chemprop with linear regressors on the following extrapolative tasks: (a) AURKA inhibitor $IC_{50}$, (b) ABL1 inhibitor $IC_{50}$, (c) MIF inhibitor $logS$

# 4 Methods

## 4.1 Single-component tasks

**Datasets**: We trained single-component Chemprop models on AURKA and ABL1 inhibitor datasets from the Kinase Knowledgebase. For both tasks, the datasets were randomly split into training ($80\%$) and validation ($20\%$) sets. We do not provide the validation set to Causal-Chemprop when learning the SCM. We test intervention-based inference with Causal-Chemprop and compare it to Chemprop by predicting the $IC_{50}$ values on the validation set. We then test counterfactual-based inference and compare it to Chemprop by extrapolating the $IC_{50}$ values of AURKA inhibitors conceptualized by Bavetsias, et al. and ABL1 inhibitors conceptualized by Huang, et al.

**Chemprop**: We used bond message passing layers with a hidden size of 800, depth of 3, and dropout rate of 0.50. We used a mean aggregation function to join the atomic embeddings. We used a feed-forward network with 4 layers, a hidden size of 200, and a dropout rate of 0.50. We enabled batch normalization. We set the optimization initial learning rate to $1E-5$, maximum learning rate to $1E-4$, and final learning rate to $1E-5$. The training batch size was 32, and the validation batch size was 16. Models were trained for up to 200 epochs, with early stopping after 10 validation epochs.

**Causal-Chemprop**: We ran DAGMA with linear structural equations to learn the weighted adjacency matrix for the SCM. We used 5 DAGMA iterations (T), an L1 regularization of 0.002, a weight threshold of 0.3, a learning rate of 0.005, 700 warm-up iterations, and 7000 maximum iterations. We then created a causal graph from the weighted adjacency matrix using NetworkX [6], instantiated an invertible causal model using the causal graph, then fit its causal mechanisms using either Scikit-Learn's histogram-based gradient boosting regressor trees or linear regressors.

## 4.2 Multi-component tasks

**Datasets**: We trained a multi-component Chemprop model on a combination of aqueous solutions from BigSolDB by Krasnov, et al. and SolProp by Vermeire, et al. The dataset was randomly split into training ($85\%$) and validation ($15\%$) sets. We do not provide the validation set to Causal-Chemprop when learning the SCM. We test intervention-based inference with Causal-Chemprop and compare it to Chemprop by predicting the $logS$ values on the validation set. We then test counterfactual-based inference by extrapolating the $logS$ values of MIF inhibitors conceptualized by Cisneros, et al.

**Chemprop**: We trained a multi-component Chemprop model using the same parameters as the single-component Chemprop model for the $IC_{50}$ task, except using a larger batch size of 256 for training and 64 for validation.

**Causal-Chemprop**: We learned a weighted adjacency matrix using the same DAGMA parameters as the $IC_{50}$ task. When constructing the causal graph, we forced temperature to be a root node, then instantiated an invertible causal model and fit its causal mechanisms like before.

### 4.3 Molecular optimization

EvoMol is an evolutionary algorithm that sequentially builds molecular graphs independent of starting data. EvoMol uses a set of 7 generic mutations close to the atomic level in order to search a large part of the chemical space [12]. EvoMol aims to optimize an objective function, which in our case is to maximize the molecular property predicted by Chemprop or Causal-Chemprop. We run EvoMol with only one objective: to maximize solubility. We added a constraint that disables the algorithm from breaking and creating bonds. We penalize the algorithm if it generates a molecule without the MIF seed structure. We run the EvoMol algorithm for 50 steps.

## 5 Conclusions

We introduced Causal-Chemprop, a causal machine learning framework for molecular property prediction and optimization. Causal-Chemprop addresses key limitations of GNN models like Chemprop, particularly with generalizing to OOD molecules when trained on small datasets.

Through *intervention-based inference*, Causal-Chemprop demonstrated strong predictive performance on $IC_{50}$ values from the Kinase Knowledgebase. By intervening directly on the parents of $IC_{50}$, Causal-Chemprop was able to correctly rank-order the AURKA and ABL1 inhibitors. Causal-Chemprop also demonstrated strong predictive performance on $logS$ values from the aqueous solubility dataset, successfully capturing temperature gradients as per domain knowledge.

Through *counterfactual-based inference*, Causal-Chemprop showed strong extrapolative performance in predicting solubility across derivatives of a quinolinyltriazole MIF inhibitor seed structure. By observing the seed structure and intervening on intrinsic causal contributors, Causal-Chemprop accurately captured the contribution of decorators on the quinolinyltriazol scaffold. Integration with the molecular optimization algorithm EvoMol as a scoring function unlocked robust inverse molecular design, yielding realistic and highly soluble analogs of the quinolinyltriazol seed structure.

### 5.1 Future work

Since we do not have access to the ground-truth causal model, we cannot guarantee that the SCM and its causal mechanisms encode real causal relations between variables. Future work is to develop methods to evaluate the causal sufficiency of our SCM for any hidden confounders, then back-propagate these evaluation metrics to Chemprop to learn a more causal molecular representation.

It is also worth experimenting with other causal mechanisms to model non-root nodes, like Post-Nonlinear Causal Models [18] and Causal Location-Scale Noise Models [10]. These could be better methods of incorporating noise into the causal mechanisms, allowing better causal identifiability than the additive noise model. Finally, we could introduce molecular descriptors and fingerprints to Causal-Chemprop as confounding variables in the SCM, which could help capture noise and further improve generalization ability to OOD molecules.

## References


[1] L. Attia, J. W. Burns, P. S. Doyle, and W. H. Green. Organic solubility prediction at the limit of aleatoric uncertainty. ChemRxiv, 2024.

[2] V. Bavetsias, S. Linardopoulos, R. Bayliss, P. Workman, J. Blagg, et al. Aurora isoform selectivity: Design and synthesis of imidazo[4,5-b]pyridine derivatives as highly selective inhibitors of aurora-a kinase in cells. *Journal of Medicinal Chemistry*, 56(22):9122–9135, 2013.

[3] K. Bello, B. Aragam, and P. Ravikumar. Causal representation learning for out-of-distribution generalization. In *Advances in Neural Information Processing Systems*, volume 35, pages 8226–8239, 2022.

[4] J. A. Cisneros, M. J. Robertson, B. Q. Mercado, and W. L. Jorgensen. Systematic study of effects of structural modifications on the aqueous solubility of drug-like molecules. *ACS Medicinal Chemistry Letters*, 8(1):124–127, 2017.

[5] C. Fang, Y. Wang, R. Grater, S. Kapadnis, C. Black, P. Trapa, and S. Sciabola. Prospective validation of machine learning algorithms for absorption, distribution, metabolism, and excretion prediction: An industrial perspective. *Journal of Chemical Information and Modeling*, 63(11):3263–3274, 2023.

[6] A. Hagberg, P. Swart, and D. Chult. Exploring network structure, dynamics, and function using networkx. In *Proceedings of the 7th Python in Science Conference*, 2008.

[7] E. Heid, K. P. Greenman, Y. Chung, S.-C. Li, D. E. Graff, F. H. Vermeire, H. Wu, W. H. Green, and C. J. McGill. Chemprop: A machine learning package for chemical property prediction. *Journal of Chemical Information and Modeling*, 64(1):9–17, 2024.

[8] P. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21 (NeurIPS 2008)*, Vancouver, Canada, 2009. Curran Associates, Inc.

[9] Ting-Ting Huang, Xin Wang, Shao-Jia Qiang, Zhen-Nan Zhao, Zhuo-Xun Wu, Charles R. Jr. Ashby, Jia-Zhong Li, and Zhe-Sheng Chen. The discovery of novel bcr-abl tyrosine kinase inhibitors using a pharmacophore modeling and virtual screening approach. *Frontiers in Cell and Developmental Biology*, 9:649434, 2021.

[10] A. Immer, C. Schultheiss, J. E. Vogt, B. Schölkopf, P. Bühlmann, and A. Marx. On the identifiability and estimation of causal location-scale noise models. arXiv preprint, 2022.

[11] L. Krasnov, S. Mikhaylov, M. Fedorov, and S. Sosnin. Bigsoldb: Solubility dataset of compounds in organic solvents and water in a wide range of temperatures. ChemRxiv, 2023.

[12] J. Leguy, T. Cauchy, M. Glavatskikh, B. Duval, and B. Da Mota. Evomol: a flexible and interpretable evolutionary algorithm for unbiased de novo molecular generation. *Journal of Cheminformatics*, 12:1–19, 2020.

[13] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2nd edition, 2009.

[14] Judea Pearl. A structural theory of causation. In M. Knauff and W. Spohn, editors, *The Handbook of Rationality*, pages 427–438. The MIT Press, Cambridge, MA, 2021.

[15] R. Sharma, S. C. Schürer, and S. M. Muskal. High-quality, small molecule–activity datasets for kinase research. *F1000Research*, 5(Chem Inf Sci):1366, 2016.

[16] F. Vermeire, Y. Chung, and W. Green. Predicting solubility limits of organic solutes for a wide range of solvents and temperatures. ChemRxiv, 2022.

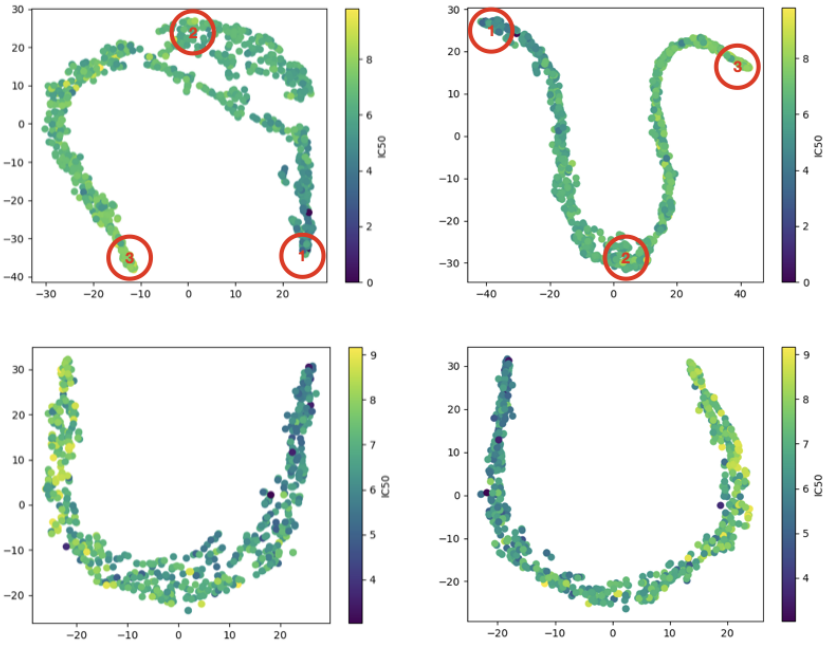
[17] X. Wen, Y. Guo, S. Wei, W. Long, L. Zhu, and R. Zhu. Causal invariant hierarchical molecular representation for out-of-distribution molecular property prediction. 2024.

[18] K. Zhang and A. Hyvärinen. On the identifiability of the post-nonlinear causal model. arXiv preprint, 2012.

# A Technical Appendices and Supplementary Material

## A.1 Representation space t-SNE plots

### A.1.1 AURKA and ABL1



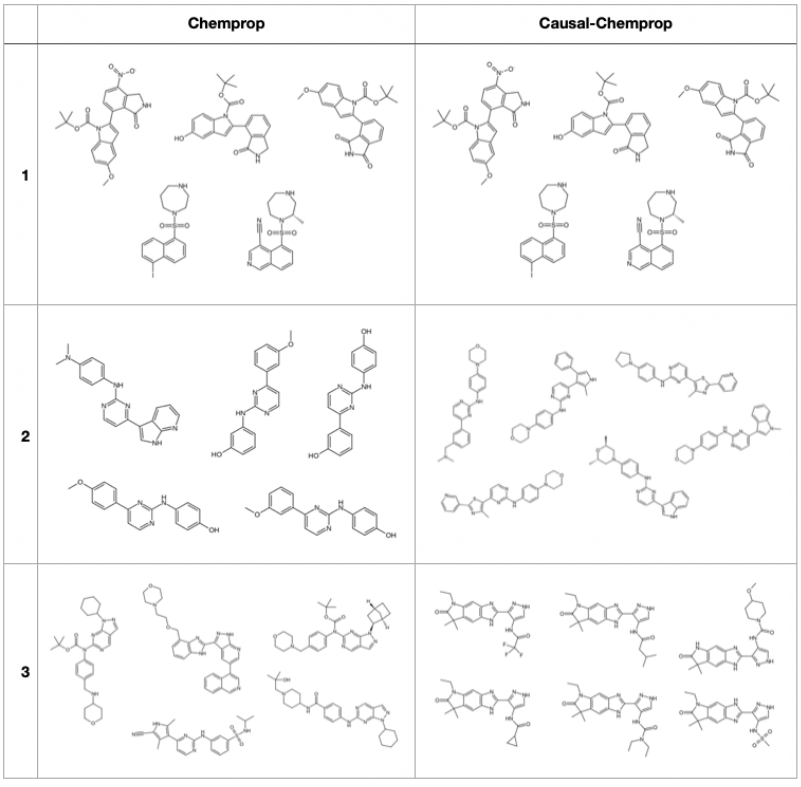




Figure 8: t-SNE visualization of the AURKA dataset (top) and ABL1 dataset (bottom) using the Chemprop embedding *(left)* and the Causal-Chemprop features *(right)*

### A.1.2 logS

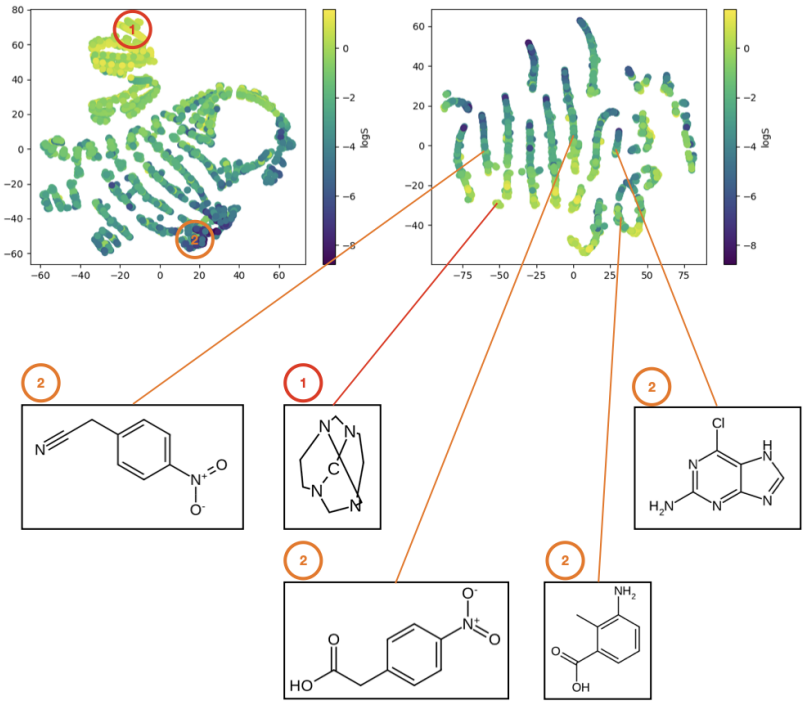



Figure 9: t-SNE visualization of the combined aqueous dataset using the full Chemprop embedding *(left)* and the Causal-Chemprop features *(right)*

### A.2 Molecular structures for counterfactual inference

#### A.2.1 Quinolinyltriazole MIF inhibitor

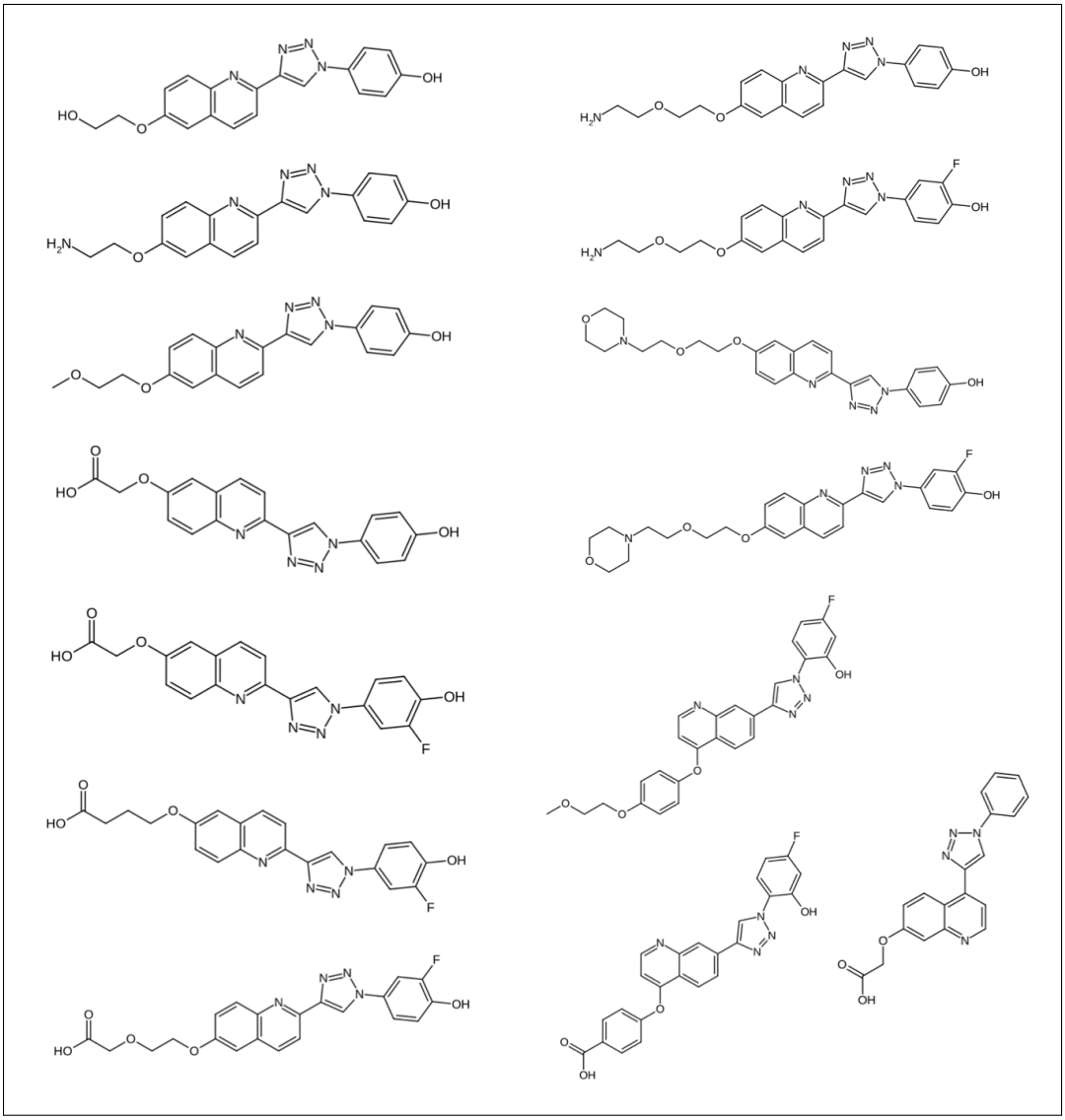


Figure 10: Molecular derivatives of the quinolinyltriazole MIF inhibitor seed structure

### A.2.2 AURKA kinase inhibitor

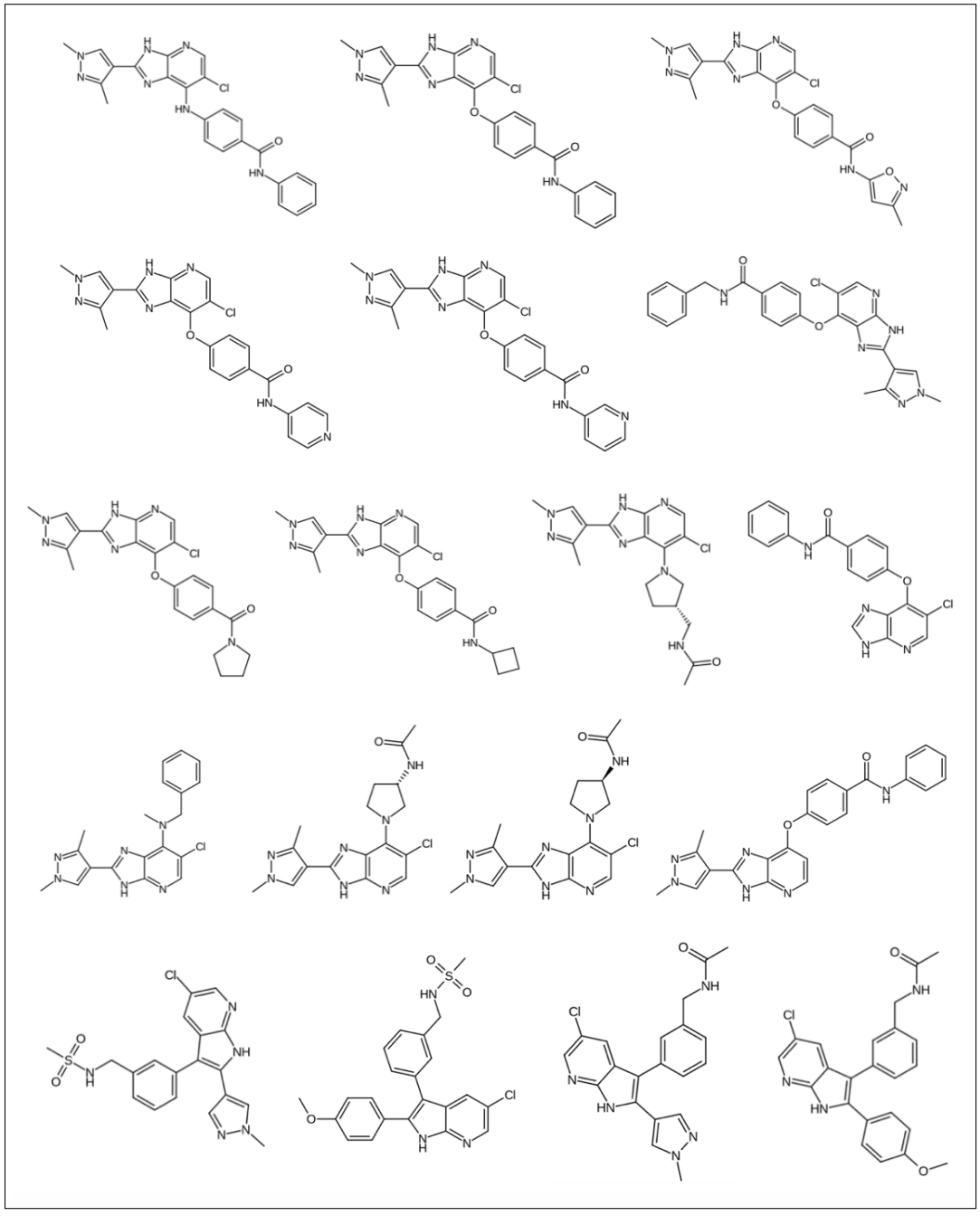


Figure 11: Molecular structures of AURKA inhibitors used for counterfactual-based inference

### A.2.3 ABL1 kinase inhibitor

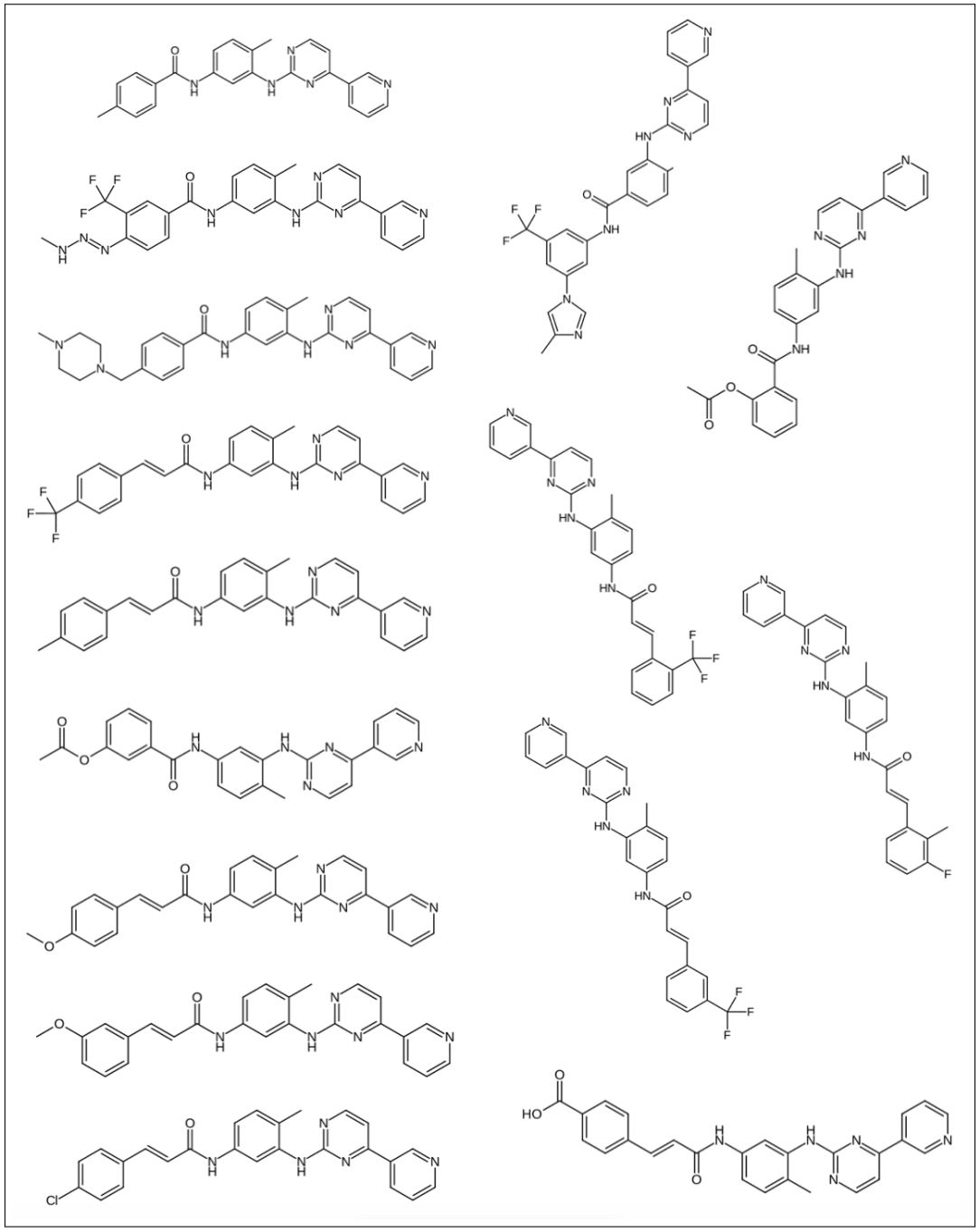


Figure 12: Molecular structures of ABL1 inhibitors used for counterfactual-based inference

### A.3 Compute resources

Training and inferencing were all conducted on an Apple M2 chip and 16GB of memory, signifying the computational viability of both Chemprop and Causal-Chemprop models. Training Chemprop took approximately 2 minutes for both $IC_{50}$ task and $logS$ tasks. The time complexity of DAGMA is dominated by the computing the log determinant acyclicity constraint, which takes $O(n^3)$ time and scales with the number of nodes in the causal graph. This then scales linearly with the number of DAGMA iterations. Learning the causal graph with DAGMA over a 200-dimension Chemprop representation with 7000 total iterations took 30 seconds for both $IC_{50}$ and $logS$ tasks, and fitting the causal mechanisms to each node takes 1 minute for both tasks. Performing intervention-based inferences with Causal-Chemprop is trivial to compute, though counterfactual-based inferences take approximately 1 second to compute per sample.