JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

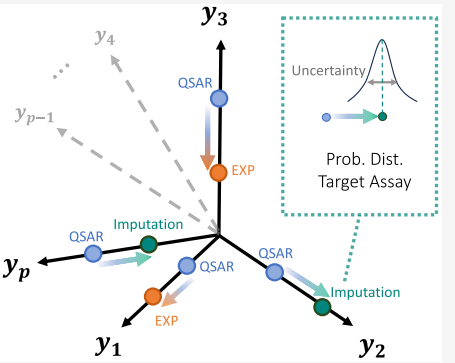
QComp: A QSAR-Based Imputation Framework for Drug Discovery

Bingjia Yang,* Yunsie Chung, Archer Y. Yang, Bo Yuan, Tianchi Chen, and Xiang Yu*

Cite This: https://doi.org/10.1021/acs.jcim.5c00059 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: In drug discovery, in vitro and in vivo experiments generate biochemical activity data that are crucial for evaluating the efficacy and toxicity of compounds. These data sets are massive, sparse, and ever-evolving. Quantitative structure−activity relationship (QSAR) models, which predict biochemical activities from compound structures, face challenges in integrating the evolving experimental data agilely as studies progress. We developed QSAR-Complete (QComp), an imputation framework, to address these challenges. While QSAR models are updated at a slow pace through extensive retraining on enlarging data sets, QComp leverages existing QSAR models to immediately exploit new experimental data and improves the imputation of missing data. We demonstrate that the improvement is robust and substantial for imputing in vivo assays with only in vitro experimental data. Additionally, QComp assists in finding the optimal sequence of experiments by quantifying the reduction in statistical uncertainty for specific end points, aiding in rational decision-making throughout the drug discovery process.



1. INTRODUCTION

Quantitative structure−activity relationship (QSAR) modeling is one of the most important approaches for data-driven prediction of molecular properties,[1−4] with recent progress led by deep learning.[5−10] Sophisticated deep learning methods can model various chemical properties with a unified (multitask) neural network model.[5,11−14] QSAR finds major applications in material and drug discovery,[10,15] where QSAR models are trained on existing data sets of molecules with known properties. The models are then utilized for high-throughput screening[14,16] of a massive database of molecules. For virtual screening, it is advantageous that QSAR takes only the structure of a molecule to make the prediction. This simplicity becomes less desirable in stages past virtual modeling, where experimental data on a few chemical properties become available for some compounds. QSAR models cannot dynamically incorporate these newly acquired data toward improved prediction.[17] Extensive retraining of the models with both the original training set and the newly acquired data has to be carried out. Such retraining is not economical for large deep learning QSAR models when the number of newly acquired data is negligible compared to the size of the original training set, a common scenario in industrial practice of material and drug discovery due to the cost of experiments and the massive size of historical data. Therefore, it is desirable to have an imputation model that can leverage pre-existing QSAR models and dynamically incorporate any amount of newly acquired data without retraining imputation/QSAR models on these new data.

For this purpose, we develop a QSAR-based imputation framework, named "QSAR-Complete" or "QComp" for brevity. QComp treats biochemical activities $\mathbf{y}$ of a molecule as a probability distribution $\mathcal{P}(\mathbf{y}|\mathbf{x})$ decided by the chemical structure $\mathbf{x}$ of the molecule. Typical structure-based QSAR models can be understood as to directly predict $\operatorname{argmax}_{\mathbf{y}}\mathcal{P}(\mathbf{y}|\mathbf{x})$ as a function of $\mathbf{x}$. QComp addresses instead the case in which some entries of $\mathbf{y}$ are determined by newly acquired experimental data. To do so, QComp parameterizes the probability distribution of the missing entries of $\mathbf{y}$ as a function of known entries and $\mathbf{x}$. The maximum likelihood of such a function yields optimal imputation. Moreover, QComp incorporates a pre-existing QSAR model in a natural way such that QComp can reproduce the structure-based QSAR prediction when $\mathbf{y}$ is entirely unknown. In other words, the maximum of the distribution $\mathcal{P}(\mathbf{y}|\mathbf{x})$ as a function of free variable $\mathbf{y}$ is constructed as the prediction of a base QSAR model. With a partially missing data set, our approach models the conditional distribution of $\mathcal{P}(\mathbf{y}|\mathbf{x})$ where some entries of $\mathbf{y}$ have been fixed by experimental data.

Because QComp is based on leveraging an existing QSAR model, it is distinguished from general imputation algorithms[18−24] that are built from scratch. Multivariate imputation by chained equations (MICE)[22] and MissForest[23] are leading members in the category of general iterative imputers. They

Received: January 14, 2025
Revised: July 14, 2025
Accepted: July 15, 2025

model each feature as a function of others, starting by replacing missing values with statistical means or the most frequent values. Then, the imputed entries are updated iteratively in a round-robin fashion. Another major category of imputers is matrix factorization-based methods.[25] Macau,[24] a member of this category, has been applied to drug discovery tasks.[17] Although these general algorithms do not base imputation on another predictive model, they are flexible enough to incorporate additional information for improved performance on sparse data sets, which allows fair comparison with QComp.

In addition to general methods, specific imputation methods have been tailored for predicting chemical properties, such as Alchemite[26] and pQSAR.[27,28] Alchemite, as an iterative imputer, updates imputed values through a multitask neural network with chemical structures and activities as input. Here, directly utilizing a neural network for imputation raises concerns about convergence,[29,30] a typical issue for iterative imputers. The risk of divergence is certain for a deep neural network that often experiences overfitting and unreliable extrapolation on insufficiently large data sets, a common scenario for in vivo properties. pQSAR as appearing in ref 27,28 addresses a scenario different from the one targeted by QComp. QComp mainly focuses on cases where there are a fixed number of assays (or columns) in a long-standing data set. One wants to better impute the sparse measurement of new compounds (new rows) added to the data set by leveraging well-developed existing QSAR models. pQSAR focuses on an orthogonal scenario: new assays (columns) are introduced to a long-standing data set. The new assays accumulate few data points compared to long-standing assays, and there are no trustworthy QSAR models directly trained on these new assays. So, pQSAR establishes an indirect QSAR model for new assays through partial-least-squares optimization of a linear mapping from existing QSAR models of old assays to the prediction of new assays. Apparently, pQSAR and QComp cannot be directly applied to each other's target scenario, as there are missing base QSAR models in the former, while in the latter, there have been nonlinear, multitask base QSAR models[8,31] that should work better than a linear combination of single-task models.

In our target scenario, considering challenges in iterative imputation, the QComp approach instead builds the imputation on a probabilistic framework with a well-defined optimum.

In this work, we mainly apply QComp to model the absorption, distribution, metabolism, elimination, and toxicity (ADMET) for small molecules. These properties are tightly bound to the efficacy and safety of the drug candidates. We will demonstrate that QComp systematically improves upon structure-based QSAR for ADMET imputation. In particular, we show that QComp is promising in improving the prediction of in vivo assays with the knowledge of in vitro data, a feature that is highly desirable in high-throughput screening. Meanwhile, QComp shows advantages in accuracy, robustness, and interpretability compared to several standard imputation methods fed with the same side information from the pre-existing QSAR model. And we show that QComp leads to a simple strategy for the optimization of decision-making in drug discovery. In addition, because QComp itself is a general algorithm, it is not limited to ADMET tasks. We demonstrate the advantages of QComp also on the QM8 data set[32] of electronic structure, reported in the Supporting Information.

The rest of the article is organized as follows. In Section 2, the QComp approach is formulated. In Section 3, we first introduce the data sets and the adopted base QSAR model. Then, we benchmark QComp against general imputers MICE and MissForest under multiple scenarios on a large ADMET data set with around 7,80,000 molecules. We also benchmark QComp on a public ADMET data set for reproducibility. Then, we show how QComp can assist decision-making in planning the sequence of experiments. In the end, we summarize our findings and conclude in Section 4.

2. METHODS

2.1. Probabilistic Framework of QComp. We begin by formally defining the imputation problem using probabilistic terminology. For a molecule uniquely labeled as i in a molecular database $\mathcal{I}$, let $\mathbf{x}^{(i)}$ be its chemical structure and the row vector $\mathbf{y}^{(i)} = (y_1^{(i)}, y_2^{(i)}, \cdots, y_p^{(i)})$ represents its p target activities. The QSAR is described by the probability distribution $\mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$. In the database, some entries of $\mathbf{y}^{(i)}$ were determined from experiments. We use $\mathbf{y}^{\mathrm{O}(i)}$ to denote the subvector (of length $p_{\mathrm{O}}^{(i)}$) of $\mathbf{y}^{(i)}$ containing those known (observed) activities from experiments, and $\mathbf{y}^{\mathrm{M}(i)}$ the subvector (of length $p_{\mathrm{M}}^{(i)} = p - p_{\mathrm{O}}^{(i)}$) containing unknown (missing) activities as random variables. So for arbitrary molecule i in the database, we have the partition $\mathbf{y}^{(i)} = (\mathbf{y}^{\mathrm{M}(i)}, \mathbf{y}^{\mathrm{O}(i)})$.

The task of QComp is to determine $\mathcal{P}(\mathbf{y}^{\mathrm{M}(i)}|\mathbf{y}^{\mathrm{O}(i)}, \mathbf{x}^{(i)})$ as a conditional distribution of $\mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$ for each molecule i. The optimal imputation is the conditional expectation $\tilde{\mathbf{y}}^{\mathrm{M}(i)} = \mathbb{E}(\mathbf{y}^{\mathrm{M}(i)}|\mathbf{y}^{\mathrm{O}(i)}, \mathbf{x}^{(i)})$. Note that, when there is no known data, the imputation task falls back to the vanilla QSAR problem, i.e., determining $\mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$ entirely from chemical structures. The situation in a realistic pharmaceutical setting is the following. With all known data $\{\mathbf{y}^{\mathrm{O}(i)}|i \in \mathcal{I}\}$, one has already trained a set of deterministic QSAR models, giving access to an estimation of $\mathrm{argmax}_{\mathbf{y}}\mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$ as a function of $\mathbf{x}^{(i)}$, denoted by $\boldsymbol{f}^{(i)} = (f_1(\mathbf{x}^{(i)}), f_2(\mathbf{x}^{(i)}), \cdots, f_p(\mathbf{x}^{(i)}))$. QComp utilizes this estimation and assumes that $\mathbf{y}^{(i)}$ conditional on $\mathbf{x}^{(i)}$ follows a multivariate Gaussian distribution $\mathbf{y}^{(i)}|\mathbf{x}^{(i)} \sim N(\boldsymbol{\mu}^{(i)}, \boldsymbol{\Sigma})$ with the probability density function $\mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})$ given by

$$\mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}) = ((2\pi)^p|\boldsymbol{\Sigma}|)^{-1/2} \exp\left(-\frac{1}{2}(\mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)})\boldsymbol{\Sigma}^{-1}(\mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)})^{\mathrm{T}}\right) \quad (1)$$

This is not to be confused with assuming the activity $\mathbf{y}^{(i)}$ itself is normally distributed, which is a much stronger assumption (see Section 3.2 for details). The row vector $\boldsymbol{\mu}^{(i)} = \boldsymbol{f}^{(i)}\mathbf{B} + \mathbf{b}$ is a linear transformation of the QSAR prediction $\boldsymbol{f}^{(i)}$, serving as a multitask calibration of given QSAR models. $\mathbf{B}$ is a $p \times p$ matrix, and $\mathbf{b}$ a $1 \times p$ vector. The covariance matrix $\boldsymbol{\Sigma}$ is a positive-definite $p \times p$ matrix. $|\boldsymbol{\Sigma}|$ denotes its determinant. Specifically, $\boldsymbol{\Sigma}$ is represented by its Cholesky decomposition, and only the resulting lower triangle matrix is treated as free parameters. In the following, we use θ to represent the group of parameters determining $\mathbf{B}$, $\mathbf{b}$, and $\boldsymbol{\Sigma}$.

For each i and the partition $\mathbf{y}^{(i)} = (\mathbf{y}^{\mathrm{M}(i)}, \mathbf{y}^{\mathrm{O}(i)})$, the calibrated QSAR prediction $\boldsymbol{\mu}^{(i)}$ can be correspondingly partitioned as $(\boldsymbol{\mu}^{\mathrm{M}(i)}, \boldsymbol{\mu}^{\mathrm{O}(i)})$. Note that the indices of the missing and observed elements of the vector $\mathbf{y}^{(i)}$ can be different for different

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

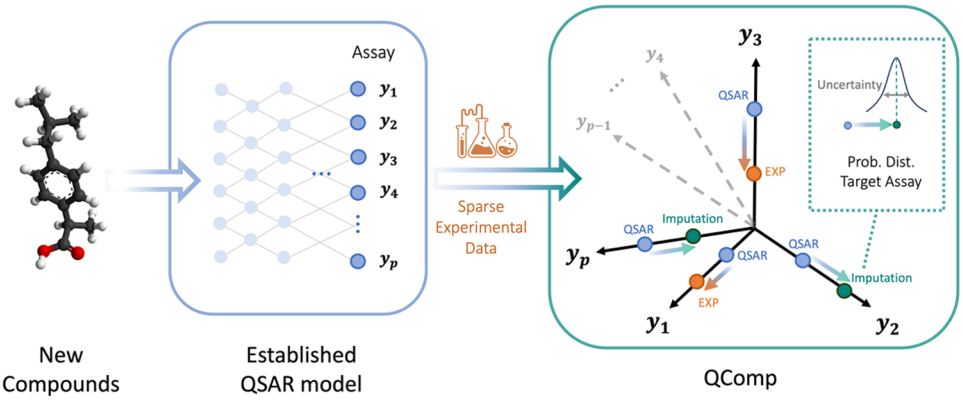


Figure 1. Imputation procedure of QComp. For a set of new compounds with assays $\mathbf{y}_1$, ⋯, and $\mathbf{y}_p$ under consideration, QSAR predictions are available for all assays, while experimental data is partially available. QComp utilizes both QSAR and sparse experimental data to predict the probability distribution of the missing assays and the corresponding optimal imputation values.

observation i. And the covariance matrix $\boldsymbol{\Sigma}$ can be partitioned as the block matrix

$$\begin{bmatrix} \boldsymbol{\Sigma}^{\mathrm{MM}(i)} & \boldsymbol{\Sigma}^{\mathrm{MO}(i)} \\ [\boldsymbol{\Sigma}^{\mathrm{MO}(i)}]^{\mathrm{T}} & \boldsymbol{\Sigma}^{\mathrm{OO}(i)} \end{bmatrix} \tag{2}$$

Here, $\boldsymbol{\Sigma}^{\mathrm{MM}(i)}$ represents the $p_{\mathrm{M}}^{(i)} \times p_{\mathrm{M}}^{(i)}$ submatrix of $\boldsymbol{\Sigma}$ associated with the covariance of $\mathbf{y}^{\mathrm{M}(i)}$. Similarly, $\boldsymbol{\Sigma}^{\mathrm{MO}(i)}$ is the submatrix of $\boldsymbol{\Sigma}$ whose rows correspond to $\mathrm{M}(i)$ and columns correspond to $\mathrm{O}(i)$, i.e., $\boldsymbol{\Sigma}^{\mathrm{MO}(i)} = \boldsymbol{\Sigma}^{(\mathrm{M}(i),\mathrm{O}(i))}$. Finally, $\boldsymbol{\Sigma}^{\mathrm{OO}(i)}$ represents the $p_{\mathrm{O}}^{(i)} \times p_{\mathrm{O}}^{(i)}$ submatrix of $\boldsymbol{\Sigma}$ corresponding to the observed coordinates of $\mathbf{y}^{(i)}$, i.e., $\boldsymbol{\Sigma}^{\mathrm{OO}(i)} = \boldsymbol{\Sigma}^{(\mathrm{O}(i),\mathrm{O}(i))}$.

2.2. Training. Within the QComp model, the likelihood of the observation $\mathbf{y}^{\mathrm{O}(i)}$ follows the marginal Gaussian distribution

$$\mathcal{P}(\mathbf{y}^{\mathrm{O}(i)}|\mathbf{x}^{(i)}) = \int \mathcal{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)})d\mathbf{y}^{\mathrm{M}(i)} = \frac{\exp\left(-\frac{1}{2}(\mathbf{y}^{\mathrm{O}(i)} - \boldsymbol{\mu}^{\mathrm{O}(i)})(\boldsymbol{\Sigma}^{\mathrm{OO}(i)})^{-1}(\mathbf{y}^{\mathrm{O}(i)} - \boldsymbol{\mu}^{\mathrm{O}(i)})^{\mathrm{T}}\right)}{\sqrt{(2\pi)^{p_{\mathrm{O}}^{(i)}}|\boldsymbol{\Sigma}^{\mathrm{OO}(i)}|}} \tag{3}$$

We define the following log-likelihood loss function with respect to $\theta = (\mathbf{B}, \mathbf{b}, \boldsymbol{\Sigma})$

$$l(\theta) = -\log \prod_{i \in \mathcal{I}} \mathcal{P}(\mathbf{y}^{\mathrm{O}(i)}|\mathbf{x}^{(i)}) = -\sum_{i \in \mathcal{I}} \log \mathcal{P}(\mathbf{y}^{\mathrm{O}(i)}|\mathbf{x}^{(i)}) \tag{4}$$

$\hat{\theta} = (\hat{\mathbf{B}}, \hat{\mathbf{b}}, \hat{\boldsymbol{\Sigma}})$ denotes the optimal values of θ, defined as $\hat{\theta} = \arg\min_{\theta} l(\theta)$. This optimization problem can be solved by performing a gradient descent on θ. The complexity of forward propagation is bound by $O(N_{\mathcal{I}}p^3)$. $N_{\mathcal{I}}$ is the number of molecules in the training set. p is the total number of chemical activities/assays. Once $\hat{\theta}$ is obtained, the calibrated QSAR prediction is $\hat{\boldsymbol{\mu}}^{(i)} = \boldsymbol{f}^{(i)}\hat{\mathbf{B}} + \hat{\mathbf{b}}$.

2.3. Imputation. After training QComp on a database, one can get an estimation of $\hat{\theta}$ and use it to do one-shot imputation. Note that $\mathbf{y}^{\mathrm{M}(i)}$ conditioned on $\mathbf{y}^{\mathrm{O}(i)}$ and $\mathbf{x}^{(i)}$ follows a Gaussian distributioni.e.,

$$\mathbf{y}^{\mathrm{M}(i)}|\mathbf{y}^{\mathrm{O}(i)}, \mathbf{x}^{(i)} \sim N(\tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)}, \tilde{\boldsymbol{\Sigma}}^{\mathrm{MM}(i)}) \tag{5}$$

where

$$(\tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)})^{\mathrm{T}} = (\hat{\boldsymbol{\mu}}^{\mathrm{M}(i)})^{\mathrm{T}} + \hat{\boldsymbol{\Sigma}}^{\mathrm{MO}(i)}(\hat{\boldsymbol{\Sigma}}^{\mathrm{OO}(i)})^{-1}(\mathbf{y}^{\mathrm{O}(i)} - \hat{\boldsymbol{\mu}}^{\mathrm{O}(i)})^{\mathrm{T}} \tag{6}$$

and

$$\tilde{\boldsymbol{\Sigma}}^{\mathrm{MM}(i)} = \hat{\boldsymbol{\Sigma}}^{\mathrm{MM}(i)} - \hat{\boldsymbol{\Sigma}}^{\mathrm{MO}(i)}(\hat{\boldsymbol{\Sigma}}^{\mathrm{OO}(i)})^{-1}[\hat{\boldsymbol{\Sigma}}^{\mathrm{MO}(i)}]^{\mathrm{T}} \tag{7}$$

The corresponding probability density function is

$$\mathcal{P}(\mathbf{y}^{\mathrm{M}(i)}|\mathbf{y}^{\mathrm{O}(i)}, \mathbf{x}^{(i)}) = \frac{\exp\left(-\frac{1}{2}(\mathbf{y}^{\mathrm{M}(i)} - \tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)})(\tilde{\boldsymbol{\Sigma}}^{\mathrm{MM}(i)})^{-1}(\mathbf{y}^{\mathrm{M}(i)} - \tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)})^{\mathrm{T}}\right)}{\sqrt{(2\pi)^{p_{\mathrm{M}}^{(i)}}|\tilde{\boldsymbol{\Sigma}}^{\mathrm{MM}(i)}|}} \tag{8}$$

The optimal imputation given by QComp for the missing assays is therefore

$$\mathbb{E}(\mathbf{y}^{\mathrm{M}(i)}|\mathbf{y}^{\mathrm{O}(i)}, \mathbf{x}^{(i)}) = \tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)} \tag{9}$$

The complexity for imputing a data set of $N_{\mathcal{I}}$ rows is therefore bound by $O(N_{\mathcal{I}}p^3)$. The imputation procedure outlined above is illustrated in Figure 1.

A comment on the imputation uncertainty is in order. Here, the uncertainty related to $\tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)}$ is not simply the diagonal of $\tilde{\boldsymbol{\Sigma}}^{\mathrm{MM}(i)}$, unless one can ignore the uncertainty embedded in the QSAR prediction, which is usually far from negligible. We construct a composite uncertainty in Section B of the Supporting Information to address this extra complication. However, even without further construction, here we are already able to have a clear idea of how much certainty one can gain on missing assays $\mathbf{y}^{\mathrm{M}(i)}$ by knowing the experimental measurements $\mathbf{y}^{\mathrm{O}(i)}$. The gain of certainty (GOC) is simply the diagonal terms in $\hat{\boldsymbol{\Sigma}}^{\mathrm{MO}(i)}(\hat{\boldsymbol{\Sigma}}^{\mathrm{OO}(i)})^{-1}[\hat{\boldsymbol{\Sigma}}^{\mathrm{MO}(i)}]^{\mathrm{T}}$.

3. EXPERIMENTS

3.1. Data and Model Details. *3.1.1. Data Sets.* We apply our approach to one proprietary ADMET data set, one public ADMET data set compiled from various public sources,[13,33−42] and the QM8 data set.[32] The proprietary data set (ADMET-780k data set) contains sparse data of 31 in vitro and in vivo ADMET assays for around 7,80,000 molecules, recorded internally at Merck & Co., Inc., Rahway, NJ. An earlier version of the proprietary data set has been used in ref 43 for QSAR modeling. The public ADMET data set contains data from 25

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

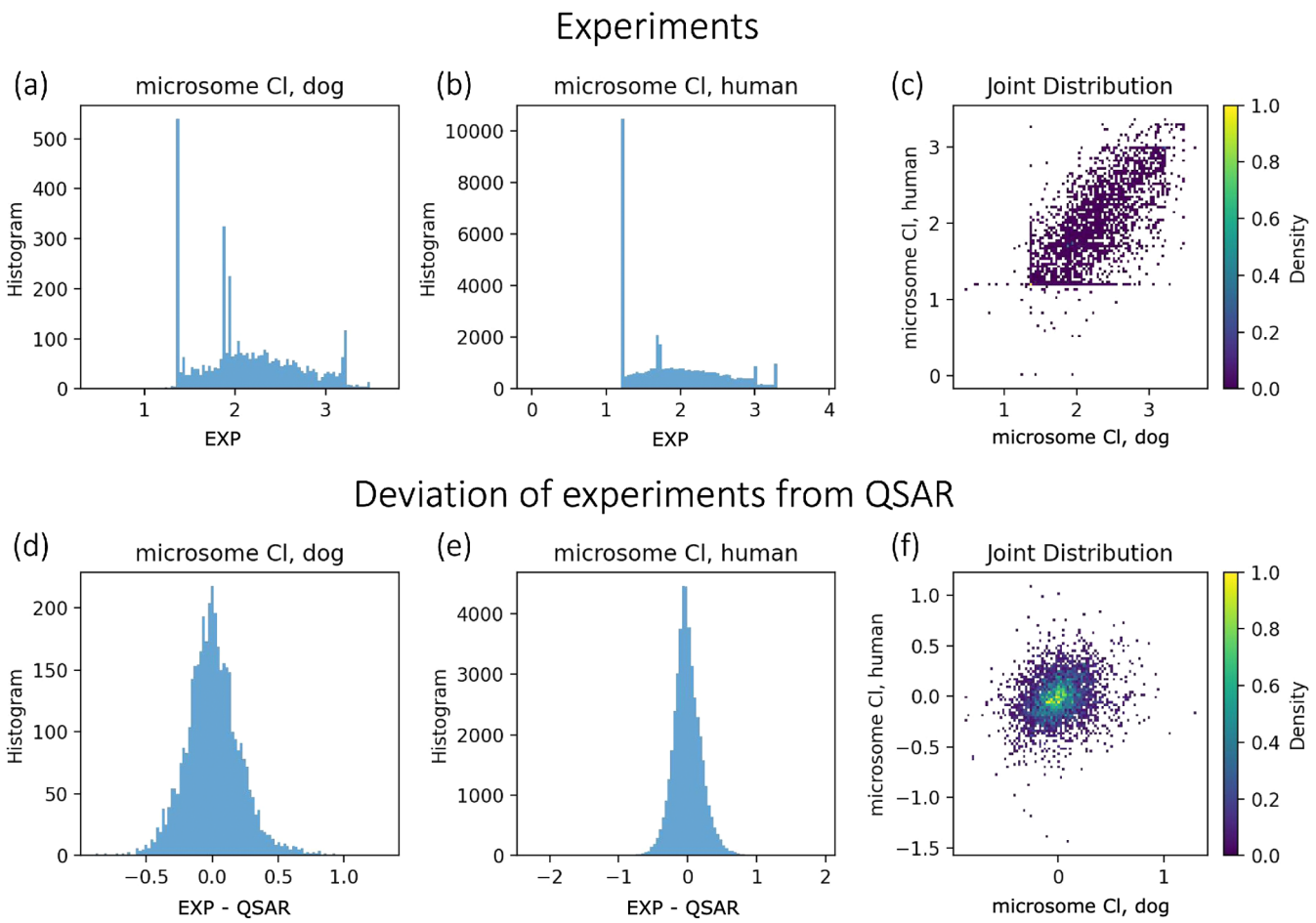


Figure 2. (a, b) Histograms of the "microsome Cl" assays for dogs and humans. (c) Heatmap of the joint distribution of "microsome Cl, dog" and "microsome Cl, human". (d, e) Histograms of the deviation of "microsome Cl" assays from the QSAR predictions. (f) The heatmap of the joint distribution associated with the quantities in panels (d) and (e).

ADMET assays for nearly 1,10,000 molecules. The details of ADMET data sets, including the list of biochemical activities and the Pearson correlation between activities, can be found in Section C of the Supporting Information. The diversity of the ADMET-780k data set and the public ADMET data set is compared against a small database of FDA-approved drugs (3480 molecules).[44] The metric of diversity is the average 5-nearest-neighbor Tanimoto distance on Morgan fingerprints. Randomly sampled subsets of size 3480 are obtained, respectively, from the ADMET-780k data set and the public data set, yielding scores of 0.48 and 0.50. The results imply that both our data sets cover larger chemical space than the set of FDA-approved drugs with a score of 0.36. The QM8 data set consists of 16 quantum mechanical properties of 21,787 molecules.[32] This data set serves to demonstrate the applicability of QComp beyond drug discovery.

We will benchmark QComp on the largest ADMET-780k data set, which is accumulated from consistent industrial drug discovery practices. A similar benchmarking procedure is performed for the public ADMET data set for the reproducibility of the QComp approach. Finally, benchmarking results on the QM8 data set are reported in Section G of the Supporting Information.

3.1.2. Data Splitting Strategies. For the ADMET-780k data set, we split the entire data set into a 90% training/validation subset and 10% test subset using a compound-based temporal split. For the public ADMET data set, we perform 5-fold 80 and 20% random splitting as the time stamp information is not available. The clustering-based splitting strategy associated with the supplemental benchmark not reported in the main text is detailed in the Supporting Information along with the benchmark results.

3.1.3. Base QSAR Models. For ADMET-780k and the public ADMET data set, we train multitask Chemprop models as the base QSAR (see Section A of the Supporting Information for details). The Chemprop model utilizes a directed message-passing neural network (D-MPNN) to predict molecular properties based on the graph representation of molecules.[8,31] Training an ensemble of four Chemprop models on the ADMET-780k data set requires 78 h on an Nvidia V100 GPU. For the QM8 data set, we fine-tune the open-sourced pretrained Uni-Mol model[45] as the base QSAR. The details are reported in Section G of the Supporting Information.

3.1.4. Baseline Imputation Models. We compare the QComp approach to two baseline imputation methods: MICE[22] and Missforest.[23] Macau,[24] also a representative imputation model, is not used for benchmarking here. Because Macau relies on sparse matrix factorization, it is much less efficient than MICE, Missforest, or QComp when working with the ADMET-780k training set (~7,80,000 rows) due to the high cost of factorizing a very large matrix. We are not able to complete the imputation within a reasonable time. A

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

workaround is to stack only a small subset of the training set with the test set for imputation. The outcomes show poor performance compared to that of all other methods. So eventually, we do not use Macau for benchmarking. We provide the two baseline methods, Mice and Missforest, with the same QSAR predictions accessed by QComp. Specifically, for MICE and MissForest, we extend the data set by appending QSAR predictions as supplementary columns. For example, the ADMET-780k data set, originally containing 31 assay columns, is extended to 62 columns, where the extra 31 columns are Chemprop predictions with no missing values. The details and the parameters for these methods are provided in Section A of the Supporting Information. These efforts ensure that all methods have access to the same knowledge, facilitating equitable comparisons and assessments of their respective performance. Training of the three imputation models on the ADMET-780k training set can be much less computationally demanding than that of the base QSAR model. To have an order-of-magnitude estimation, training QComp and MICE on the ADMET-780k training set (~7,80,000 rows) takes several minutes on an Apple M3 Pro CPU chip. Training MissForest with the same data set and CPU chip requires several hours. For comparison, training a neural network-based QSAR model on the ADMET-780k training set typically requires 10−100 h of computing time on an Nvidia V100 GPU (or newer models).

3.2. Validation of Assumption. Here, we examine the basic assumption of QComp: the deviation of the experimental value of an assay from the QSAR prediction is distributed normally (see eq 1). Evidently, the assumed distribution is subject to the quality of the QSAR model. For a trivial QSAR model that gives constant predictions independent of chemical structures, the distribution of $(\mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)})$ can be far from being Gaussian. This is exemplified by Figure 2a,b, where we show with histograms the plain distribution of the experimental values of two assays, "microsome Cl dog" and "microsome Cl human", in the ADMET-780k data set. For both assays, the peak of the histogram is located near the lower end of the distribution, in sharp contrast to a typical Gaussian distribution. Furthermore, the joint distribution of the two assays (Figure 2c) is not close to a 2D Gaussian distribution.

The situation is different when the QSAR model is properly trained. We examine the multitask Chemprop model (trained on the same data set) that serves as the base $\boldsymbol{\mu}^{(i)}$. Figure 2d (Figure 2e) shows with a histogram the distribution of the "microsome Cl, dog (human)" component of $(\mathbf{y}^{(i)} - \boldsymbol{\mu}^{(i)})$. The distributions display a close resemblance to the 1D Gaussian distribution centered at zero. Meanwhile, Figure 2f shows that the joint distribution of the two assays is similar to a zero-centered 2D Gaussian distribution with a positive off-diagonal covariance. The nonzero off-diagonal covariance, i.e., the correlation between different assays, is what is to be utilized by QComp to exceed the capability of bare QSAR. Of course, not all pairs of assays display nonzero off-diagonal covariance, since two chemical properties cannot always be statistically correlated.

Besides the two assays used as examples here, other pairs of assays in all our data sets (listed in Section 3.1) also yield satisfactory Gaussianity with a properly trained QSAR model. These observations validate the assumption of QComp in practical applications. Certainly, there are also cases where QSAR models yield low accuracy and non-Gaussian deviation from experimental data. In such a situation, QComp should not be applied. Instead, efforts should be made to improve the performance of the base QSAR model itself. In principle, Gaussianity is reflected by kurtosis (Fisher's definition) κ associated with a distribution of QSAR-experiment deviation. However, κ is very sensitive to outliers. In practice, we find that calculating κ after excluding outliers outside two standard deviations yields reasonable results that agree with the visual comparison of Gaussianity between assays. Empirically, $|\kappa| < 1$ may be considered acceptable Gaussianity. The majority of assays in our data sets fall in this range.

3.3. Benchmarking QComp for ADMET Imputation. *3.3.1. ADMET-780k Data Set.* We benchmark QComp with other two imputation methods on the ADMET-780k data set. The data set is divided into a 90% training subset and a 10% test subset using a compound-based temporal split. Here, we choose temporal split over random split because the latter may lead to overestimation of generalization capabilities.[46] We train the multitask Chemprop model as the base QSAR model on the training set (see Section A in the Supporting Information for details). Then, QComp, MICE, and Missforest models are trained on the same training set with the QSAR predictions from Chemprop as the side information. Then, these imputation methods are evaluated on the test set (around 77,000 molecules) with the following protocol. For any assay-*i*, we mask the column of assay-*i* in the test set as totally missing and impute this column with all other columns. The imputed column is then compared against available experimental data of assay-*i* with multiple metrics, including the squared Pearson correlation coefficient r^2, the coefficient of determination R^2, the mean absolute error (MAE), and the mean-squared error (MSE). In the following, we will mainly discuss results on r^2 and MSE. Results on R^2 and MAE provide similar insights; therefore, they are reported in the Supporting Information instead.

The r^2 score obtained by the three imputation methods on the test set is reported in Table 1. Overall, the base QSAR model achieves a mean r^2 score (averaged over all 32 assays) of 0.441. QComp, MICE, and Missforest achieve a mean r^2 score of 0.596, 0.530, and 0.530, respectively. QComp outperforms other methods with a 35% improvement over the base. MICE and Missforest yield the same 20% improvement.

Then, to examine the r^2 score on the individual assay, we consider a simple criterion: a successful imputation method should not reduce the r^2 score from the base QSAR model by more than 0.01. QComp meets the requirements for all assays. In contrast, all other methods can yield r^2 scores significantly lower than the base. "PAMPA" and "PXR activation" are outstanding examples where MICE reduces the base r^2 score in the order of 0.1. The comparison shows the excellent robustness of QComp. The robustness of QComp can be understood for those relatively isolated assays (such as CYP2C8, CYP2C9, CPY2D6, CYP3A4) that have negligible correlation with all other assays; the correction from QComp (second term on the right-hand side of eq 6) will be suppressed by vanishing terms in the covariance matrix. This keeps the imputed values of isolated assays close to the base QSAR prediction.

Moreover, QComp outperforms other imputation methods for all assays except "microsome Cl, dog", "CYP2C8", "Fu,p, human", "PAMPA", and "SOLY7", where QComp loses by a small margin and serves as the second best. Similarly, systematic advantages of QComp are found for other metrics,

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 1. Pearson r^2 Scores of the Base QSAR Model Chemprop, MICE, Missforest, and QComp on ADMET-780k Data set with Compound-Based Temporal Splitting[a]

assay name	Chemprop	MICE	Missforest	QComp
Papp	0.721	***0.714***	0.713	**0.725**
CaV 1.2	0.352	***0.357***	0.352	**0.372**
NaV 1.5	0.347	***0.358***	0.339	**0.361**
Cl, dog	0.222	***0.397***	0.289	**0.469**
Cl, rat	0.387	***0.965***	0.893	**0.993**
hepatocyte Cl, dog	0.430	***0.532***	0.426	**0.571**
microsome Cl, dog	0.494	**0.621**	0.472	***0.619***
hepatocyte Cl, human	0.413	***0.509***	0.410	**0.534**
microsome Cl, human	0.472	***0.545***	0.514	**0.599**
hepatocyte Cl, rat	0.365	***0.510***	0.357	**0.527**
microsome Cl, rat	0.499	***0.617***	0.559	**0.661**
CYP2C8	0.442	**0.455**	0.432	***0.453***
CYP2C9	0.400	**0.422**	***0.394***	**0.422**
CYP2D6	0.224	***0.242***	0.151	**0.247**
CYP3A4	0.405	***0.415***	0.413	**0.433**
CYP,TDI,3A4,ratio	0.140	***0.151***	0.138	**0.152**
EPSA	0.816	***0.804***	0.792	**0.813**
halflife, dog	0.334	0.530	***0.719***	**0.753**
halflife, rat	0.224	0.422	***0.721***	**0.752**
hERG MK499	0.470	0.373	***0.466***	**0.469**
Fu,p, human	0.596	**0.630**	0.585	***0.616***
LogD	0.837	0.840	***0.842***	**0.847**
PAMPA	0.494	0.004	**0.519**	***0.493***
PXR activation	0.384	0.218	***0.300***	**0.384**
Fu,p, rat	0.637	***0.658***	0.638	**0.687**
Fassif Solub	0.384	0.384	***0.435***	**0.463**
Vd, rat	0.582	***0.958***	0.857	**0.995**
MRT, dog	0.366	0.712	***0.845***	**0.916**
MRT, rat	0.165	***0.990***	0.683	**0.991**
SOLY7	0.585	0.609	**0.691**	***0.671***
PGP, rat	0.494	***0.497***	0.484	**0.501**

[a]For each assay, the highest r^2 score among different imputation methods is marked in **bold**. The second-highest r^2 score in imputation methods is marked in ***bold and italic***.

including R^2, MAE, and MSE (see Section E of the Supporting Information).

So far, the benchmarking results give access only to the performance of QComp in the long term because the metrics are evaluated on the entire test set with around 77,000 entries, a number much greater than the typical number of newly acquired data obtained within a time frame of a few months. In practice, QComp will be mainly applied to the latter scenario for the agile development of new drugs. To evaluate its performance in such a scenario, we further partitioned the test set into 50 bins with time splitting. Each bin contains 1500 molecules, corresponding to experimental data collected within 1−2 months. For each bin, we computed the changes in Pearson r^2 score and MSE brought by the imputation method over the base Chemprop model. We denote the former by Δr^2 and the latter by Δ_{MSE}. The mean and the error bars of these changes are calculated over the 50 bins and reported in Figure 3 for QComp, MICE, and MissForest, respectively.

Here, we find in the cases of imputing small data sets, QComp still robustly outperforms other imputation methods. The average improvement over the base QSAR model is statistically significant for almost half of the assays. And the results are consistent with the results in Table 1.

Furthermore, we apply the same benchmarking protocol to the ADMET-780k data set with clustering-based splitting (five clusters). The details are reported in Section F of the Supporting Information. Again, we find that QComp systematically outperforms other imputation methods, showing the advantage that QComp may generalize to chemical space that is not covered by the training set.

Comparing QComp, MICE, and Missforest, the success of QComp may be due to its constrained way of utilizing the correlation among ADMET properties: the simple Gaussian model adopted by QComp disregards nonlinear correlations and greatly reduces overfitting. This drastic simplification, however, should not impair much the capability of QComp since we are modeling the deviation of the assay from QSAR predictions. The nonlinear correlation between assays has been largely captured by the nonlinear base QSAR model. The importance of the base QSAR model can also be seen from another perspective: the mean r^2 scores obtained by MICE and Missforest will be significantly smaller if we do not provide QSAR predictions as side information.

3.3.2. Public ADMET Data Set. We benchmark QComp also on the public ADMET data set for the reproducibility of this work. We will demonstrate whether the enhancement brought by QComp is robust over an ensemble of QSAR models trained on different splitting of the same data set. We perform a 5-fold random split (80% training sets and 20% test sets) of the public ADMET data set. For each fold, we first train a Chemprop model as the base QSAR and then a QComp model with the same training set. Next, we evaluated the performance of QComp models on their respective test sets with the general protocol introduced previously. In Figure 4, we report the average change of the Pearson r^2 score (Δr^2) and the MSE (Δ_{MSE}) achieved by these five models on their respective test set (see the Supporting Information for more details). The error bars are computed accordingly for the five models. Similar results showing advantages of QComp are also found for other metrics (R^2, MAE, and MSE) and for clustering-based split (see Sections E and F of the Supporting Information).

Again, QComp shows systematic improvement over the base QSAR model for half of the assays. Note that the public ADMET data set is compiled from multiple sources with many nonoverlapping compounds. Therefore, some assays rarely or never gain experimental data simultaneously for the same compound in the public data set. The covariance among these exclusive assays cannot be accurately determined, which lowers the performance of QComp. The degradation of performance is especially severe when biologically closely related assays, such as pharmacokinetic properties associated with the same animal, have few or no overlaps in the public data set. For example, only two compounds in the public data set have "CL microsome, rat" and "CL total, rat" data simultaneously. Therefore, for biochemical properties that are present in both ADMET-780k data set and the public data set, such as "CL microsome" ("microsome Cl" in ADMET-780k) and "Vd, rat", the performance of QComp appears significantly better on the ADMET-780k data set. We believe that the single-source ADMET-780k data set should provide a more real-world setting than the public ADMET data set.

3.4. Imputing In Vivo Assays with In Vitro Data. In the last section, the simple benchmarking protocol adopted disregards the complication that the location of missing entries is not randomly distributed. In practice, some assays are

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

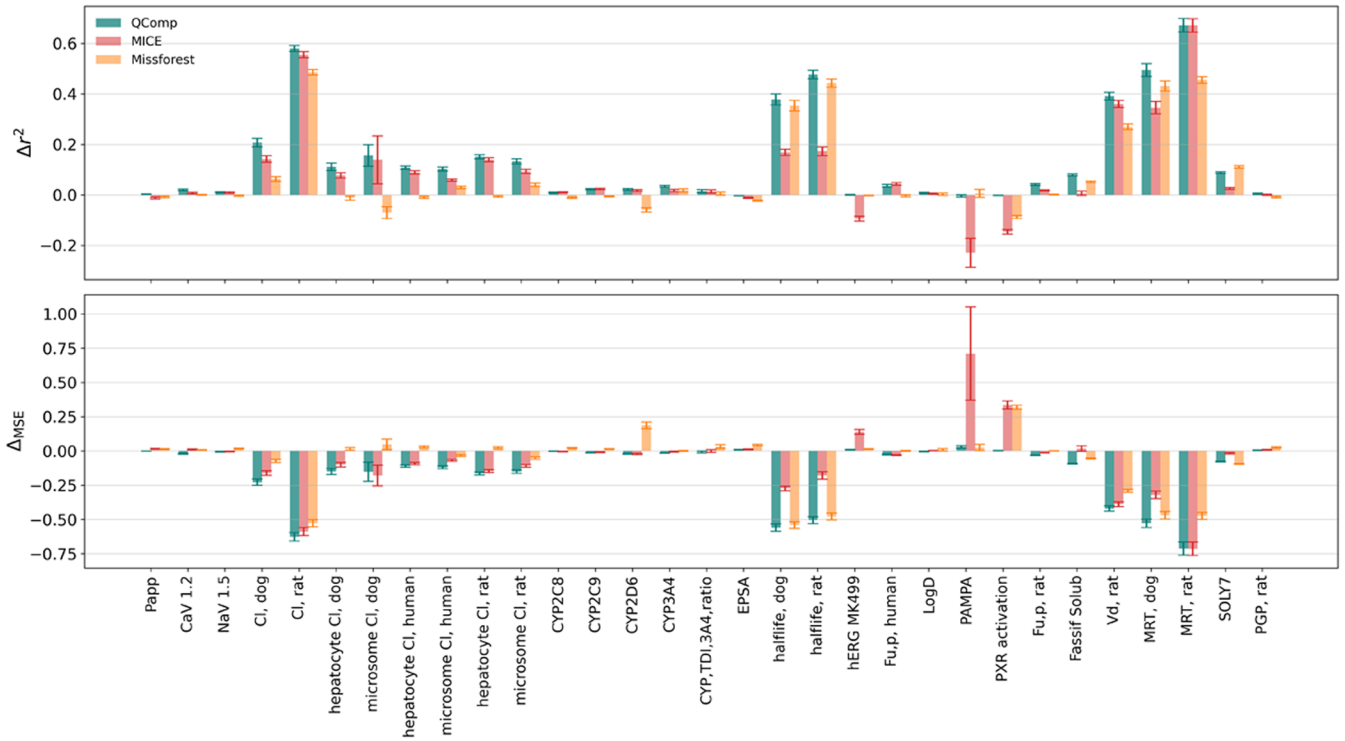


Figure 3. Average and error bar of the change on the Pearson r^2 score (a positive change means improvement) and MSE (a negative change means improvement) over the base QSAR model. The average and the error bar are calculated from the 50-bin splitting of the test set.

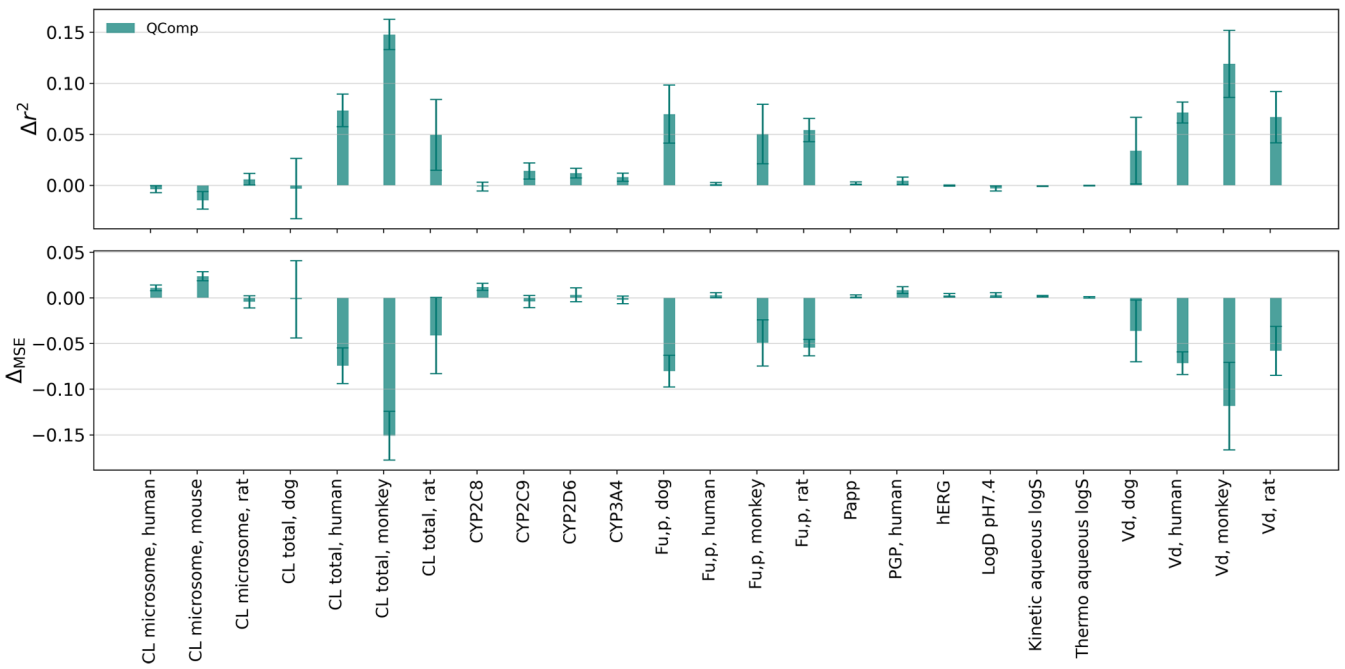


Figure 4. Average and the error bar of the change on the Pearson r^2 score (a positive change means improvement) and MSE (a negative change means improvement) over the base QSAR model. The average and the error bar are calculated over the five QComp models trained with a 5-fold random split.

typically simultaneously present or missing in a data set because they are measured from the same experiment. For example, in the ADMET-780k data set, the in vivo "MRT", "half-life", "Cl", and "Vd" assays associated with the same animal are measured in the same experiment. In vitro assays in this data set have no such issue. Our benchmarking protocol so far does not incorporate such constraints. When we use QComp to impute an in vivo assay like "MRT, rat" of a compound, available experimental data on "half-life, rat", "Cl, rat", and "Vd, rat" of the same compound are utilized. Therefore, our previous benchmarking yields unrealistically large improvements to the base QSAR model for in vivo assays. Although it demonstrated how effective QComp is in utilizing assay−assay correlation, such improvement should not be

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

expected in practice. Next, we address this issue with realistic considerations.

Specifically, in drug discovery, in vitro experiments typically precede in vivo experiments. Carrying out the latter is often expensive and restricted. It is then highly desirable to impute in vivo assays with a knowledge of in vitro assays. In the following, we demonstrate that QComp can achieve such a goal. The tasks are to impute the seven in vivo assays ("MRT, rat", "half-life, rat", "Cl, rat", and "Vd, rat", "MRT, dog", "half-life, dog", "Cl, dog") in the ADMET-780k data set with only available data of in vitro assays. Hence, all of the in vivo data in the test set are masked as missing values. We then imputed these missing entries and compared them against the experimental values. This addresses the underlying issue of previous general benchmarking protocols and puts the imputation tasks in a practical context.

With the setup introduced above, we report in Table 2 the Pearson r^2 scores for in vivo assays, obtained by QComp, Missforest, MICE, and the base QSAR model, on the entire test set of the ADMET-780k data set. For all seven in vivo assays, QComp systematically improves over the base QSAR model and exceeds the performance of all other imputation methods. The improvement is roughly of the order of 10% by percentile for each assay. In contrast to the exaggerated enhancement on predicting in vivo assays reported by Table 1, the results here reflect a reasonable gain from which one can benefit from imputation. Reasonable improvement is also found for R^2, MAE, and MSE, as reported in Section E of the Supporting Information.

Table 2. Pearson r^2 Scores of the Base QSAR Model Chemprop, MICE, Missforest, and QComp on ADMET-780k Data set with Compound-Based Temporal Splitting[a]

assay name	Chemprop	MICE	Missforest	QComp
Cl, dog	0.222	***0.241***	0.219	**0.242**
Cl, rat	0.387	***0.432***	0.385	**0.437**
halflife, dog	0.334	0.304	***0.330***	**0.342**
halflife, rat	0.224	0.193	***0.214***	**0.238**
Vd, rat	0.582	0.538	***0.570***	**0.613**
MRT, dog	0.366	0.284	***0.353***	**0.390**
MRT, rat	0.165	***0.178***	0.156	**0.181**

[a]Only in vitro data are utilized for imputation. For each assay, the highest r^2 score among different imputation methods is marked in **bold**. The second-highest r^2 score in imputation methods is marked in ***bold and italic***.

Next, to confirm the statistical significance of the improvement brought by imputation, we again partitioned equally the test set into 50 bins with time splitting. We study the changes in the Pearson r^2 score and MSE brought by the imputation method over the base QSAR model. The mean and error bars of these changes are calculated over the 50 bins and reported in Figure 5. In contrast to other imputation methods that often cannot keep up with the accuracy of the base QSAR model, the improvement brought by QComp is systematic and substantial for predicting in vivo assays. Such a conclusion applies also to the same data set with clustering-based splitting (see Section F of the Supporting Information). These results suggest that QComp is indeed a useful tool for high-throughput screening, where in vivo experiments are not carried out on a large scale.

3.5. Imputing In Vitro Assays without In Vivo Data. In addition to imputing in vivo assays with in vitro data, we address here another realistic scenario that is imputing missing in vitro data with only other available in vitro data. In fact, this situation holds for a large part of compounds in the ADMET-780k data sets. We mask all of the in vivo data as missing values in the test set. Then, to examine the performance of imputation for any in vitro assay-*i*, we also mask the column of assay-*i* in the test set as totally missing and impute it. The imputed column is then compared against available experimental data of assay-*i* with the squared Pearson correlation coefficient r^2 as a metric.

The benchmarking results on the test set are reported in Table 3. The improvement on the Pearson r^2 score brought by QComp is largely consistent with the improvement calculated from the general protocol used in Section 3.3.1. In Section 3.3.1, while in vivo data are also used to impute in vitro assays, QComp achieves an average r^2 score of 0.526 on in vitro assays, which decreases slightly to 0.518 when in vivo data are not utilized. For comparison, the average r^2 score obtained by the base QSAR model is 0.475 for in vitro assays, signifying a roughly 10% improvement achieved by QComp in both cases.

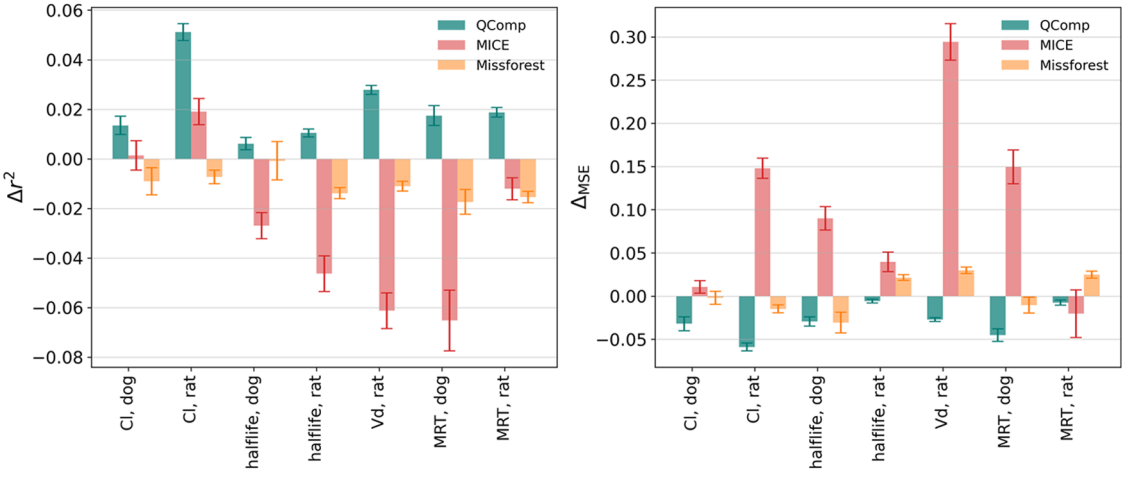


Figure 5. Average and the error bar of the change in the Pearson r^2 score (a positive change means improvement) and MSE (a negative change means improvement) over the base QSAR model for in vivo assays. The average and the error bar are calculated from the 50-bin splitting of the test set.

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Table 3. Pearson r^2 Scores of the Base QSAR Model Chemprop, MICE, Missforest, and QComp on ADMET-780k Data set with Compound-Based Temporal Splitting[a]

assay name	Chemprop	MICE	Missforest	QComp
Papp	0.721	**0.723**	***0.714***	**0.723**
CaV 1.2	0.352	***0.362***	0.353	**0.371**
NaV 1.5	0.347	***0.361***	0.340	**0.362**
hepatocyte Cl, dog	0.430	***0.493***	0.414	**0.517**
microsome Cl, dog	0.494	**0.598**	0.466	***0.594***
hepatocyte Cl, human	0.413	***0.512***	0.414	**0.537**
microsome Cl, human	0.472	***0.548***	0.514	**0.600**
hepatocyte Cl, rat	0.365	**0.476**	0.357	***0.473***
microsome Cl, rat	0.499	***0.620***	0.559	**0.642**
CYP2C8	0.442	**0.457**	0.434	***0.453***
CYP2C9	0.400	**0.423**	0.395	***0.422***
CYP2D6	0.224	***0.245***	0.152	**0.247**
CYP3A4	0.405	***0.429***	0.413	**0.433**
CYP,TDI,3A4,ratio	0.140	**0.153**	0.140	***0.152***
EPSA	0.816	***0.811***	0.792	**0.814**
hERG MK499	0.470	0.453	***0.467***	**0.469**
Fu,p, human	0.596	**0.632**	0.587	***0.616***
logD	0.837	0.841	***0.842***	**0.847**
PAMPA	0.494	0.002	**0.518**	***0.495***
PXR activation	0.384	***0.314***	0.302	**0.384**
Fu,p, rat	0.637	***0.641***	0.623	**0.645**
Fassif Solub	0.384	***0.443***	0.436	**0.463**
SOLY7	0.585	0.620	**0.691**	***0.671***
PGP, rat	0.494	***0.499***	0.485	**0.501**

[a]Only in vitro data are utilized for the input of in vitro assays. For each assay, the highest r^2 score among different imputation methods is marked in **bold**. The second-highest r^2 score in imputation methods is marked in **bold and italic**.

This shows that in vivo assays only marginally help in imputing in vitro assays. This is expected considering that in vivo properties involve a lot of biochemical processes that cannot be characterized by a few in vitro assays.

Next, following the same protocol of partitioning the test set into 50 bins as used in Sections 3.3.1 and 3.4, we study the mean and the error bars of the changes in the Pearson r^2 score and MSE brought by the imputation method over the base QSAR model. The results as reported by Figure 6 validate the statistical significance of the improvement brought by QComp for about one-third of in vitro assays, including "hepatocyte Cl", "microsome Cl", "Fassif Solub", and "SOLY7". Similar conclusions are also drawn for other metrics (R^2, MAE) and for clustering-based split (see Sections E and F of the Supporting Information).

3.6. Rational Decision-Making with QComp. When QComp predicts a missing assay, it also gives the GOC brought by available experimental data. GOC quantifies the reduction in the statistical uncertainty of a QComp prediction compared with the corresponding base QSAR prediction. In practice, the GOC can be used as an indicator of how effective an imputation process is.

Within our framework, GOC is a statistical quantity that does not depend on the chemical structures of the individual compounds. Specifically, for imputing a missing assay-k of an arbitrary compound, the GOC depends only on the indices of the other assays with available experimental data for this compound. This allows a convenient greedy scheme for the decision-making procedure in the experimental ADMET studies.

We consider the scenario that the assay-k is of primary interest for a new compound with no experimental data yet. We assume that the direct measurement of assay-k is expensive. For example, assay-k is an in vivo property. The goal here is to measure a few in vitro assays instead and impute in vivo assay-k

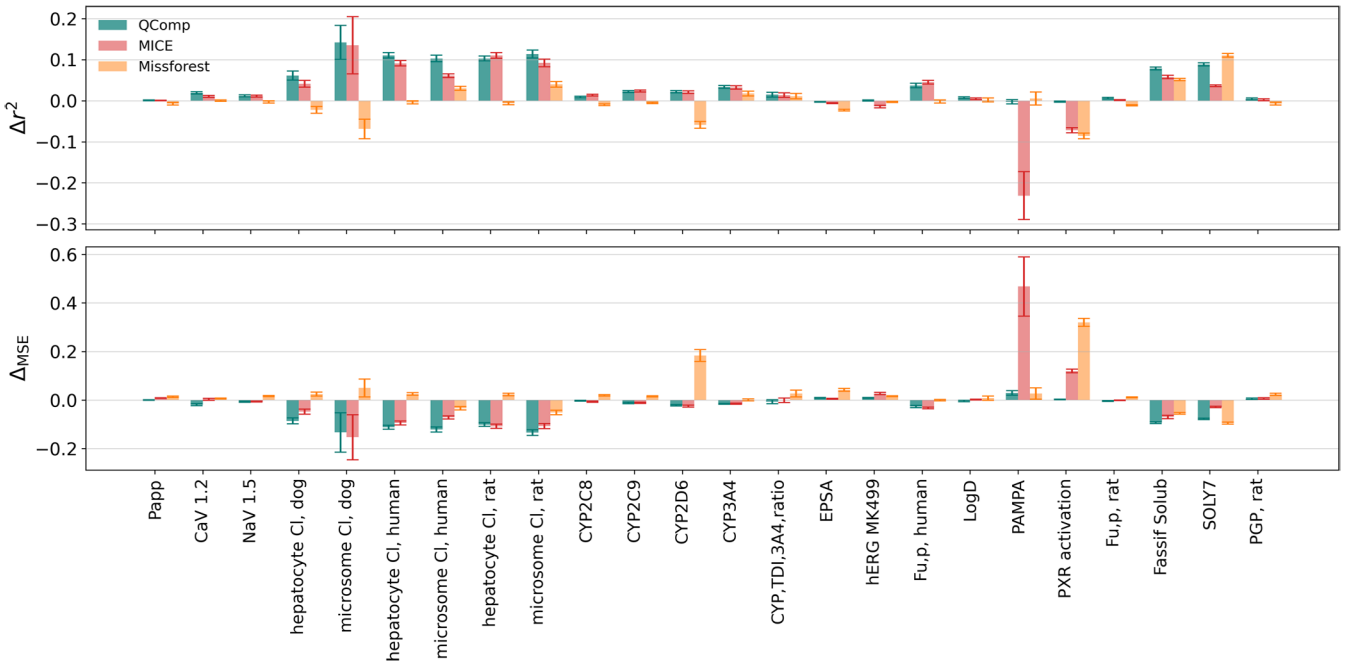


Figure 6. Average and the error bar of the change in the Pearson r^2 score (a positive change means improvement) and MSE (a negative change means improvement) over the base QSAR model for in vitro assays. The average and the error bar are calculated from the 50-bin splitting of the test set.

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

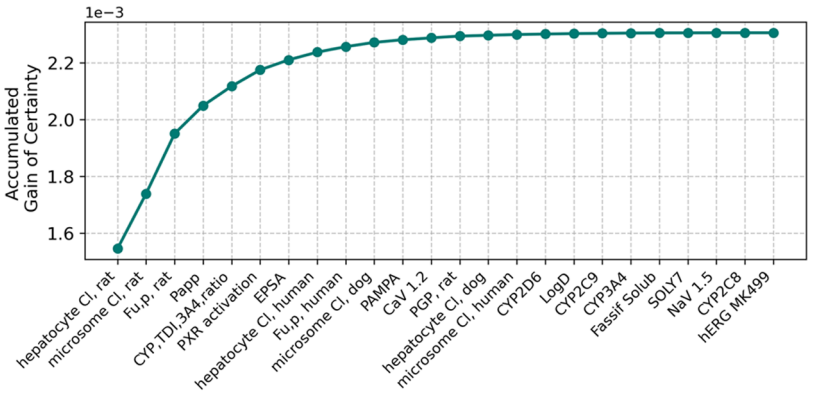


Figure 7. Gain of certainty accumulated along the optimal (gray) sequence of in vitro assays when setting "MRT, rat" as the target assay.

with the acquired in vitro data and the pre-existing QSAR prediction. For such circumstances, we propose a scheme that predicts the sequence of in vitro assays to be measured for maximizing the short-term gain. The scheme first prioritizes the measurement of the in vitro assay-k_0 that brings the highest GOC for assay-k. Then, after assay-k_0 gains experimental data, the GOC for assay-k with respect to the measurement of other in vitro assays changes. One can recalculate the GOC and prioritize again the assay that brings the highest GOC for assay-k. This procedure repeats until the GOC for assay-k is ignorable for any remaining missing in vitro assay, meaning we cannot significantly improve the quality of imputation anymore.

We illustrate this greedy scheme with the ADMET-780k data set. We let "MRT, rat" be the assay of primary interest. We assume that all in vivo experimental data ("half-life, rat", "Cl, rat", "Vd,rat", "half-life, dog", "MRT, dog", "Cl, dog") are not available, and we allow all in vitro assays to be measured. Within the greedy scheme, we determined the optimal sequence of in vitro assays to be measured. The results, along with the accumulated GOC, are given in Figure 7. The accumulated GOC is the cumulative sum of the GOC of each new measurement along the sequence.

We find that the top three assays in the optimal sequence are "hepatocyte Cl, rat", "microsome Cl, rat", and "Fu,p, rat". They contribute to more than 80% of the final accumulated GOC. The types of the top three assays also align seamlessly with the empirical expectation that the in vitro properties directly associated with rats should efficiently improve the imputation of "MRT, rat". Compared to the top three assays, other in vitro assays bring only marginal GOC. The accumulated GOC saturates around the "PAMPA" assay. Therefore, in practice, the termination of the experimental sequence can be set at any point between "Fu,p, rat" and "PAMPA", depending on the budget and the cost of individual experiments.

The optimal sequence is also largely consistent with a direct ranking of assays by their covariance with the end point "MRT, rat", which can be extracted from the trained covariance matrix Σ. The overall consistency can be understood as the definition of GOC for one end point is essentially a normalized sum of covariance between the end point and other assays. The optimal sequence therefore reveals the inner structure of the covariance matrix in a straightforward way for the chosen end point.

4. CONCLUSIONS

We developed the QComp approach for reliable imputation. QComp can dynamically exploit newly acquired data to improve the prediction of missing data without retraining the numerical model.

We benchmarked QComp for ADMET imputation. QComp systematically improves upon structure-based QSAR models Chemprop and outperforms standard iterative imputation methods, including MICE and Missforest, when they are all provided with the same side information. Notably, for assays where imputation approaches do not show an advantage over plain QSAR prediction, QComp yields similar r^2 scores as the base QSAR. Other imputation methods, however, may suffer from catastrophic failure. Moreover, we show that QComp improves by roughly 10% the prediction of all in vivo assays in our data set when only in vitro data are utilized, suggesting a promising potential of QComp for practical application. An extra advantage of QComp is providing a simple yet useful scheme for rational decision-making in preclinical drug discovery research, where acquiring in vivo assays is considerably more convenient than acquiring in vitro assays. Moreover, through the study of the QM8 data set (in the Supporting Information), we see the broad applicability of QComp in the future.

These results demonstrate that QComp is accurate, robust, interpretable, and versatile. These advantages allow QComp to be integrated into most QSAR workflows of preclinical studies without restructuring the inference stage of the existing QSAR models. In other words, QComp can be performed after base QSAR inference. QComp is also of low cost considering the $O(p^3)$ scaling for imputing a new row of data. The bottleneck is mainly the cubic-scaled Cholesky inverse of covariance matrices. For a typical AMDET data set, p is between 1 to a few hundred. So, the Cholesky inverse can be done very efficiently. For example, with an Apple M3 Pro CPU chip, it takes only a few seconds to impute the test set (around 77,000 rows with ∼83% data missing) of the 31-column ADMET-780k data set. We hence foresee more systematic incremental applications of QComp in drug discovery.

Further development of QComp may focus on overcoming the stringent assumption on the base QSAR model, i.e., the QSAR-experiment deviations should consistently obey a correlated, normal distribution over relevant chemical spaces. Case studies reported in this paper, by and large, follow such an assumption. However, there is no systematic way to enforce these assumptions for generic tasks. To address this issue, a possible generalization of QComp is to allow a structure-

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

dependent covariance matrix Σ, such that the correlation between assays can vary over chemical spaces. The training of such Σ can be incorporated into the current likelihood-based training framework. It follows that the GOC can also be estimated with structure dependence, making it a more useful tool for decision-making.

■ ASSOCIATED CONTENT

Data Availability Statement

Python implementation of QComp and the public ADMET data set are available at https://github.com/MSDLLCpapers/QComp. This work uses the proprietary ADMET data set from Merck & Co. Inc. (Rahway, NJ, USA) to provide conclusions in a real-world setting. Results using the public ADMET data set are also reported for reproducibility.

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.5c00059.

Detailed model information, descriptions of data sets, additional benchmarking results, and figures (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Bingjia Yang − *Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical, Merck & Co., Inc., South San Francisco, California 94080, United States;* orcid.org/0000-0003-4074-9553; Email: bingjia.yang@merck.com

Xiang Yu − *Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical, Merck & Co., Inc., West Point, Pennsylvania 19486, United States;* orcid.org/0000-0002-8217-5896; Email: xiang.yu2@merck.com

Authors

Yunsie Chung − *Computational and Structural Chemistry, Merck & Co., Inc, South San Francisco, California 94080, United States*

Archer Y. Yang − *Department of Mathematics and Statistics, McGill University; Mila - Quebec AI Institute, Montreal, Quebec H2S 3H1, Canada*

Bo Yuan − *Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical, Merck & Co., Inc., South San Francisco, California 94080, United States*

Tianchi Chen − *Decision Science, Merck & Co., Inc., Cambridge, Massachusetts 02141, United States*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.5c00059

Author Contributions

All authors contributed to the conception and design of the research project. All authors discussed the results, contributed to the writing, and approved the final version of the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors thank Ti-chiun Chang, Liying Zhang, and Alan C. Cheng for their insightful suggestions in preparing this manuscript. This work was funded by Merck& Co., Inc., Rahway, NJ.

■ REFERENCES

(1) von Ragué Schleyer, P.; Allinger, N. L.; Clark, T.; Gasteiger, J.; Kollman, P.; Schaefer, H. F.; Schreiner, P. R. *Encyclopedia of Computational Chemistry*; Wiley Online Library, 1998.

(2) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(3) Noble, W. S. What is a support vector machine? *Nat. Biotechnol.* **2006**, *24*, 1565−1567.

(4) Obrezanova, O.; Csányi, G.; Gola, J. M.; Segall, M. D. Gaussian processes: a method for automatic QSAR modeling of ADME properties. *J. Chem. Inf. Model.* **2007**, *47*, 1847−1857.

(5) Ma, J.; Sheridan, R. P.; Liaw, A.; Dahl, G. E.; Svetnik, V. Deep neural nets as a method for quantitative structure−activity relationships. *J. Chem. Inf. Model.* **2015**, *55*, 263−274.

(6) Gawehn, E.; Hiss, J. A.; Schneider, G. Deep learning in drug discovery. *Mol. Inf.* **2016**, *35*, 3−14.

(7) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E.*Neural Message Passing for Quantum Chemistry*, International Conference on Machine Learning, 2017; pp 1263−1272.

(8) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370−3388.

(9) Huang, K.; Fu, T.; Glass, L. M.; Zitnik, M.; Xiao, C.; Sun, J. DeepPurpose: a deep learning library for drug−target interaction prediction. *Bioinformatics* **2021**, *36*, 5545−5547.

(10) Tropsha, A.; Isayev, O.; Varnek, A.; Schneider, G.; Cherkasov, A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. *Nat. Rev. Drug Discovery* **2023**, *23*, 141−155.

(11) Dahl, G. E.; Jaitly, N.; Salakhutdinov, R. Multi-task neural networks for QSAR predictions, arXiv:1406.1231. arXiv.org e-Print archive. https://arxiv.org/abs/1406.1231, 2014.

(12) Kearnes, S.; Goldman, B.; Pande, V. Modeling industrial ADMET data with multitask networks, arXiv:1606.08793. arXiv.org e-Print archive. https://arxiv.org/abs/1606.08793, 2016.

(13) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *J. Chem. Inf. Model.* **2019**, *59*, 1253−1268.

(14) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET prediction with multitask deep featurization. *J. Med. Chem.* **2020**, *63*, 8835−8848.

(15) Muratov, E. N.; Bajorath, J.; Sheridan, R. P.; Tetko, I. V.; Filimonov, D.; Poroikov, V.; Oprea, T. I.; Baskin, I. I.; Varnek, A.; Roitberg, A.; et al. QSAR without borders. *Chem. Soc. Rev.* **2020**, *49*, 3525−3564.

(16) Jia, L.; Gao, H. Machine Learning for in silico ADMET prediction. *Artif. Intell. Drug Des.* **2022**, *2390*, 447−460.

(17) Walter, M.; Allen, L. N.; de la Vega de León, A.; Webb, S. J.; Gillet, V. J. Analysis of the benefits of imputation models over traditional QSAR models for toxicity prediction. *J. Cheminf.* **2022**, *14*, 1−27.

(18) Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520−525.

(19) Oba, S.; Sato, M.-a.; Takemasa, I.; Monden, M.; Matsubara, K.-i.; Ishii, S. A Bayesian missing value estimation method for gene expression profile data. *Bioinformatics* **2003**, *19*, 2088−2096.

(20) Liew, A. W.-C.; Law, N.-F.; Yan, H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. *Briefings Bioinf.* **2011**, *12*, 498−513.

(21) Kim, H.; Golub, G. H.; Park, H. Missing value estimation for DNA microarray gene expression data: local least squares imputation. *Bioinformatics* **2005**, *21*, 187−198.

(22) Van Buuren, S.; Oudshoorn, K. *Flexible Multivariate Imputation by MICE*; TNO: Leiden, 1999.
(23) Stekhoven, D. J.; Bühlmann, P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* **2012**, *28*, 112−118.
(24) Simm, J.; Arany, A.; Zakeri, P.; Haber, T.; Wegner, J. K.; Chupakhin, V.; Ceulemans, H.; Moreau, Y. *Macau: scalable bayesian multi-relational factorization with side information using MCMC*, arXiv:1509.04610. arXiv.org e-Print archive. https://arxiv.org/abs/1509.04610, 2015.
(25) Koren, Y.; Bell, R.; Volinsky, C. Matrix factorization techniques for recommender systems. *Computer* **2009**, *42*, 30−37.
(26) Whitehead, T. M.; Irwin, B. W.; Hunt, P.; Segall, M. D.; Conduit, G. J. Imputation of assay bioactivity data using deep learning. *J. Chem. Inf. Model.* **2019**, *59*, 1197−1204.
(27) Martin, E.; Mukherjee, P.; Sullivan, D.; Jansen, J. Profile-QSAR: a novel meta-QSAR method that combines activities across the kinase family to accurately predict affinity, selectivity, and cellular activity. *J. Chem. Inf. Model.* **2011**, *51*, 1942−1956.
(28) Martin, E. J.; Polyakov, V. R.; Tian, L.; Perez, R. C. Profile-QSAR 2.0: Kinase Virtual Screening Accuracy Comparable to Four-Concentration IC50s for Realistically Novel Compounds. *J. Chem. Inf. Model.* **2017**, *57*, 2077−2088.
(29) Oberman, H. I.; van Buuren, S.; Vink, G.et al. *Missing the point: Non-convergence in iterative imputation algorithms*, First Workshop on the Art of Learning with Missing Values (Artemiss) hosted by the 37th International Conference on Machine Learning (ICML), 2020.
(30) Nguyen, C. D.; Carlin, J. B.; Lee, K. J. Practical strategies for handling breakdown of multiple imputation procedures. *Emerg. Themes Epidemiol.* **2021**, *18*, No. 5.
(31) Heid, E.; Greenman, K. P.; Chung, Y.; Li, S.-C.; Graff, D. E.; Vermeire, F. H.; Wu, H.; Green, W. H.; McGill, C. J. Chemprop: A Machine Learning Package for Chemical Property Prediction. *J. Chem. Inf. Model.* **2024**, *64*, 9−17.
(32) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **2018**, *9*, 513−530.
(33) Iwata, H.; Matsuo, T.; Mamada, H.; Motomura, T.; Matsushita, M.; Fujiwara, T.; Maeda, K.; Handa, K. Predicting Total Drug Clearance and Volumes of Distribution Using the Machine Learning-Mediated Multimodal Method through the Imputation of Various Nonclinical Data. *J. Chem. Inf. Model.* **2022**, *62*, 4057−4065.
(34) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Res.* **2023**, *51*, D1373−D1380.
(35) Watanabe, R.; Esaki, T.; Kawashima, H.; Natsume-Kitatani, Y.; Nagao, C.; Ohashi, R.; Mizuguchi, K. Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Mol. Pharmaceutics* **2018**, *15*, 5302−5311.
(36) Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M. Á. Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches. *Pharmaceutics* **2022**, *14*, No. 1998, DOI: 10.3390/pharmaceutics14101998.
(37) Esposito, C.; Wang, S.; Lange, U. E.; Oellien, F.; Riniker, S. Combining machine learning and molecular dynamics to predict P-glycoprotein substrates. *J. Chem. Inf. Model.* **2020**, *60*, 4730−4749.
(38) Braga, R. C.; Alves, V. M.; Silva, M. F.; Muratov, E.; Fourches, D.; Lião, L. M.; Tropsha, A.; Andrade, C. H. Pred-hERG: ANovel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Mol. Inf.* **2015**, *34*, 698−701.
(39) Aliagas, I.; Gobbi, A.; Lee, M. L.; Sellers, B. D. Comparison of logP and logD correction models trained with public and proprietary data sets. *J. Comput.-Aided Mol. Des.* **2022**, *36*, 253−262.
(40) Perryman, A. L.; Inoyama, D.; Patel, J. S.; Ekins, S.; Freundlich, J. S. Pruned Machine Learning Models to Predict Aqueous Solubility. *ACS Omega* **2020**, *5*, 16562−16567.
(41) Meng, J.; Chen, P.; Wahib, M.; Yang, M.; Zheng, L.; Wei, Y.; Feng, S.; Liu, W. Boosting the predictive performance with aqueous solubility dataset curation. *Sci. Data* **2022**, *9*, No. 71, DOI: 10.1038/s41597-022-01154-3.
(42) Vermeire, F. H.; Chung, Y.; Green, W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *J. Am. Chem. Soc.* **2022**, *144*, 10785−10797.
(43) Sheridan, R. P.; Karnachi, P.; Tudor, M.; Xu, Y.; Liaw, A.; Shah, F.; Cheng, A. C.; Joshi, E.; Glick, M.; Alvarez, J. Experimental error, kurtosis, activity cliffs, and methodology: What limits the predictivity of quantitative structure−activity relationship models? *J. Chem. Inf. Model.* **2020**, *60*, 1969−1982.
(44) Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Félix, E.; Magariños, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; et al. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930−D940.
(45) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. *Uni-Mol: A Universal 3D Molecular Representation Learning Framework*, 11th International Conference on Learning Representations, 2023.
(46) Sheridan, R. P. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *J. Chem. Inf. Model.* **2013**, *53*, 783−790.

https://doi.org/10.1021/acs.jcim.5c00059
J. Chem. Inf. Model. XXXX, XXX, XXX−XXX

Supporting Information for QComp: A QSAR-Based Imputation Framework for Drug Discovery

Bingjia Yang,[*,†] Yunsie Chung,[‡] Archer Y. Yang,[¶] Bo Yuan,[†] Tianchi Chen,[§] and Xiang Yu[*,‖]

†*Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical, Merck & Co., Inc., South San Francisco, CA, 94080, USA*

‡*Computational and Structural Chemistry, Merck & Co., Inc, South San Francisco, CA 94080, USA*

¶*Department of Mathematics and Statistics, McGill University; Mila - Quebec AI Institute, Montreal, Quebec, Canada*

§*Decision Science, Merck & Co., Inc., Cambridge, MA, 02141, USA*

‖*Pharmacokinetics, Dynamics, Metabolism, and Bioanalytical, Merck & Co., Inc., West Point, PA, 19486, USA*

E-mail: bingjia.yang@merck.com; xiang.yu2@merck.com

A Model details

Chemprop Chemprop consists of (1) a message passing network in which a graph structure of a molecule is transformed into a molecular latent representation and (2) a feed forward network which makes property predictions from the latent representation. A multitask model is employed to predict all ADMET assays simultaneously as former studies have shown that multitask models achieve better performance than single-task models when multiple tasks are correlated with each other.[1,2] For the ADMET-780k dataset, the model is trained with an ensemble of 4 models, each initialized with a different random seed, and an epoch of 60. 10% of the training set is randomly chosen as a validation set and used to determine the best epoch for the model during training. A hidden size of 600 and a depth of 4 are selected for the message-passing network. A hidden size of 1300 and a depth of 4 are selected for the feed-forward network. A normalized sum is used to aggregate the atomic embedding into a molecular embedding during the message-passing phase. For the public ADMET dataset, the models are trained using the same hyperparameters and ensembles with an epoch of 40.

QComp The parameters of a QComp model are initialized as follows:

- $\mathbf{B}$: A $p \times p$ identity matrix. p is the total number of biochemical activities.
- $\mathbf{b}$: A zero vector of size p.
- $\mathbf{\Sigma}$: $\mathbf{\Sigma}$ is parameterized by its Cholesky decomposition $\mathbf{\Sigma} = \boldsymbol{L}\boldsymbol{L}^T$. The lower triangle matrix $\boldsymbol{L}$ is initialized as a $p \times p$ diagonal matrix, where the i-th diagonal term is the standard deviation of the available data of the i-th biochemical activity.

To train the QComp model for the ADMET-780k dataset, we let the total number of epochs be 5 and the batch size be 5000. We use the ADAM optimizer[3] for gradient descent in all our studies. Here, the initial learning rate is 0.01. The learning rate decays by 0.5 every epoch. For the public ADMET dataset, the number of epoch is 10, with a batch size of 100. We use an initial learning rate of 0.001. The learning rate decays by 0.5 every epoch.

MICE We use the IterativeImputer implemented in the fancyimpute[4] package for MICE[5] imputation. All parameters are default values (max_iter=10, tol=0.001).

Missforest Missforest is an iterative imputation method similar to MICE. The difference is that the regression model in Missforest is random forest. In our study, we use the class "IterativeImputer" in the scikit-learn package for Missforest imputation. The regression model (estimator) is the random forest regressor (n_estimators=4, max_depth=10. max_samples=0.5) in scikit-learn. The maximal iteration for iterative imputation is 25 with tol=0.1.

B Composite Uncertainty

We let the uncertainty of $\boldsymbol{f}^{(i)}$ be denoted by $\boldsymbol{\sigma}^{(i)} = (\sigma_1^{(i)}, \sigma_2^{(i)}, \cdots, \sigma_p^{(i)})$. In practice, both $\boldsymbol{f}^{(i)}$ and $\boldsymbol{\sigma}^{(i)}$ are calculated from an ensemble of deterministic QSAR models trained on the same dataset but initialized differently. $\boldsymbol{f}^{(i)}$ and $\boldsymbol{\sigma}^{(i)}$ are respectively the ensemble average and the standard deviation of QSAR predictions. Assuming the components of $\boldsymbol{\sigma}^{(i)}$ are not correlated with each other, the ensemble covariance matrix associated with $\boldsymbol{f}^{(i)}$ is a diagonal matrix $\boldsymbol{\Sigma}_f^{(i)}$ with $(\sigma_j^{(i)})^2$ as j-th diagonal terms. Through propagation of uncertainty, the ensemble covariance matrix of $\boldsymbol{\mu}^{(i)}$ is $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(i)} = \mathbf{B}^\top \boldsymbol{\Sigma}_f^{(i)} \mathbf{B}$. We can use $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(i)}$ to compute the ensemble deviation associated with $\tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)}$. But before that, we need to define some extra notations. Let $(\mathbf{y}^{(i)})_j$ be an arbitrary missing assay and $(\mathbf{y}^{(i)})_k$ any known assay. $1 \leq j, k \leq p$ are the indices of the assays in the whole collection of p assays. In terms of the partition $\mathbf{y}^{(i)} = (\mathbf{y}^{\mathrm{M}(i)}, \mathbf{y}^{\mathrm{O}(i)})$, we use $j^{\mathrm{M}(i)}$ to denote the index of the assay-j in the sub-vector $\mathbf{y}^{\mathrm{M}(i)}$, and $k^{\mathrm{O}(i)}$ the index of the assay-k in $\mathbf{y}^{\mathrm{O}(i)}$. So there is an one-to-one mapping between j and $j^{\mathrm{M}(i)}$, k and $k^{\mathrm{O}(i)}$. Additionally, we define $D^{(i)} = \boldsymbol{\Sigma}^{\mathrm{MO}(i)} (\boldsymbol{\Sigma}^{\mathrm{OO}(i)})^{-1}$. So we can express the

ensemble deviation associated to $\tilde{\boldsymbol{\mu}}^{\mathrm{M}(i)}$ in simple terms:

$$(\sigma^{(i)}_{\tilde{\boldsymbol{\mu}}^{\mathrm{M}}})^2_{j^{\mathrm{M}(i)}} = (\boldsymbol{\Sigma}^{(i)}_{\mu})_{jj} + \sum_{k^{\mathrm{O}(i)}=1}^{p^{(i)}_{\mathrm{O}}} (D^{(i)}_{j^{\mathrm{M}(i)}k^{\mathrm{O}(i)}})^2 (\boldsymbol{\Sigma}^{(i)}_{\mu})_{kk}. \tag{1}$$

To incorporate the Gaussian statistical uncertainty assumed by QComp, we construct the composite uncertainty

$$(\varsigma^{(i)}_{\tilde{\boldsymbol{\mu}}^{\mathrm{M}}})^2_{j^{\mathrm{M}(i)}} = (\sigma^{(i)}_{\tilde{\boldsymbol{\mu}}^{\mathrm{M}}})^2_{j^{\mathrm{M}(i)}} + (\widetilde{\boldsymbol{\Sigma}}^{\mathrm{MM}(i)})_{j^{\mathrm{M}(i)}j^{\mathrm{M}(i)}}. \tag{2}$$

This final expression serves as a practical but rough estimation for the error of optimal imputation.

C Dataset details

C.1 Proprietary ADMET-780k dataset

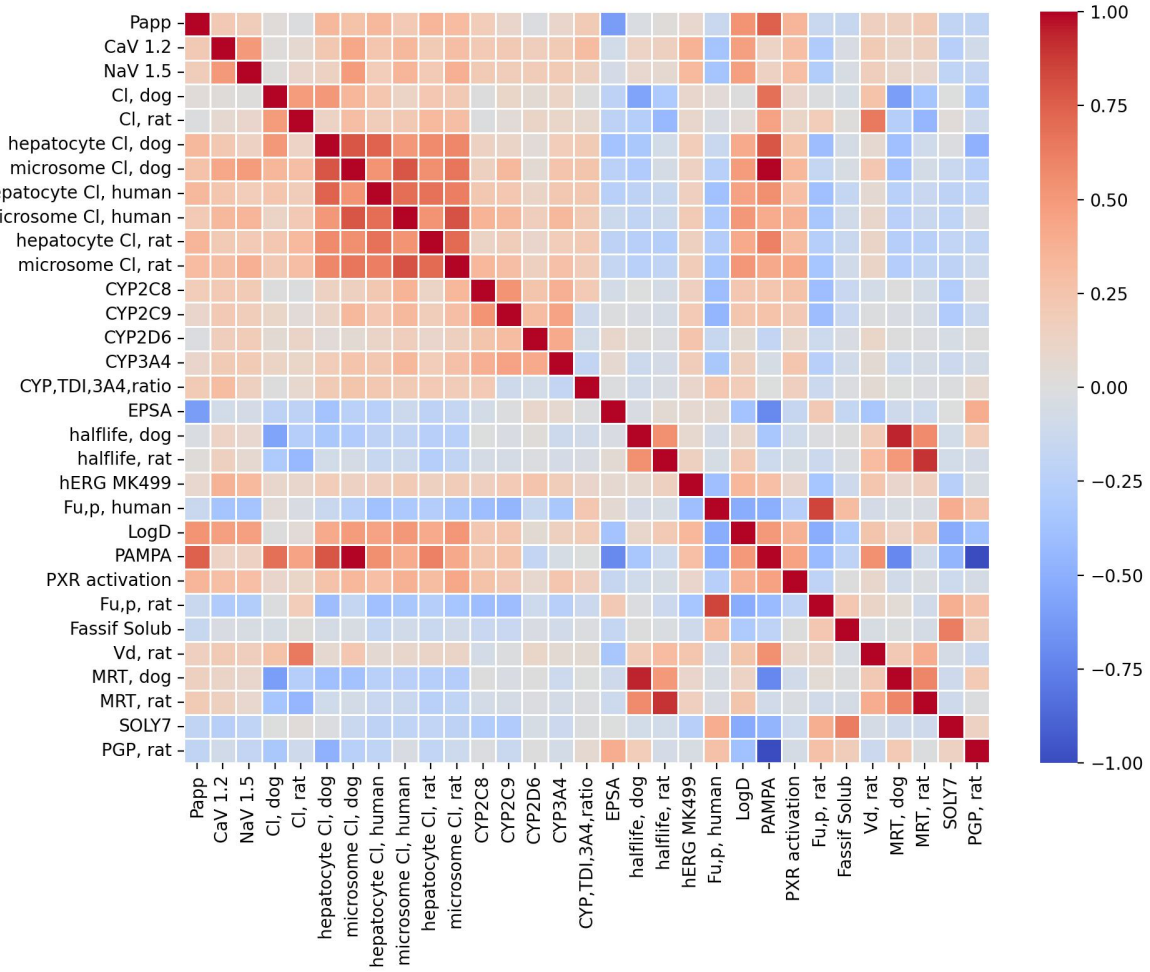


Figure S1: ADMET-780k dataset: Pearson correlation heatmap.

We split the dataset temporally according to the synthesis date of each compound.

Table S1: ADMET-780k dataset: Number of compounds in training and test sets for compound-based temporal split

Assay	Short Name	Train Size	Test Size	Description
Absorption Papp	Papp	49272	5774	Apparent permeability through cell monolayers
Ca Na Ion Channel CaV 1.2 Inhibition	CaV 1.2	142473	11266	Inhibition of CaV1.2 ion channels
Ca Na Ion Channel NaV 1.5 Inhibition	NaV 1.5	135994	11449	Inhibition of NaV1.5 ion channels
Clearance Dog	Cl, dog	19633	2203	In vivo plasma clearance in dog
Clearance Rat	Cl, rat	64395	11505	In vivo plasma clearance in rat
CLint Dog hepatocyte	hepatocyte Cl, dog	7232	1144	In vitro dog hepatocyte intrinsic clearance
CLint Dog microsome	microsome Cl, dog	3946	467	In vitro dog microsome intrinsic clearance
CLint Human hepatocyte	hepatocyte Cl, human	36476	5717	In vitro human hepatocyte intrinsic clearance
CLint Human microsome	microsome Cl, human	44252	4443	In vitro human microsome intrinsic clearance
CLint Rat hepatocyte	hepatocyte Cl, rat	33531	5644	In vitro rat hepatocyte intrinsic clearance
CLint Rat microsome	microsome Cl, rat	41609	4240	In vitro rat microsome intrinsic clearance
CYP Inhibition 2C8	CYP2C8	58548	11791	Inhibition of CYP2C8
CYP Inhibition 2C9	CYP2C9	211790	11087	Inhibition of CYP2C9
CYP Inhibition 2D6	CYP2D6	211776	8732	Inhibition of CYP2D6
CYP Inhibition 3A4	CYP3A4	213576	10146	Inhibition of CYP3A4
CYP TDI 3A4 Ratio	CYP,TDI,3A4,ratio	38477	1848	Time-dependent CYP3A4 inhibition via NADPH IC50 shift
EPSA	EPSA	18863	21937	Exposed polarity measurement (Experimental Polar Surface Area)
Halflife Dog	halflife, dog	21541	2345	Half-life in vivo in dog
Halflife Rat	halflife, rat	70892	11996	Half-life in vivo in rat
hERG MK499	hERG MK499	349226	14993	Binding to the HERG channel through the displacement of MK499
Human fraction unbound plasma-current	Fu,p, human	19478	2775	Fraction of unbound drug in human plasma
LogD HPLC pH 7.0	LogD	457038	2663	LogD at pH 7
PAMPA	PAMPA	2907	1094	Parallel artificial membrane permeability assay
PXR activation	PXR activation	219501	14739	PXR receptor activation
Rat fraction unbound plasma-current	Fu,p, rat	43382	11081	Fraction of unbound drug in rat plasma
Solubility Fassif	Fassif Solub	247693	68504	Solubility in FaSSIF (Fasted State Simulated Intestinal Fluid)
Volume of Distribution Rat	Vd, rat	64431	11490	In vivo volume of distribution in rat
Dog MRT	MRT, dog	17506	2196	Mean residence time in dog
Rat MRT	MRT, rat	60538	11467	Mean residence time in rat
SOLY 7	SOLY7	412744	58766	Kinetic solubility at pH 7
Rat PGP 1uM	PGP, rat	25214	2417	Rat BA:AB efflux ratio P-gp

C.2 Public ADMET dataset

The public dataset (Sec. D) used in this work is compiled from various public sources including Ref. 6 (ChEMBL, CC BY-SA 3.0 DEED), Ref. 7 (CC-BY-NC-ND 4.0), Ref. 8(PubChem), Ref. 9 (from PharmaPendium and ChEMBL), Ref. 10 (CC BY 4.0 DEED), Ref. 11 (ChEMBL), Ref. 12(ChEMBL), Ref. 13(ChEMBL), Ref. 14, Ref. 15(CC BY 4.0 DEED), and Ref. 16(CC BY 4.0 DEED).

Each assay data is converted to an appropriate unit as indicated in the Table S2. The SMILES identifiers from different data sources are validated and canonicalized using RDKit.[17] The mean values are used when multiple data points are found for the same compound.

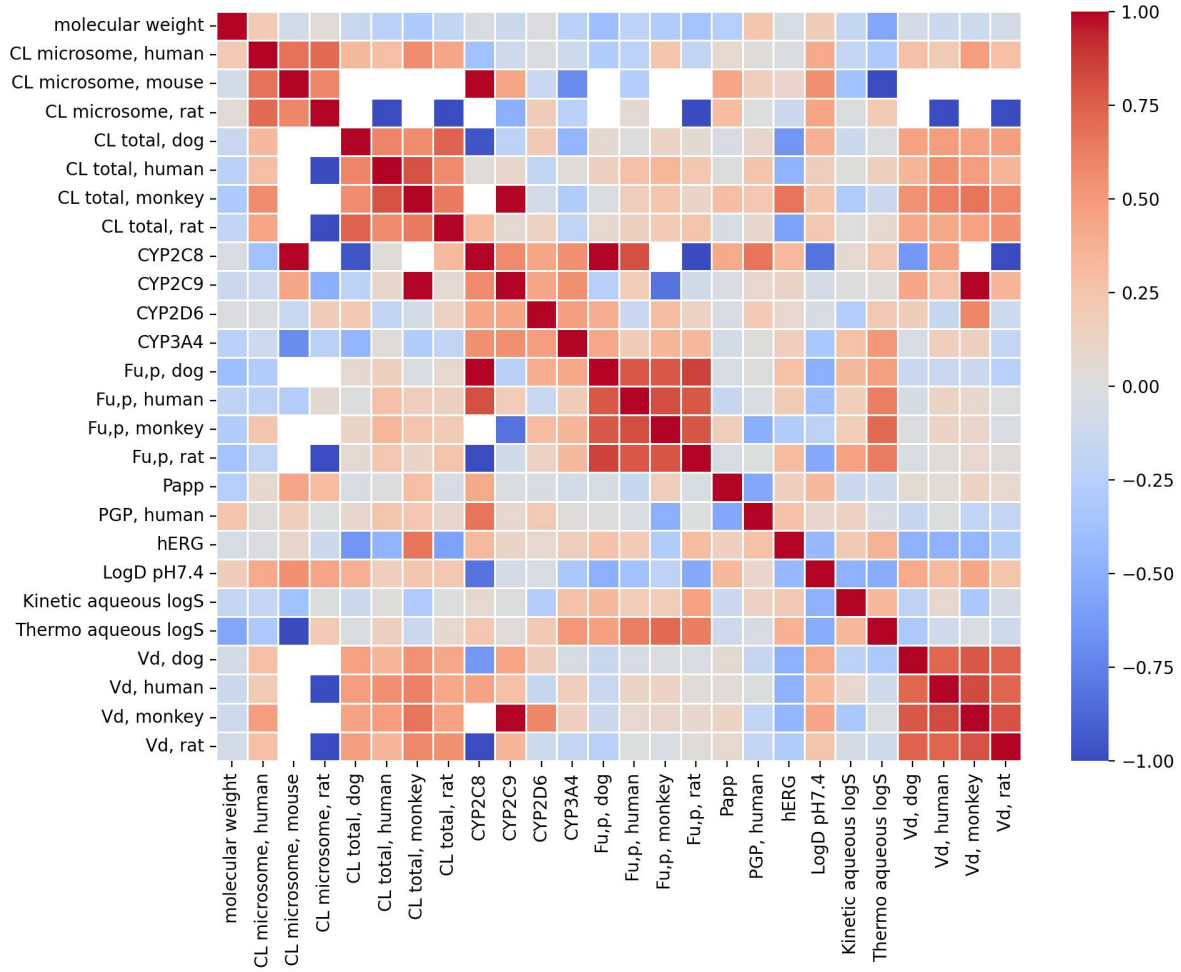


Figure S2: The public dataset: Pearson correlation heatmap. Blank blocks indicate missing values (assays appearing mutually exclusively in the dataset).

Table S2: Public dataset: size, assay name mapping, units and authoritative descriptions

Assay	Data Count	Short Name	Units	Description
CL_microsome_human	5218	CL microsome, human	$\log_{10}$(mL/min/kg)	In vitro intrinsic clearance measured in human microsomes
CL_microsome_mouse	663	CL microsome, mouse	$\log_{10}$(mL/min/kg)	In vitro intrinsic clearance measured in mouse microsomes
CL_microsome_rat	1798	CL microsome, rat	$\log_{10}$(mL/min/kg)	In vitro intrinsic clearance measured in rat microsomes
CL_total_dog	284	CL total, dog	$\log_{10}$(mL/min/kg)	Total body clearance in dog
CL_total_human	741	CL total, human	$\log_{10}$(mL/min/kg)	Total body clearance in human
CL_total_monkey	129	CL total, monkey	$\log_{10}$(mL/min/kg)	Total body clearance in monkey
CL_total_rat	387	CL total, rat	$\log_{10}$(mL/min/kg)	Total body clearance in rat
CYP2C8_inhibition	328	CYP2C8	$\log_{10}$(nM IC_{50})	Inhibition of CYP2C8
CYP2C9_inhibition	2374	CYP2C9	$\log_{10}$(nM IC_{50})	Inhibition of CYP2C9
CYP2D6_inhibition	2539	CYP2D6	$\log_{10}$(nM IC_{50})	Inhibition of CYP2D6
CYP3A4_inhibition	4403	CYP3A4	$\log_{10}$(nM IC_{50})	Inhibition of CYP3A4
Dog_fraction_unbound_plasma	179	Fu,p, dog	$\log_{10}$(fraction)	Fraction of drug unbound in dog plasma
Human_fraction_unbound_plasma	2717	Fu,p, human	$\log_{10}$(fraction)	Fraction of drug unbound in human plasma
Monkey_fraction_unbound_plasma	88	Fu,p, monkey	$\log_{10}$(fraction)	Fraction of drug unbound in monkey plasma
Rat_fraction_unbound_plasma	237	Fu,p, rat	$\log_{10}$(fraction)	Fraction of drug unbound in rat plasma
Papp_Caco2	6457	Papp	$\log_{10}(10^{-6}$ cm/s)	Apparent permeability coefficient across Caco-2 cell monolayers
Pgp_human	2073	PGP, human	$\log_{10}$(efflux ratio)	Human BA:AB efflux ratio P-gp
hERG_binding	5108	hERG	$\log_{10}$(nM IC_{50})	hERG binding affinity
LogD_pH_7.4	4190	LogD pH7.4	$\log_{10}$(M/M)	logD measured at pH 7.4
kinetic_logSaq	74895	Kinetic aqueous logS	$\log_{10}$(M)	High-throughput (kinetic) aqueous solubility
thermo_logSaq	11804	Thermo aqueous logS	$\log_{10}$(M)	Equilibrium (thermodynamic) aqueous solubility
VDss_dog	274	Vd, dog	$\log_{10}$(L/kg)	Steady-state volume of distribution determined in dogs
VDss_human	751	Vd, human	$\log_{10}$(L/kg)	Steady-state volume of distribution determined in humans
VDss_monkey	125	Vd, monkey	$\log_{10}$(L/kg)	Steady-state volume of distribution determined in monkeys
VDss_rat	351	Vd, rat	$\log_{10}$(L/kg)	Steady-state volume of distribution determined in rats

D Results on public ADMET dataset

Table S3: Improvement brought by QComp on the public ADMET dataset with 5-fold random splitting. The second column is the Pearson r^2 scores of Chemprop, averaged over five folds. The third column is the Standard error of the mean (SEM) of the r^2 scores. The fourth column is the average improvement of r^2 scores brought by QComp. The fifth column is the corresponding SEM.

Assay name	Chemprop	SEM	Δr^2	SEM
CL microsome, human	0.618	0.012	-0.004	0.003
CL microsome, mouse	0.602	0.015	-0.015	0.009
CL microsome, rat	0.616	0.017	**0.006**	0.006
CL total, dog	0.380	0.030	-0.003	0.030
CL total, human	0.355	0.020	**0.073**	0.016
CL total, monkey	0.452	0.046	**0.148**	0.015
CL total, rat	0.344	0.033	**0.049**	0.035
CYP2C8	0.364	0.072	-0.001	0.004
CYP2C9	0.451	0.026	**0.014**	0.008
CYP2D6	0.445	0.021	**0.012**	0.005
CYP3A4	0.574	0.012	**0.008**	0.004
Fu,p, dog	0.479	0.096	**0.070**	0.029
Fu,p, human	0.712	0.011	**0.001**	0.001
Fu,p, monkey	0.589	0.066	**0.050**	0.029
Fu,p, rat	0.535	0.071	**0.054**	0.011
Papp	0.684	0.011	**0.002**	0.001
PGP, human	0.460	0.014	**0.004**	0.004
hERG	0.645	0.007	0.000	0.001
LogD pH7.4	0.811	0.006	-0.003	0.003
Thermo aqueous logS	0.870	0.004	0.000	0.000
Vd, dog	0.424	0.056	**0.034**	0.033
Vd, human	0.526	0.029	**0.071**	0.010
Vd, monkey	0.512	0.052	**0.119**	0.033
Vd, rat	0.495	0.064	**0.067**	0.025

We adopt a 5-fold random split for the public ADMET dataset. Five base QSAR models yield an average r^2 score of 0.548 for all assays. The QComp models improve the mean r^2 score by 0.03, amounting to a $\sim 5\%$ improvement.

Among all assays, "CL total" (clearance total), "Fu,p" (fraction unbound in plasma), and "Vd" (volume of distribution), associated with dog, human, monkey, and rat (12 assays in total), benefit considerably from QComp imputation with an average 0.061 gain in Pearson

r^2 scores. For assays in this category, Δr^2 is typically larger than half of the SEM of the r^2 scores obtained by the base QSAR model, suggesting the improvement brought by QComp is statistically significant. The assays not in this category, such as "Cl microsome" and "Papp", do not receive considerable improvement from QComp. At the same time, no harm is done by QComp either — Δr^2 and the associated SEM shows no statistical significance compared to the r^2-SEM obtained by Chemprop.

We conclude that QComp works on the public dataset also robustly and efficiently, without one case of catastrophic imputation displayed previously by other methods on the proprietary ADMET-780k dataset. QComp is also robust against the deviation of base QSAR models trained on different splitting of the dataset. Note that, the public dataset is compiled from multiple resources, which does not represent a typical use case of QComp in the industrial setting as reported in the main text.

E Other evaluation metrics for compound-based temporal split

E.1 Benmarking QComp on ADMET-780k dataset

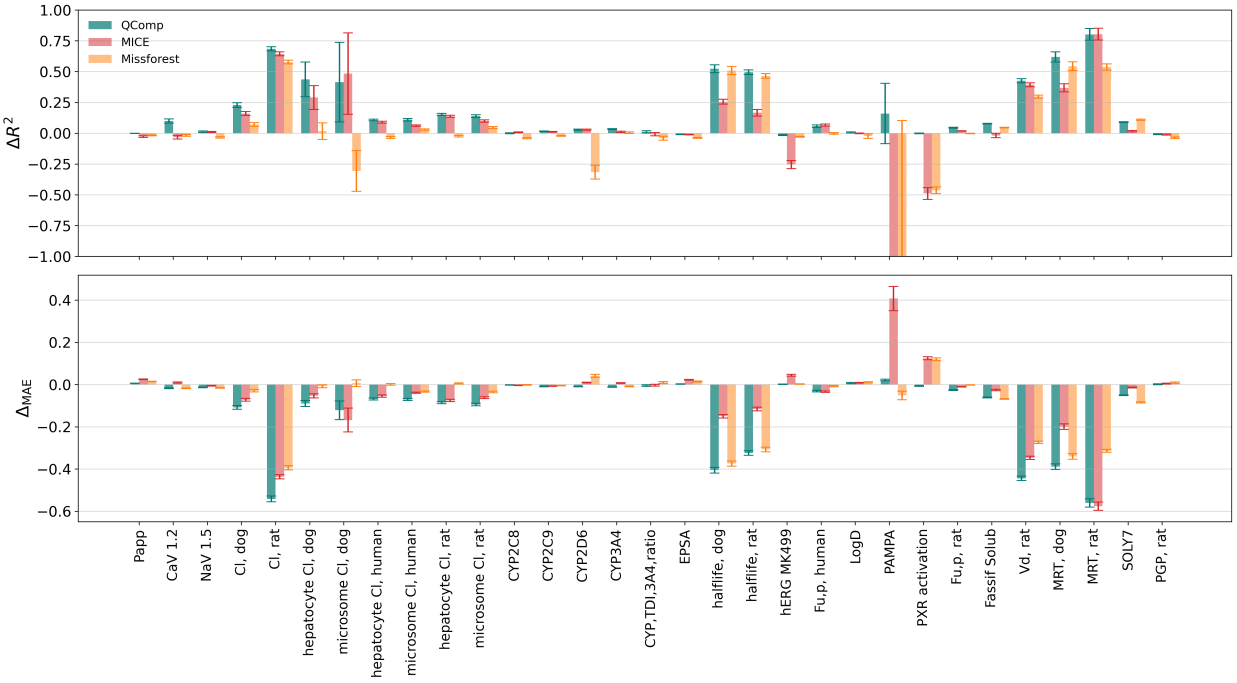


Figure S3: The average and the error bar of the change on R^2 score (a positive change means improvement) and MAE (a negative change means improvement) over the base QSAR model. The average and the error bar are calculated from the 50-bin splitting of the test set.

Table S4: R^2 **scores** of the base QSAR model Chemprop, MICE, Missforest and QComp on ADMET-780k dataset with compound-based temporal splitting. For each assay, the highest R^2 score among different imputation methods is marked in **bold**. The second highest R^2 score in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.646	**0.628**	**0.628**	**0.646**
CaV 1.2	-0.221	-0.267	**-0.255**	**-0.170**
NaV 1.5	0.322	**0.331**	0.293	**0.335**
Cl, dog	0.169	**0.354**	0.247	**0.428**
Cl, rat	0.291	**0.944**	0.882	**0.992**
hepatocyte Cl, dog	0.320	**0.452**	0.308	**0.518**
microsome Cl, dog	0.401	**0.561**	0.389	**0.542**
hepatocyte Cl, human	0.408	**0.502**	0.388	**0.524**
microsome Cl, human	0.471	**0.543**	0.510	**0.597**
hepatocyte Cl, rat	0.322	**0.465**	0.299	**0.482**
microsome Cl, rat	0.493	**0.610**	0.554	**0.650**
CYP2C8	0.416	**0.421**	0.384	**0.417**
CYP2C9	0.381	**0.393**	0.364	**0.396**
CYP2D6	0.202	**0.223**	**0.024**	**0.223**
CYP3A4	0.394	**0.397**	0.388	**0.419**
CYP,TDI,3A4,ratio	0.119	**0.116**	0.086	**0.123**
EPSA	0.812	**0.802**	0.776	**0.803**
halflife, dog	0.285	0.507	**0.716**	**0.741**
halflife, rat	0.198	0.414	**0.715**	**0.744**
hERG MK499	0.383	0.145	**0.358**	**0.368**
Fu,p, human	0.542	**0.589**	0.533	**0.581**
LogD	0.819	0.817	**0.819**	**0.825**
PAMPA	0.406	-0.447	**0.385**	**0.372**
PXR activation	0.372	-0.114	**-0.020**	**0.370**
Fu,p, rat	0.631	**0.651**	0.628	**0.682**
Fassif Solub	0.378	0.370	**0.422**	**0.455**
Vd, rat	0.559	**0.958**	0.855	**0.995**
MRT, dog	0.320	0.687	**0.828**	**0.909**
MRT, rat	0.000	**0.988**	0.623	**0.986**
SOLY7	0.584	0.603	**0.690**	**0.670**
PGP, rat	0.477	**0.468**	0.441	**0.470**

Table S5: **Mean absolute error (MAE)** of the base QSAR model Chemprop, MICE, Missforest and QComp on ADMET-780k dataset with compound-based temporal splitting. For each assay, the lowest MAE among different imputation methods is marked in **bold**. The second lowest MAE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.364	0.387	**0.379**	**0.370**
CaV 1.2	0.466	**0.475**	**0.450**	**0.450**
NaV 1.5	0.410	**0.402**	**0.395**	**0.395**
Cl, dog	0.716	**0.633**	0.682	**0.593**
Cl, rat	0.637	**0.170**	0.209	**0.050**
hepatocyte Cl, dog	0.531	**0.477**	0.516	**0.442**
microsome Cl, dog	0.622	**0.510**	0.612	**0.530**
hepatocyte Cl, human	0.624	**0.563**	0.620	**0.547**
microsome Cl, human	0.603	**0.558**	0.567	**0.522**
hepatocyte Cl, rat	0.695	**0.613**	0.703	**0.602**
microsome Cl, rat	0.593	**0.522**	0.553	**0.486**
CYP2C8	0.444	**0.441**	**0.442**	0.443
CYP2C9	0.462	**0.454**	0.457	**0.452**
CYP2D6	0.421	**0.432**	0.454	**0.411**
CYP3A4	0.329	0.339	**0.322**	**0.319**
CYP,TDI,3A4,ratio	0.651	**0.650**	0.661	**0.647**
EPSA	0.261	0.284	**0.276**	**0.265**
halflife, dog	0.737	0.585	**0.360**	**0.324**
halflife, rat	0.703	0.563	**0.363**	**0.349**
hERG MK499	0.466	0.511	**0.469**	**0.468**
Fu,p, human	0.383	**0.353**	0.377	**0.354**
LogD	0.245	0.254	**0.253**	**0.250**
PAMPA	0.547	0.976	**0.496**	**0.568**
PXR activation	0.514	0.643	**0.631**	**0.506**
Fu,p, rat	0.387	**0.375**	0.384	**0.357**
Fassif Solub	0.580	0.551	**0.510**	**0.517**
Vd, rat	0.506	**0.148**	0.222	**0.040**
MRT, dog	0.635	0.420	**0.288**	**0.229**
MRT, rat	0.708	**0.062**	0.375	**0.078**
SOLY7	0.423	0.409	**0.338**	**0.373**
PGP, rat	0.495	**0.498**	0.504	**0.496**

Table S6: **Mean square error (MSE)** of the base QSAR model Chemprop, MICE, Missforest and QComp on ADMET-780k dataset with compound-based temporal splitting. For each assay, the lowest RMSE among different imputation methods is marked in **bold**. The second lowest RMSE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.341	**0.358**	**0.358**	**0.340**
CaV 1.2	0.396	0.411	**0.408**	**0.380**
NaV 1.5	0.423	**0.417**	0.441	**0.415**
Cl, dog	0.895	**0.696**	0.811	**0.616**
Cl, rat	0.736	**0.058**	0.122	**0.008**
hepatocyte Cl, dog	0.463	**0.373**	0.471	**0.328**
microsome Cl, dog	0.585	**0.429**	0.596	**0.448**
hepatocyte Cl, human	0.647	**0.544**	0.669	**0.521**
microsome Cl, human	0.619	**0.535**	0.573	**0.472**
hepatocyte Cl, rat	0.763	**0.602**	0.789	**0.582**
microsome Cl, rat	0.580	**0.447**	0.510	**0.400**
CYP2C8	0.395	**0.392**	0.416	**0.395**
CYP2C9	0.520	**0.510**	0.534	**0.508**
CYP2D6	0.700	**0.682**	**0.856**	**0.682**
CYP3A4	0.341	**0.340**	0.345	**0.328**
CYP,TDI,3A4,ratio	0.848	**0.851**	0.880	**0.844**
EPSA	0.217	**0.229**	0.259	**0.228**
halflife, dog	0.928	0.640	**0.369**	**0.336**
halflife, rat	0.851	0.622	**0.302**	**0.272**
hERG MK499	0.383	0.531	**0.399**	**0.393**
Fu,p, human	0.254	**0.228**	0.259	**0.232**
LogD	0.168	0.170	**0.168**	**0.163**
PAMPA	0.561	1.366	**0.580**	**0.593**
PXR activation	0.497	0.881	**0.808**	**0.498**
Fu,p, rat	0.257	**0.243**	0.259	**0.221**
Fassif Solub	0.771	0.781	**0.716**	**0.675**
Vd, rat	0.461	**0.044**	0.151	**0.006**
MRT, dog	0.678	0.312	**0.171**	**0.091**
MRT, rat	0.909	**0.011**	0.342	**0.013**
SOLY7	0.375	0.357	**0.279**	**0.297**
PGP, rat	0.418	**0.425**	0.447	**0.424**

E.2 Benmarking QComp on public ADMET dataset

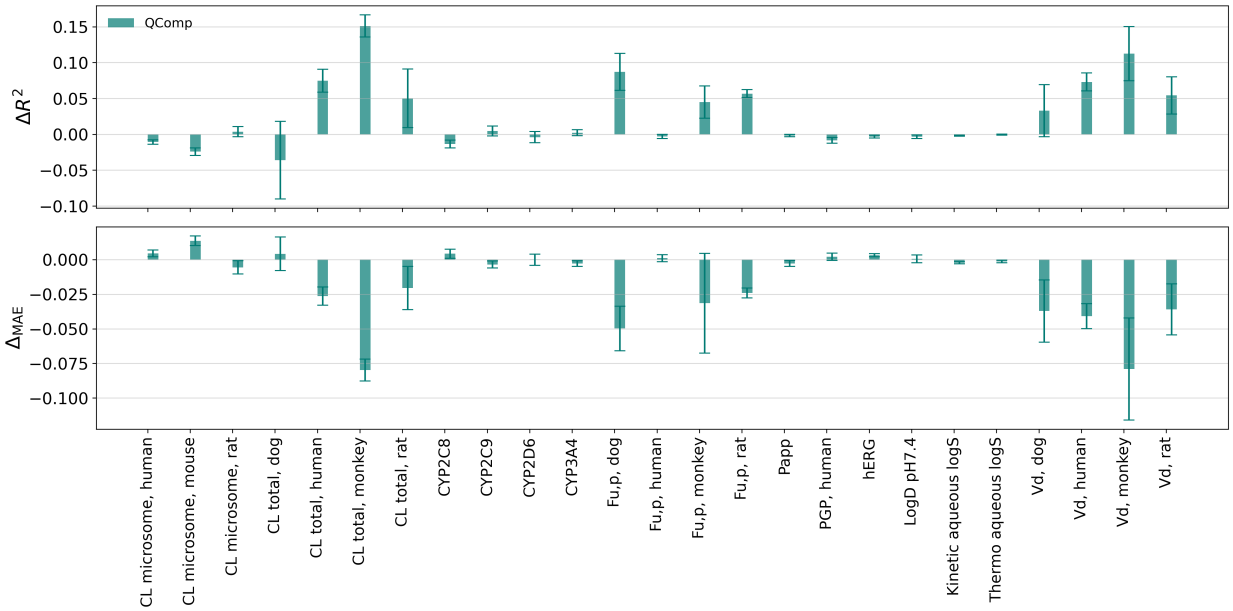


Figure S4: The average and the error bar of the change on R^2 score (a positive change means improvement) and MAE (a negative change means improvement) over the base QSAR model, for in vitro assays. The average and the error bar are calculated over the five QComp models trained with 5-fold random split.

E.3 Imputing in vivo assays with in vitro data on ADMET-780k dataset

Table S7: R^2 scores of the base QSAR model Chemprop, MICE, Missforest, and QComp on ADMET-780k dataset with compound-based temporal splitting. Only in vitro data are utilized for imputation. For each assay, the highest R^2 score among different imputation methods is marked in **bold**. The second highest R^2 score in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Cl, dog	0.169	0.172	**0.177**	**0.204**
Cl, rat	0.291	0.144	**0.315**	**0.357**
halflife, dog	0.285	0.232	**0.299**	**0.306**
halflife, rat	0.198	**0.185**	0.178	**0.207**
Vd, rat	0.559	0.311	**0.520**	**0.586**
MRT, dog	0.320	0.190	**0.327**	**0.366**
MRT, rat	0.000	**0.140**	-0.031	**0.002**

Table S8: MAE of the base QSAR model Chemprop, MICE, Missforest, and QComp on ADMET-780k dataset with compound-based temporal splitting. Only in vitro data are utilized for imputation. For each assay, the lowest MAE among different imputation methods is marked in **bold**. The second lowest MAE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Cl, dog	0.716	**0.717**	0.718	**0.700**
Cl, rat	0.637	0.704	**0.631**	**0.609**
halflife, dog	0.737	0.776	**0.725**	**0.728**
halflife, rat	0.703	0.715	**0.712**	**0.698**
Vd, rat	0.506	0.660	**0.529**	**0.488**
MRT, dog	0.635	0.709	**0.625**	**0.611**
MRT, rat	0.708	**0.708**	0.718	**0.699**

Table S9: MSE of the base QSAR model Chemprop, MICE, Missforest, and QComp on ADMET-780k dataset with compound-based temporal splitting. Only in vitro data are utilized for imputation. For each assay, the lowest MSE among different imputation methods is marked in **bold**. The second lowest MSE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Cl, dog	0.895	0.892	**0.886**	**0.857**
Cl, rat	0.736	0.890	**0.712**	**0.668**
halflife, dog	0.928	0.997	**0.911**	**0.901**
halflife, rat	0.851	**0.865**	0.872	**0.842**
Vd, rat	0.461	0.720	**0.502**	**0.432**
MRT, dog	0.678	0.808	**0.671**	**0.632**
MRT, rat	0.909	**0.781**	0.937	**0.906**

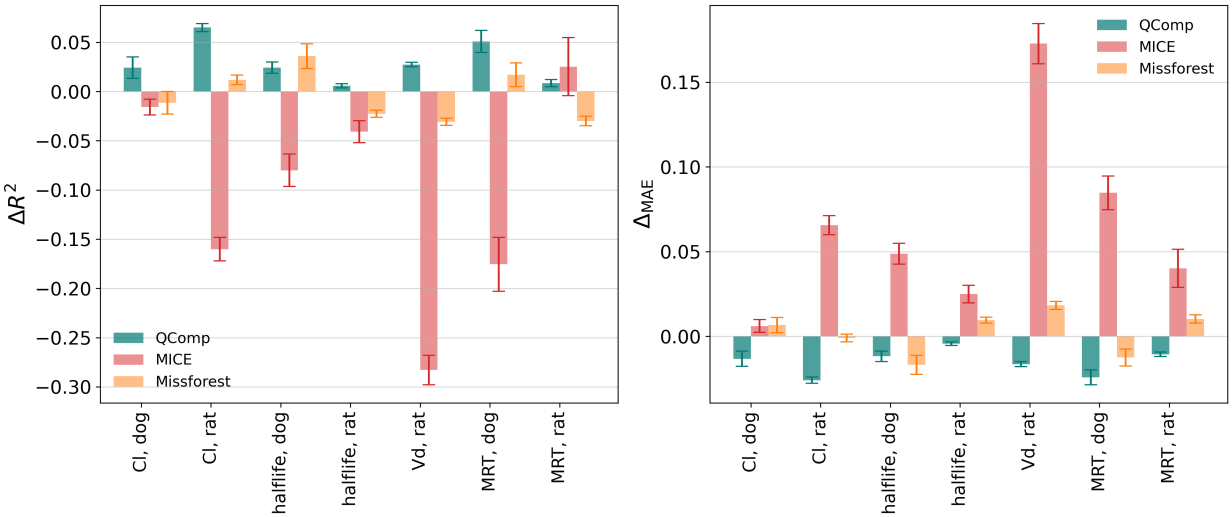


Figure S5: The average and the error bar of the change in R^2 score (a positive change means improvement) and MAE (a negative change means improvement) over the base QSAR model, for in vivo assays. The average and the error bar are calculated from the 50-bin splitting of the test set.

E.4 Imputing in vitro assays without in vivo data on ADMET-780k dataset

Table S10: R^2 scores of the base QSAR model Chemprop, MICE, Missforest, and QComp on ADMET-780k dataset with compound-based temporal splitting. Only in vitro data are utilized for imputation. For each assay, the highest R^2 score among different imputation methods is marked in **bold**. The second highest R^2 score in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.646	**0.637**	0.629	**0.646**
CaV 1.2	-0.221	**-0.243**	-0.252	**-0.171**
NaV 1.5	0.322	**0.336**	0.295	**0.337**
hepatocyte Cl, dog	0.320	**0.385**	0.296	**0.436**
microsome Cl, dog	0.401	**0.514**	0.383	**0.513**
hepatocyte Cl, human	0.408	**0.505**	0.391	**0.526**
microsome Cl, human	0.471	**0.546**	0.510	**0.598**
hepatocyte Cl, rat	0.322	**0.426**	0.298	**0.420**
microsome Cl, rat	0.493	**0.612**	0.554	**0.635**
CYP2C8	0.416	**0.424**	0.387	**0.418**
CYP2C9	0.381	**0.396**	**0.365**	**0.396**
CYP2D6	0.202	**0.228**	0.028	**0.224**
CYP3A4	0.394	**0.416**	0.388	**0.419**
CYP,TDI,3A4,ratio	0.119	**0.119**	0.089	**0.123**
EPSA	0.812	**0.808**	0.776	**0.804**
hERG MK499	0.383	0.333	**0.358**	**0.367**
Fu,p, human	0.542	**0.590**	0.538	**0.580**
LogD	0.819	0.817	**0.819**	**0.825**
PAMPA	0.406	-0.158	**0.384**	**0.373**
PXR activation	0.372	**0.218**	-0.018	**0.370**
Fu,p, rat	0.631	**0.634**	0.612	**0.637**
Fassif Solub	0.378	**0.437**	0.423	**0.455**
SOLY7	0.584	0.616	**0.690**	**0.671**
PGP, rat	0.477	**0.470**	**0.442**	**0.470**

Table S11: MAE of the base QSAR model Chemprop, MICE, Missforest, and QComp on ADMET-780k dataset with compound-based temporal splitting. Only in vitro data are utilized for imputation. For each assay, the lowest MAE among different imputation methods is marked in **bold**. The second lowest MAE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.364	**0.378**	0.379	**0.369**
CaV 1.2	0.466	0.468	**0.449**	**0.450**
NaV 1.5	0.410	0.396	**0.395**	**0.393**
hepatocyte Cl, dog	0.531	**0.508**	0.516	**0.479**
microsome Cl, dog	0.622	**0.540**	0.614	**0.545**
hepatocyte Cl, human	0.624	**0.561**	0.618	**0.545**
microsome Cl, human	0.603	**0.555**	0.567	**0.521**
hepatocyte Cl, rat	0.695	**0.636**	0.703	**0.637**
microsome Cl, rat	0.593	**0.520**	0.553	**0.499**
CYP2C8	0.444	**0.438**	**0.441**	0.442
CYP2C9	0.462	**0.451**	0.457	**0.452**
CYP2D6	0.421	**0.427**	0.453	**0.412**
CYP3A4	0.329	0.328	**0.322**	**0.319**
CYP,TDI,3A4,ratio	0.651	**0.649**	0.660	**0.647**
EPSA	0.261	0.278	**0.276**	**0.264**
hERG MK499	0.466	0.472	**0.469**	**0.468**
Fu,p, human	0.383	**0.352**	0.376	**0.355**
LogD	0.245	0.254	**0.253**	**0.250**
PAMPA	0.547	0.964	**0.496**	**0.568**
PXR activation	0.514	**0.570**	0.630	**0.506**
Fu,p, rat	0.387	**0.384**	0.392	**0.381**
Fassif Solub	0.580	0.534	**0.510**	**0.517**
SOLY7	0.423	0.404	**0.338**	**0.372**
PGP, rat	0.495	**0.497**	0.503	**0.496**

Table S12: MSE of the base QSAR model Chemprop, MICE, Missforest, and QComp on ADMET-780k dataset with compound-based temporal splitting. Only in vitro data are utilized for imputation. For each assay, the lowest MSE among different imputation methods is marked in **bold**. The second lowest MSE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.341	**0.350**	0.357	**0.340**
CaV 1.2	0.396	**0.404**	0.407	**0.380**
NaV 1.5	0.423	**0.414**	**0.440**	**0.414**
hepatocyte Cl, dog	0.463	**0.418**	0.479	**0.383**
microsome Cl, dog	0.585	**0.475**	0.603	**0.476**
hepatocyte Cl, human	0.647	**0.541**	0.665	**0.518**
microsome Cl, human	0.619	**0.531**	0.574	**0.471**
hepatocyte Cl, rat	0.763	**0.646**	0.789	**0.652**
microsome Cl, rat	0.580	**0.444**	0.511	**0.418**
CYP2C8	0.395	**0.389**	0.415	**0.394**
CYP2C9	0.520	**0.507**	**0.534**	**0.507**
CYP2D6	0.700	**0.678**	0.853	**0.681**
CYP3A4	0.341	**0.329**	0.345	**0.327**
CYP,TDI,3A4,ratio	0.848	**0.848**	0.877	**0.844**
EPSA	0.217	**0.222**	0.259	**0.226**
hERG MK499	0.383	0.414	**0.398**	**0.393**
Fu,p, human	0.254	**0.227**	0.256	**0.233**
LogD	0.168	0.170	**0.168**	**0.163**
PAMPA	0.561	1.093	**0.581**	**0.592**
PXR activation	0.497	**0.619**	0.805	**0.498**
Fu,p, rat	0.257	**0.255**	0.270	**0.252**
Fassif Solub	0.771	**0.698**	0.715	**0.675**
SOLY7	0.375	0.346	**0.279**	**0.296**
PGP, rat	0.418	**0.423**	**0.446**	**0.423**

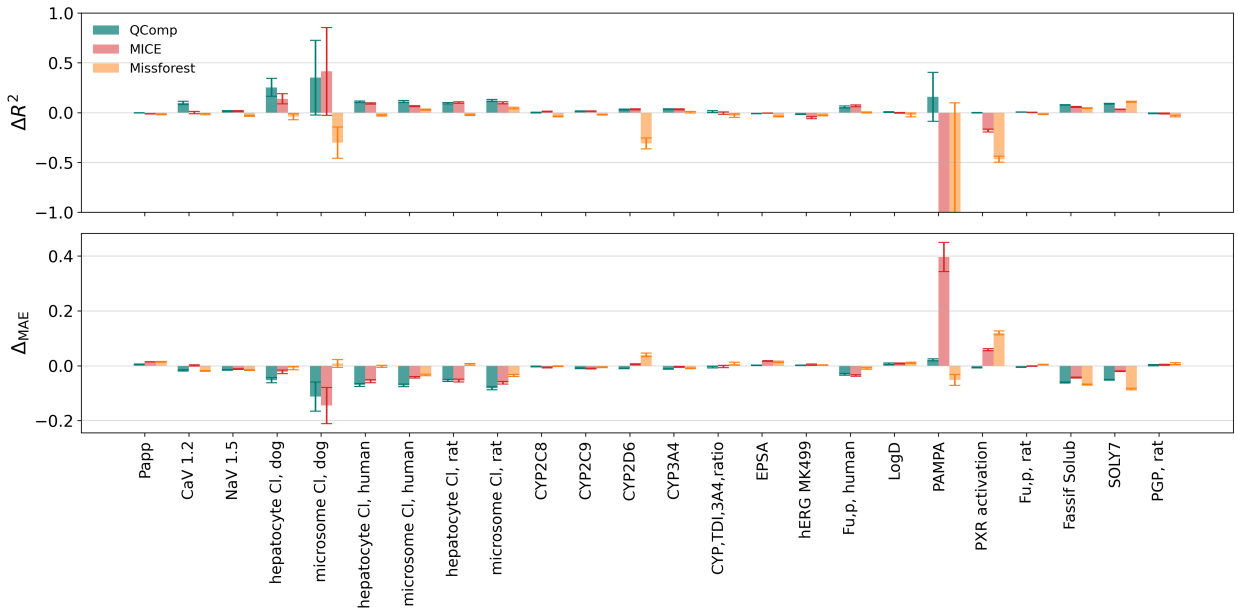


Figure S6: The average and the error bar of the change on R^2 score (a positive change means improvement) and MAE (a negative change means improvement) over the base QSAR model, for in vitro assays. The average and the error bar are calculated from the 50-bin splitting of the test set.

F Results of Clustering-based Split

In this section, we present the experiments using datasets with clustering-based split. We applied k-means clustering using Morgan fingerprints to partition the compounds into five clusters. For the internal ADMET-780k dataset, we selected one cluster as the test set, with the split details summarized in Table S13. For the public ADMET dataset, we performed five different train-test splits based on the clusters.

Table S13: ADMET-780k dataset: Number of compounds in training and test sets for clustering-based split

Assay	Train Size	Test Size	Test Size[%]
Papp	41092	13954	25.35
CaV 1.2	111810	41929	27.27
NaV 1.5	106747	40696	27.60
Cl, dog	14202	7634	34.96
Cl, rat	57555	18345	24.17
hepatocyte Cl, dog	5872	2504	29.89
microsome Cl, dog	3618	795	18.01
hepatocyte Cl, human	31328	10865	25.75
microsome Cl, human	38414	10281	21.11
hepatocyte Cl, rat	29165	10010	25.55
microsome Cl, rat	35718	10131	22.10
CYP2C8	50586	19753	28.08
CYP2C9	172650	50227	22.54
CYP2D6	171722	48786	22.12
CYP3A4	173944	49778	22.25
CYP,TDI,3A4,ratio	30439	9886	24.52
EPSA	33784	7016	17.20
halflife, dog	15721	8165	34.18
halflife, rat	62483	20405	24.62
hERG MK499	286177	78042	21.43
Fu,p, human	15708	6545	29.41
LogD	366587	93114	20.26
PAMPA	2480	1521	38.02
PXR activation	179565	54675	23.34
Fu,p, rat	41169	13294	24.41
kinetic FaSSIF solub	252622	63575	20.11
Vd, rat	57580	18341	24.16
MRT, dog	12559	7143	36.26
MRT, rat	54301	17704	24.59
SOLY7	376008	95502	20.25
PGP, rat	21281	6350	22.98

F.1 Benchmarking results on ADMET-780k dataset

The training and imputing processes follow the same protocol stated in the main text. The results are reported using four evaluation metrics: Pearson r^2 score, coefficient of determination R^2, MAE, and MSE.

Table S14: Pearson r^2 scores of the base QSAR model Chemprop, MICE, Missforest and QComp on ADMET-780k dataset with clustering-based splitting. For each assay, the highest r^2 score among different imputation methods is marked in **bold**. The second highest r^2 score in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.849	**0.849**	0.845	**0.851**
CaV 1.2	0.654	**0.657**	0.646	**0.670**
NaV 1.5	0.573	**0.586**	0.557	**0.588**
Cl, dog	0.350	**0.513**	0.416	**0.578**
Cl, rat	0.438	**0.959**	0.900	**0.982**
hepatocyte Cl, dog	0.663	**0.694**	0.637	**0.711**
microsome Cl, dog	0.614	**0.683**	0.590	**0.702**
hepatocyte Cl, human	0.546	**0.597**	0.542	**0.619**
microsome Cl, human	0.683	**0.736**	0.700	**0.765**
hepatocyte Cl, rat	0.555	**0.631**	0.554	**0.647**
microsome Cl, rat	0.651	**0.724**	0.665	**0.744**
CYP2C8	0.567	**0.579**	0.557	**0.580**
CYP2C9	0.612	**0.618**	0.605	**0.623**
CYP2D6	0.513	**0.521**	0.473	**0.523**
CYP3A4	0.695	0.642	**0.694**	**0.703**
CYP,TDI,3A4,ratio	0.634	**0.636**	0.632	**0.637**
EPSA	0.799	0.772	**0.786**	**0.798**
halflife, dog	0.451	0.640	**0.765**	**0.803**
halflife, rat	0.382	0.601	**0.715**	**0.755**
hERG MK499	0.630	**0.627**	0.623	**0.635**
Fu,p, human	0.676	**0.716**	0.674	**0.724**
LogD	0.904	**0.907**	**0.904**	**0.907**
PAMPA	0.477	0.103	**0.408**	**0.485**
PXR activation	0.583	0.575	**0.576**	**0.587**
Fu,p, rat	0.718	**0.738**	0.726	**0.772**
Fassif Solub	0.416	**0.481**	0.459	**0.492**
Vd, rat	0.551	**0.953**	0.909	**0.981**
MRT, dog	0.510	**0.909**	0.851	**0.917**
MRT, rat	0.500	**0.982**	0.846	**0.987**
SOLY7	0.687	0.701	**0.719**	**0.729**
PGP, rat	0.768	**0.770**	0.765	**0.771**

Table S15: R^2 scores of the base QSAR model Chemprop, MICE, Missforest and QComp on ADMET-780k dataset with clustering-based splitting. For each assay, the highest R^2 score among different imputation methods is marked in **bold**. The second highest R^2 score in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.849	**0.849**	0.845	**0.851**
CaV 1.2	0.648	**0.655**	0.644	**0.664**
NaV 1.5	0.573	**0.584**	0.549	**0.587**
Cl, dog	0.348	**0.512**	0.409	**0.568**
Cl, rat	0.427	**0.956**	0.891	**0.982**
hepatocyte Cl, dog	0.662	**0.694**	0.631	**0.709**
microsome Cl, dog	0.600	**0.679**	0.576	**0.690**
hepatocyte Cl, human	0.544	**0.597**	0.533	**0.619**
microsome Cl, human	0.681	**0.735**	0.695	**0.763**
hepatocyte Cl, rat	0.551	**0.629**	0.546	**0.646**
microsome Cl, rat	0.649	**0.721**	0.660	**0.741**
CYP2C8	0.567	**0.579**	**0.553**	**0.579**
CYP2C9	0.610	**0.605**	0.595	**0.621**
CYP2D6	0.512	**0.518**	0.454	**0.522**
CYP3A4	0.693	0.634	**0.694**	**0.702**
CYP,TDI,3A4,ratio	0.621	**0.630**	**0.625**	0.622
EPSA	0.799	0.771	**0.786**	**0.798**
halflife, dog	0.429	0.625	**0.759**	**0.798**
halflife, rat	0.361	0.599	**0.712**	**0.755**
hERG MK499	0.630	**0.624**	0.621	**0.635**
Fu,p, human	0.675	**0.714**	0.669	**0.724**
LogD	0.903	**0.907**	0.904	**0.906**
PAMPA	0.389	-1.167	**0.322**	**0.392**
PXR activation	0.580	0.571	**0.572**	**0.584**
Fu,p, rat	0.717	**0.737**	0.724	**0.771**
Fassif Solub	0.413	**0.474**	0.447	**0.491**
Vd, rat	0.550	**0.951**	0.907	**0.981**
MRT, dog	0.498	**0.906**	0.849	**0.913**
MRT, rat	0.485	**0.982**	0.840	**0.986**
SOLY7	0.686	0.698	**0.715**	**0.728**
PGP, rat	0.766	**0.769**	0.764	**0.768**

Table S16: MAE of the base QSAR model Chemprop, MICE, Missforest and QComp on ADMET-780k dataset with clustering-based splitting. For each assay, the lowest MAE among different imputation methods is marked in **bold**. The second lowest MAE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.281	**0.281**	0.284	**0.280**
CaV 1.2	0.465	**0.450**	0.456	**0.454**
NaV 1.5	0.410	0.407	**0.398**	**0.406**
Cl, dog	0.589	**0.521**	0.567	**0.487**
Cl, rat	0.547	**0.133**	0.191	**0.046**
hepatocyte Cl, dog	0.391	**0.367**	0.396	**0.358**
microsome Cl, dog	0.471	**0.412**	0.440	**0.409**
hepatocyte Cl, human	0.475	**0.447**	0.472	**0.433**
microsome Cl, human	0.425	**0.386**	0.398	**0.368**
hepatocyte Cl, rat	0.494	**0.451**	0.491	**0.439**
microsome Cl, rat	0.447	**0.395**	0.427	**0.383**
CYP2C8	0.465	**0.455**	0.461	**0.458**
CYP2C9	0.460	0.454	**0.451**	**0.453**
CYP2D6	0.355	**0.351**	0.357	**0.352**
CYP3A4	0.349	0.353	**0.331**	**0.344**
CYP,TDI,3A4,ratio	0.601	**0.591**	**0.577**	0.599
EPSA	0.157	0.174	**0.161**	**0.158**
halflife, dog	0.571	0.445	**0.311**	**0.276**
halflife, rat	0.566	0.435	**0.341**	**0.311**
hERG MK499	0.424	**0.422**	**0.425**	**0.422**
Fu,p, human	0.355	**0.328**	0.358	**0.326**
LogD	0.174	**0.171**	**0.173**	0.174
PAMPA	0.604	1.324	**0.624**	**0.604**
PXR activation	0.416	**0.415**	0.416	**0.414**
Fu,p, rat	0.366	**0.347**	0.361	**0.330**
Fassif Solub	0.424	**0.386**	0.387	**0.383**
Vd, rat	0.516	**0.141**	0.171	**0.048**
MRT, dog	0.545	**0.227**	0.271	**0.220**
MRT, rat	0.554	**0.053**	0.273	**0.050**
SOLY7	0.359	0.344	**0.324**	**0.338**
PGP, rat	0.377	**0.370**	**0.370**	**0.376**

Table S17: MSE of the base QSAR model Chemprop, MICE, Missforest and QComp on ADMET-780k dataset with clustering-based splitting. For each assay, the lowest MSE among different imputation methods is marked in **bold**. The second lowest MSE in imputation methods is marked in **bold and grey**.

Assay name	Chemprop	MICE	Missforest	QComp
Papp	0.193	**0.193**	0.198	**0.190**
CaV 1.2	0.465	**0.456**	0.470	**0.444**
NaV 1.5	0.450	**0.439**	0.475	**0.435**
Cl, dog	0.650	**0.487**	0.589	**0.431**
Cl, rat	0.603	**0.047**	0.115	**0.019**
hepatocyte Cl, dog	0.273	**0.247**	0.298	**0.235**
microsome Cl, dog	0.377	**0.303**	0.400	**0.292**
hepatocyte Cl, human	0.395	**0.349**	0.405	**0.330**
microsome Cl, human	0.335	**0.278**	0.321	**0.249**
hepatocyte Cl, rat	0.420	**0.347**	0.425	**0.332**
microsome Cl, rat	0.356	**0.284**	0.346	**0.263**
CYP2C8	0.451	**0.439**	**0.465**	**0.439**
CYP2C9	0.456	**0.462**	0.473	**0.443**
CYP2D6	0.460	**0.455**	0.514	**0.451**
CYP3A4	0.363	0.433	**0.363**	**0.353**
CYP,TDI,3A4,ratio	0.723	**0.707**	**0.715**	0.720
EPSA	0.071	0.081	**0.076**	**0.072**
halflife, dog	0.589	0.388	**0.249**	**0.209**
halflife, rat	0.593	0.372	**0.267**	**0.228**
hERG MK499	0.340	**0.345**	0.348	**0.336**
Fu,p, human	0.267	**0.236**	0.273	**0.227**
LogD	0.081	**0.078**	0.080	**0.079**
PAMPA	0.617	2.185	**0.684**	**0.613**
PXR activation	0.365	**0.372**	**0.372**	**0.361**
Fu,p, rat	0.265	**0.246**	0.259	**0.214**
Fassif Solub	0.472	**0.423**	0.444	**0.409**
Vd, rat	0.494	**0.054**	0.102	**0.021**
MRT, dog	0.521	**0.098**	0.157	**0.091**
MRT, rat	0.551	**0.020**	0.171	**0.015**
SOLY7	0.289	0.278	**0.262**	**0.250**
PGP, rat	0.243	**0.240**	0.245	**0.241**

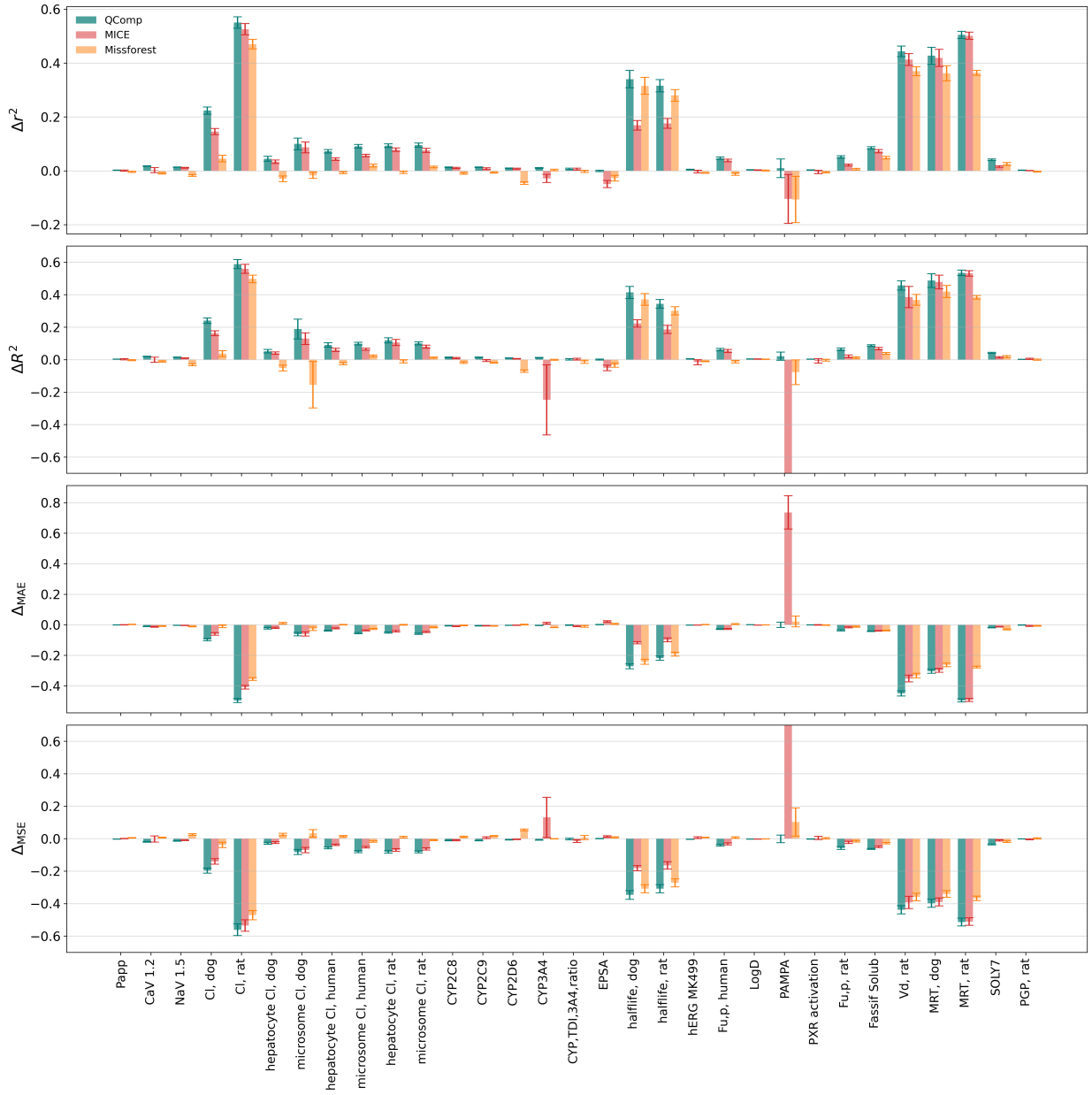


Figure S7: The average and the error bar of the change on Pearson r^2 score, R^2 score, MAE and MSE over the base QSAR model. The average and the error bar are calculated from the 100-bin splitting of the test set.

F.2 Benchmarking results on public ADMET dataset

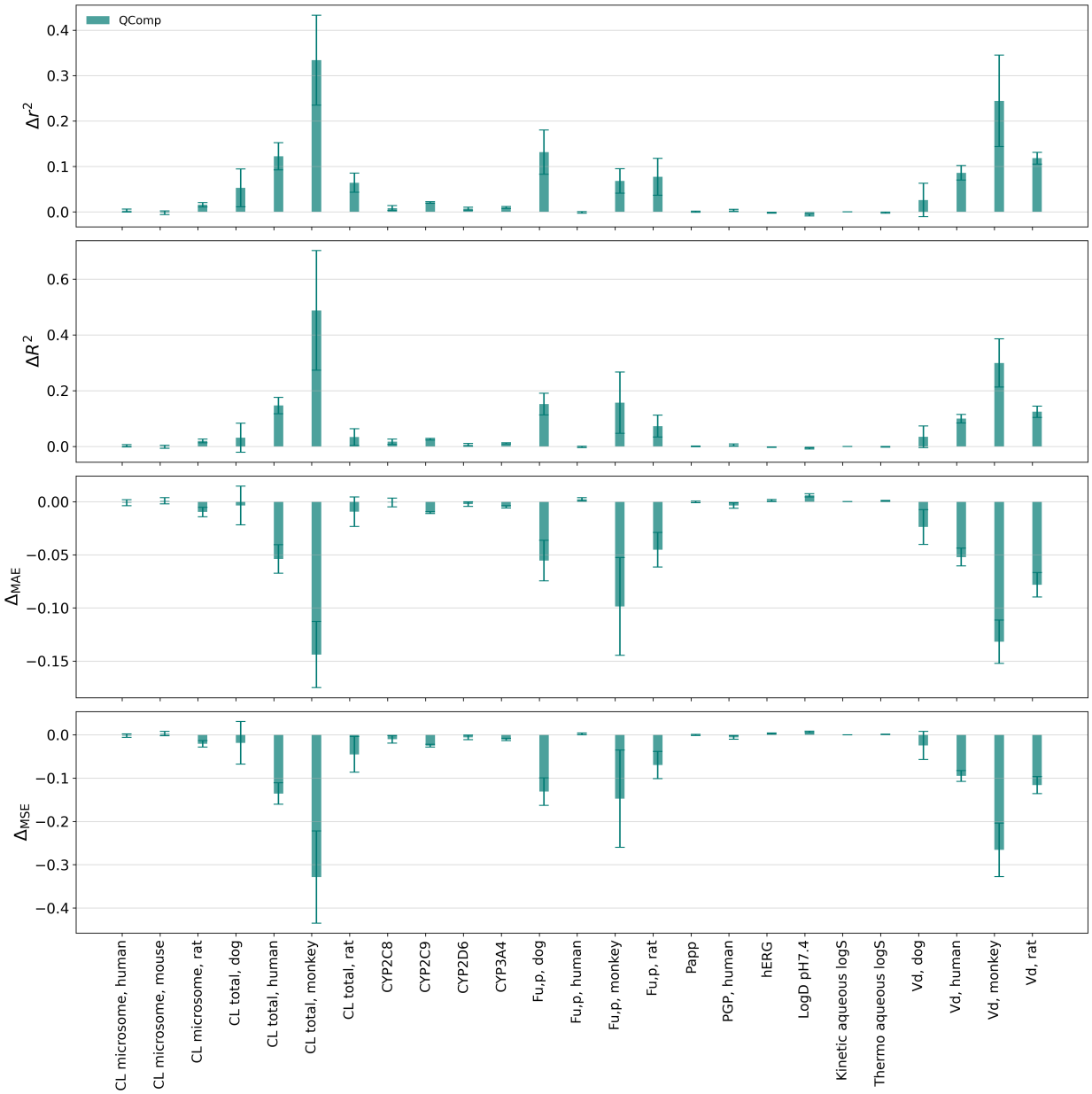


Figure S8: The average and the error bar of the change on Pearson r^2 score, R^2 score, MAE and MSE over the base QSAR model. The average and the error bar are calculated over the QComp models trained with 5 different train-test cluster splits.

F.3 Imputing in vivo assays with in vitro data on ADMET-780k dataset

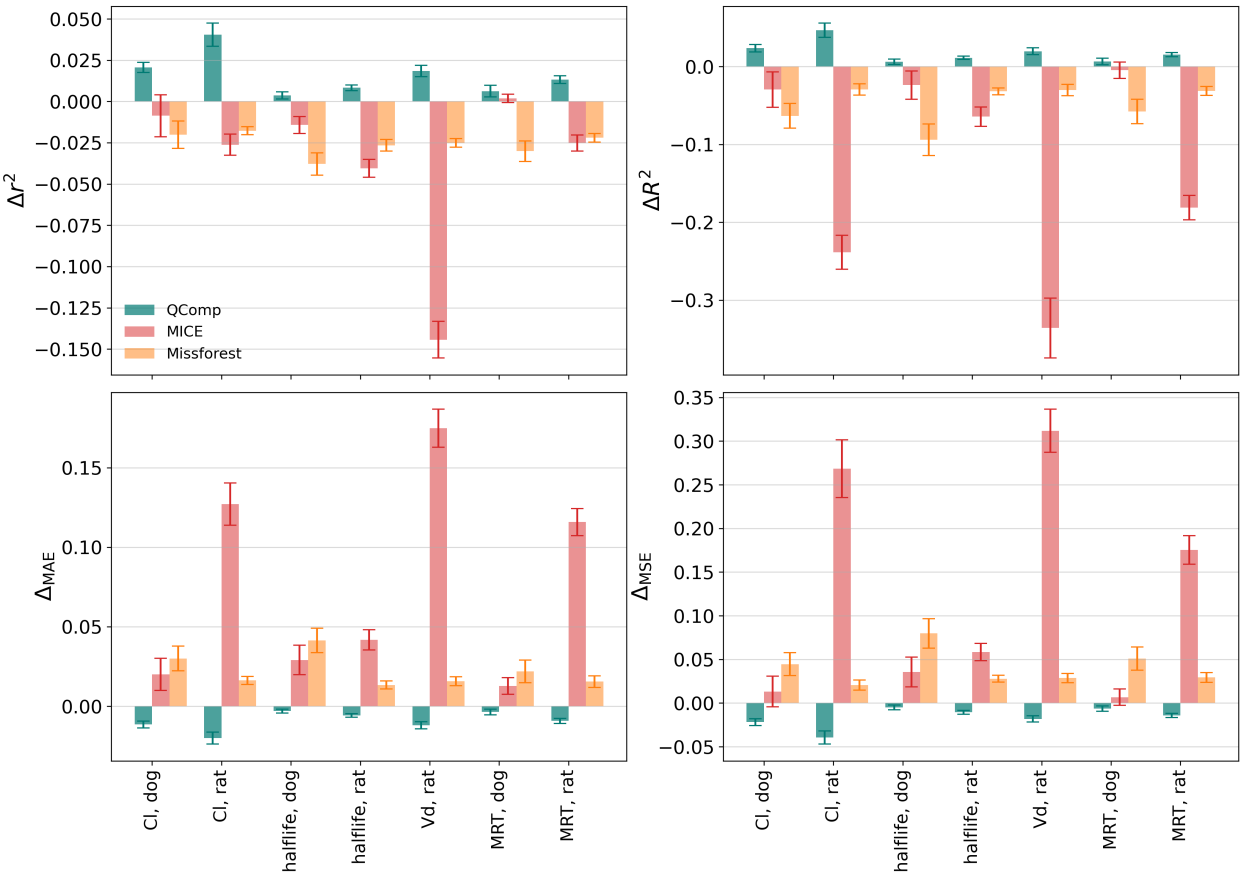


Figure S9: The average and the error bar of the change on Pearson r^2 score, R^2 score, MAE and MSE over the base QSAR model, for in vivo assays. The average and the error bar are calculated from the 100-bin splitting of the test set.

F.4 Imputing in vitro assays without in vivo data on ADMET-780k dataset

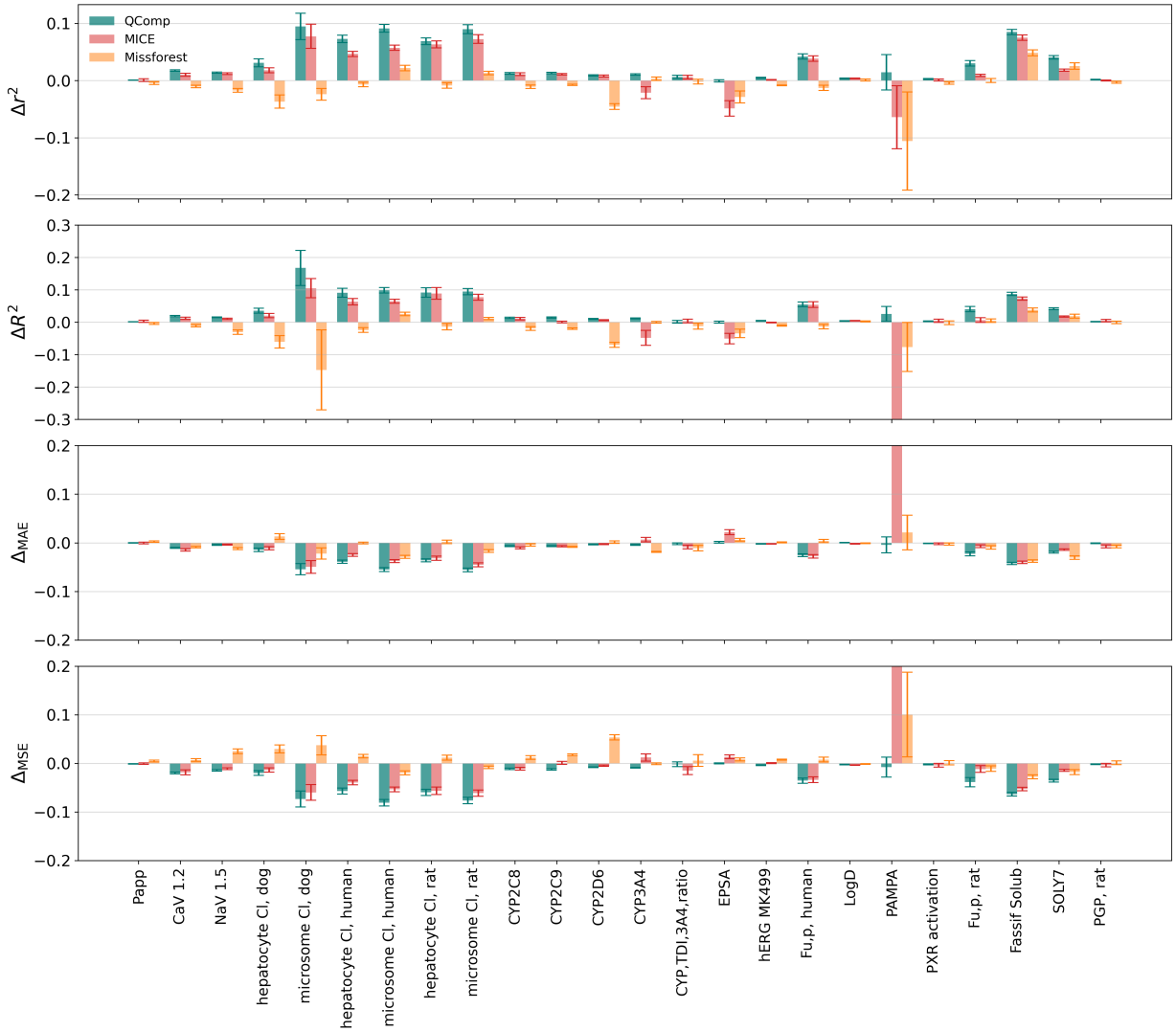


Figure S10: The average and the error bar of the change on Pearson r^2 score, R^2 score, MAE and MSE over the base QSAR model, for in vivo assays. The average and the error bar are calculated from the 100-bin splitting of the test set.

G Results on QM8 dataset

Here, we explore the broader applicability of QComp. We study the imputation of the quantum chemistry dataset QM8.[18] We split the QM8 dataset into an 80% training/validation subset and a 20% test subset using random split. As for the base QSAR model, we adopt a pre-trained, state-of-the-art Uni-Mol model.[19] We fine-tune the Uni-Mol model with the training set. And we train QComp using the same training set. The results from imputation is shown in Figure S11. The average MSE, MAE, Pearson r^2 and R^2 of the Uni-Mol model are 0.019, 0.009, 0.853 and 0.847, respectively. The average MSE, MAE, Pearson r^2 and R^2 achieved by QComp imputation are 0.013, 0.006, 0.923 and 0.922.

Note that the benchmarking results in the Uni-Mol paper were obtained using a scaffold-based split, while we use a random split for simplicity. Thus, the reported performance of Uni-Mol in our experiments is better than the published SOTA score.

Our results show success of QComp on tasks other than ADMET. Again, the success relies on the fact that there are correlations between chemical properties (see Figure S12), and on well-behaved base QSAR model (see Figure S13).

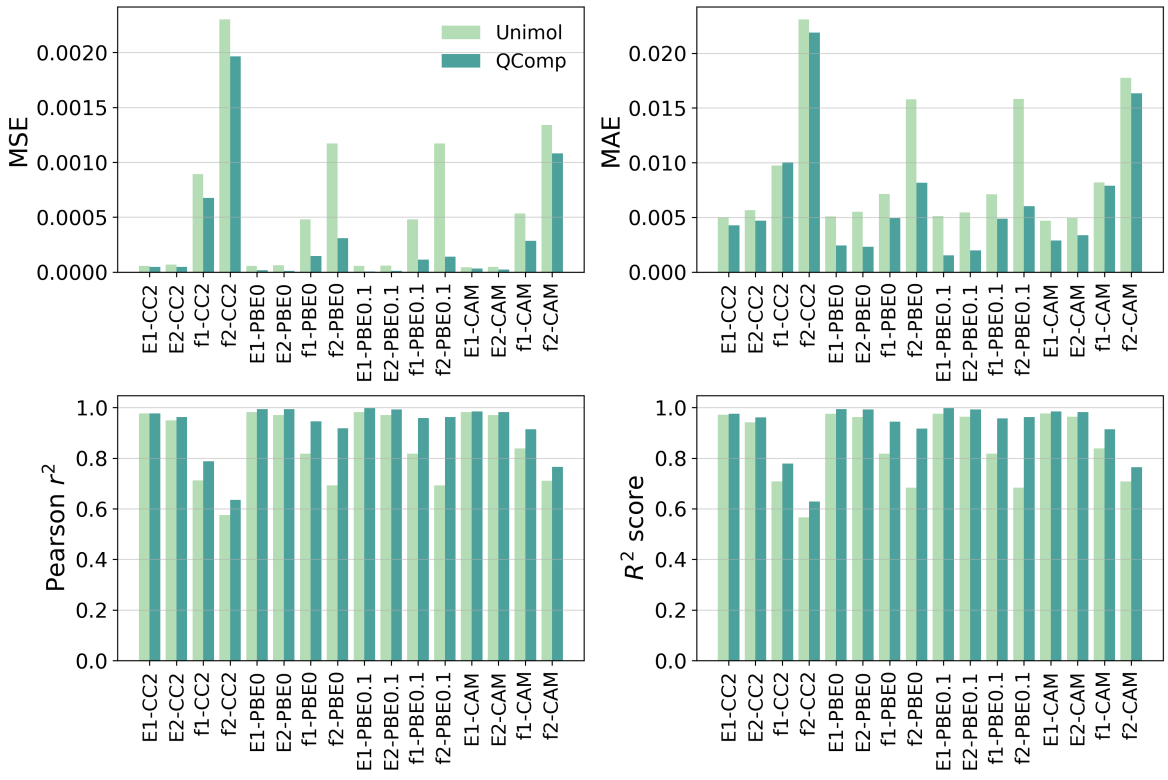


Figure S11: The MSE, MAE, Pearson r^2 and R^2 score of QComp and the base QSAR model (Uni-Mol) for 16 tasks in the QM8 dataset.

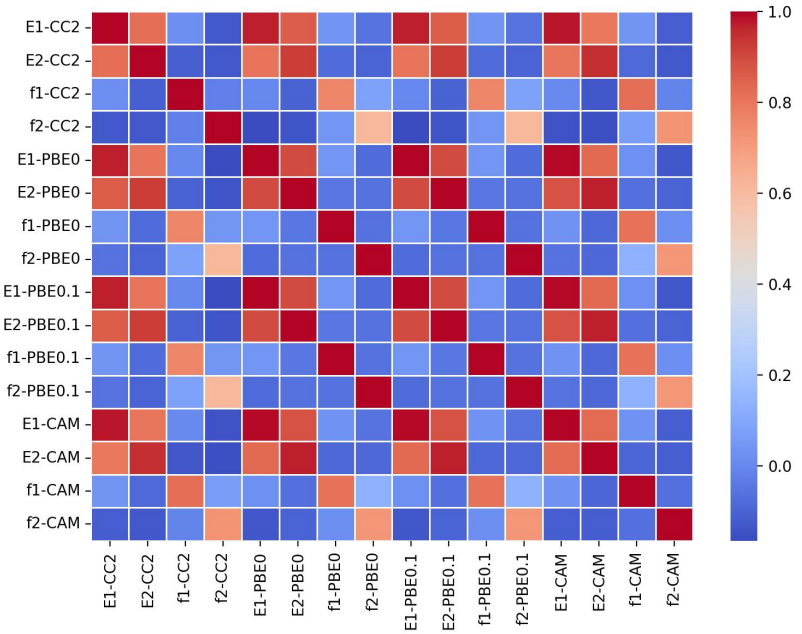


Figure S12: The QM8 dataset: Pearson correlation heatmap.

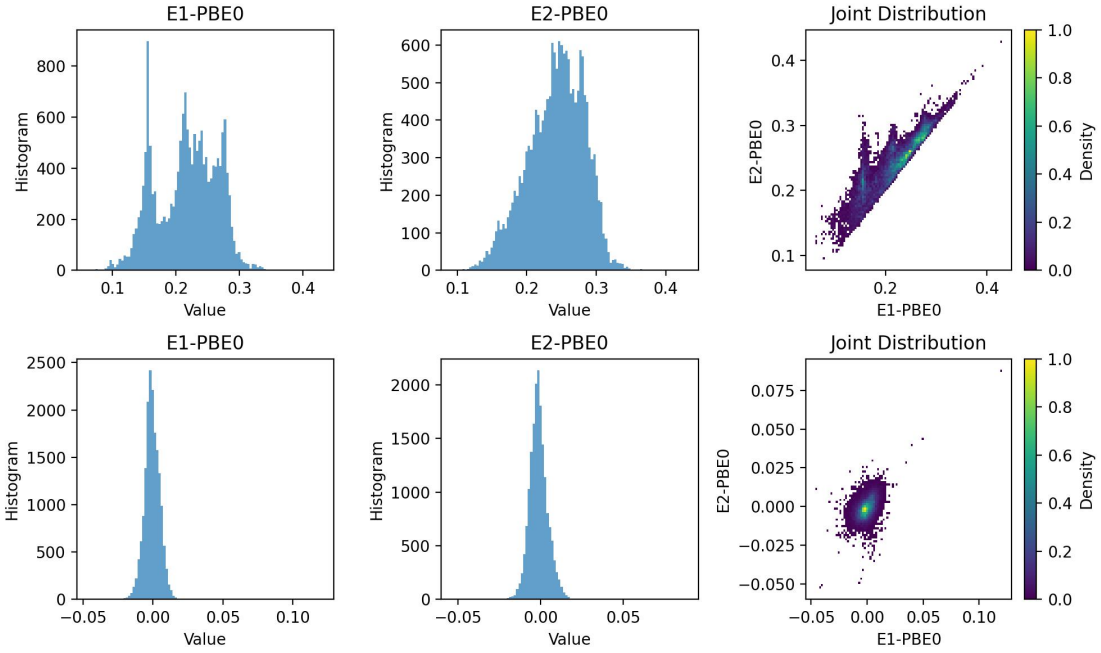


Figure S13: Upper panel: Histograms of "E1-PBE0" and "E2-PBE0"; the heatmap of the joint distribution of them. Lower panel: Histograms of the deviation of "E1-PBE0" and "E2-PBE0" properties from the base QSAR (Uni-Mol) predictions; the corresponding heatmap of the joint distribution.

References

(1) Feinberg, E. N.; Joshi, E.; Pande, V. S.; Cheng, A. C. Improvement in ADMET prediction with multitask deep featurization. *Journal of Medicinal Chemistry* **2020**, *63*, 8835–8848.

(2) Biswas, S.; Chung, Y.; Ramirez, J.; Wu, H.; Green, W. H. Predicting Critical Properties and Acentric Factors of Fluids Using Multitask Machine Learning. *Journal of Chemical Information and Modeling* **2023**, *63*, 4574–4588.

(3) Kingma, D. P.; Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* **2014**,

(4) Rubinsteyn, A.; Feldman, S. fancyimpute: An imputation library for python. *URL https://github. com/iskandr/fancyimpute* **2016**,

(5) Van Buuren, S.; Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software* **2011**, *45*, 1–67.

(6) Wenzel, J.; Matter, H.; Schmidt, F. Predictive Multitask Deep Neural Network Models for ADME-Tox Properties: Learning from Large Data Sets. *Journal of Chemical Information and Modeling* **2019**, *59*, 1253–1268.

(7) Iwata, H.; Matsuo, T.; Mamada, H.; Motomura, T.; Matsushita, M.; Fujiwara, T.; Maeda, K.; Handa, K. Predicting Total Drug Clearance and Volumes of Distribution Using the Machine Learning-Mediated Multimodal Method through the Imputation of Various Nonclinical Data. *Journal of Chemical Information and Modeling* **2022**, *62*, 4057–4065.

(8) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2023 update. *Nucleic Acids Research* **2023**, *51*, D1373–D1380.

(9) Watanabe, R.; Esaki, T.; Kawashima, H.; Natsume-Kitatani, Y.; Nagao, C.; Ohashi, R.; Mizuguchi, K. Predicting Fraction Unbound in Human Plasma from Chemical Structure: Improved Accuracy in the Low Value Ranges. *Molecular Pharmaceutics* **2018**, *15*, 5302–5311.

(10) Falcón-Cano, G.; Molina, C.; Cabrera-Pérez, M. Á. Reliable Prediction of Caco-2 Permeability by Supervised Recursive Machine Learning Approaches. *Pharmaceutics* **2022**, *14*.

(11) Esposito, C.; Wang, S.; Lange, U. E.; Oellien, F.; Riniker, S. Combining machine learning and molecular dynamics to predict P-glycoprotein substrates. *Journal of Chemical Information and Modeling* **2020**, *60*, 4730–4749.

(12) Braga, R. C.; Alves, V. M.; Silva, M. F.; Muratov, E.; Fourches, D.; Lião, L. M.; Tropsha, A.; Andrade, C. H. Pred-hERG: A Novel web-Accessible Computational Tool for Predicting Cardiac Toxicity. *Molecular Informatics* **2015**, *34*, 698–701.

(13) Aliagas, I.; Gobbi, A.; Lee, M. L.; Sellers, B. D. Comparison of logP and logD correction models trained with public and proprietary data sets. *Journal of Computer-Aided Molecular Design* **2022**, *36*, 253–262.

(14) Perryman, A. L.; Inoyama, D.; Patel, J. S.; Ekins, S.; Freundlich, J. S. Pruned Machine Learning Models to Predict Aqueous Solubility. *ACS Omega* **2020**, *5*, 16562–16567.

(15) Meng, J.; Chen, P.; Wahib, M.; Yang, M.; Zheng, L.; Wei, Y.; Feng, S.; Liu, W. Boosting the predictive performance with aqueous solubility dataset curation. *Scientific Data* **2022**, *9*.

(16) Vermeire, F. H.; Chung, Y.; Green, W. H. Predicting Solubility Limits of Organic Solutes for a Wide Range of Solvents and Temperatures. *Journal of the American Chemical Society* **2022**, *144*, 10785–10797.

(17) Landrum, G. RDKit: Open-Source Cheminformatics. 2006; `http://www.rdkit.org/`, (accessed November 29, 2023).

(18) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a benchmark for molecular machine learning. *Chemical science* **2018**, *9*, 513–530.

(19) Zhou, G.; Gao, Z.; Ding, Q.; Zheng, H.; Xu, H.; Wei, Z.; Zhang, L.; Ke, G. Uni-Mol: A Universal 3D Molecular Representation Learning Framework. The Eleventh International Conference on Learning Representations. 2023.