ACS OMEGA

Open Access
This article is licensed under CC-BY-NC-ND 4.0

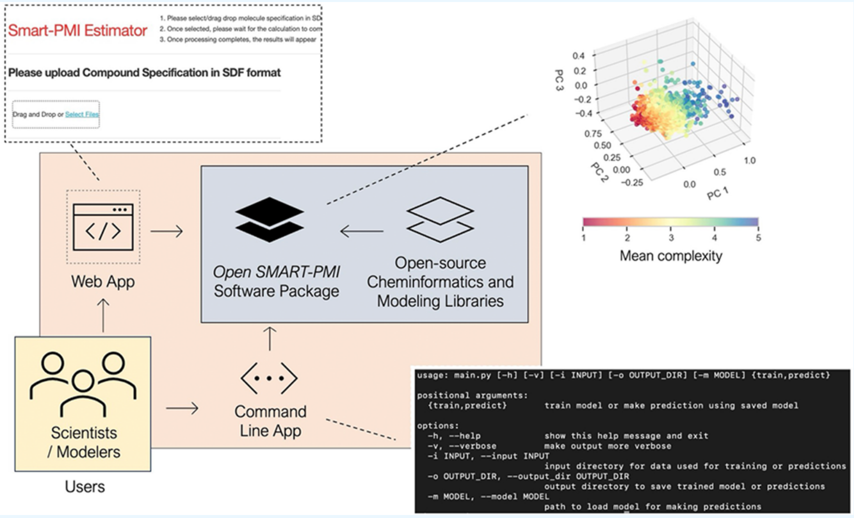
Evaluating Molecular Complexity with Open-Source Machine Learning Approaches to Predict Process Mass Intensity

Nicole Tin,* Mandeep Chauhan, Kennedy Agwamba, Yibai Sun, Astrid Parsons, Philippa Payne,# and Remus Osan*,#

Cite This: *ACS Omega* 2024, 9, 28476−28484

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information



Smart-PMI Estimator

1. Please select/drag drop molecule specification in SD
2. Once selected, please wait for the calculation to com
3. Once processing completes, the results will appear

Please upload Compound Specification in SDF format

Drag and Drop or Select Files

```
usage: main.py [-h] [-v] [-i INPUT] [-o OUTPUT_DIR] [-m MODEL] {train,predict}

positional arguments:
  {train,predict}       train model or make prediction using saved model

options:
  -h, --help            show this help message and exit
  -v, --verbose         make output more verbose
  -i INPUT, --input INPUT
                        input directory for data used for training or predictions
  -o OUTPUT_DIR, --output_dir OUTPUT_DIR
                        output directory to save trained model or predictions
  -m MODEL, --model MODEL
                        path to load model for making predictions
```

ABSTRACT: The application of green chemistry is critical for cultivating environmental responsibility and sustainable practices in pharmaceutical manufacturing. Process mass intensity (PMI) is a key metric that quantifies the resource efficiency of a manufacturing process, but determining what constitutes a successful PMI of a specific molecule is challenging. A recent approach correlated molecular features to a crowdsourced definition of molecular complexity to determine PMI targets. While recent machine learning tools show promise in predicting molecular complexity, a more extensive application could significantly optimize manufacturing processes. To this end, we refine and expand upon the SMART-PMI tool by Sheridan et al. to create an open-source model and application. Our solution emphasizes explainability and parsimony to facilitate a nuanced understanding of prediction and ensure informed decision-making. The resulting model uses four descriptors—the heteroatom count, stereocenter count, unique topological torsion, and connectivity index chi4n—to compute molecular complexity with a comparable 82.6% predictive accuracy and 0.349 RMSE. We develop a corresponding app that takes in structured data files (SDF) to rapidly quantify molecular complexity and provide a PMI target that can be used to drive process development activities. By integrating machine learning explainability and open-source accessibility, we provide flexible tools to advance the field of green chemistry and sustainable pharmaceutical manufacturing.

1. INTRODUCTION

1.1. Process Mass Intensity. Process mass intensity (PMI) is a key metric for measuring the efficiency of chemical processes, as benchmarked by the ACS Green Chemistry Institute Pharmaceutical Roundtable.[1] As drug candidates advance throughout the drug development lifecycle, PMI calculations are often used to capture progress toward more sustainable manufacturing and to drive further development.

PMI for a given synthesis is calculated as

$$\mathrm{PMI} = \frac{\text{total mass of raw materials used (kg)}}{\text{total product obtained (kg)}}$$

Received: March 12, 2024
Revised: June 3, 2024
Accepted: June 6, 2024
Published: June 24, 2024

© 2024 The Authors. Published by
American Chemical Society

https://doi.org/10.1021/acsomega.4c02427

Forecasting what constitutes a "successful" PMI for the manufacturing process for a given molecule is an unresolved industry challenge. In the absence of a "target" PMI, it is challenging for chemists to gauge if the resource efficiency of the process is on target, exceeding expectations, or in need of additional optimization. The comparison of the PMI of a process under development to an ambitious target PMI will help drive the incorporation of green chemistry and deliver sustainable manufacturing processes.

Despite not having a precise theoretical way to predict this feature, its estimates are extremely important in decision-making and process sustainability evaluation. Recently, it has been shown that a "target" PMI can be set for a given molecule through the molecular weight (m_w) and the calculation of its estimated molecular complexity (c).[2] This is given by

$$\text{SMART-PMI} = (0.13 \times m_w) + (177 \times c) - 252$$

The two variables, overall PMI and molecular complexity, share a positive linear correlation, where it is generally known that complex molecules have worse resource efficiency. Thus, the complexity of an active pharmaceutical ingredient (API) has far-reaching implications in terms of the challenges and efficiency of its manufacturing.

1.2. Molecular Complexity. Molecular complexity is commonly agreed to be an attribute intrinsic to a molecule that can affect chemical, material, and biological processes such as synthetic ease[3−8] and sustainability.[9,10] Generally the term refers to the intricacy and interactions in the internal structure of a molecule, however, the multidimensional nature of molecules represent a challenge to creating one summary metric of complexity.

Creating a definition of molecular complexity is a multidisciplinary problem, with applications and solutions coming from domains in math, physics, chemistry, and biology. While many solutions exist, none have been universally adopted.[11] Molecules are naturally viewed as graphs; approaches in graph theory analyze topology to develop graph invariants, such as the Zagreb or Weiner Indices, that can determine the structural properties of molecules.[12] Information theoretic measures, such as Shannon entropy, are also widely used in complexity science to capture diversity or degree of uniformity.[13−16]

In pharmaceutical chemistry, linear combinations of factors have been used as a proxy for complexity. Those factors usually include molecular weight and counts of attributes, such as stereocenters, rings, and functional groups.[3,5,10,17] Increases in computational power and advances in AI/ML have enabled researchers to develop data-driven approaches that can model these complex intuitions. A pivotal method by Sheridan et al. proposed one such model based on crowdsourced votes on a diverse set of molecules.[18] From the votes of 386 chemists, they found that the notion of complexity is independent of the chemistry subfield (i.e., process, analytical, medicinal, computational, etc.). A different approach focused on creating a model that is time dependent, asserting that complexity changes over time with respect to the available synthetic technology.[19]

There is a gap between the former theoretical and latter data-driven approaches to complexity. Theoretical models like those developed by Bertz, Bonchev, and Proudfoot use our inherent understanding of graphs or information theory to create explicit and transparent solutions.[13−15,20] The need for similar transparency in data-driven models is increasingly being recognized, with suggested standards for AI/ML in pharmaceutical development emerging recently from the FDA.[21] There is a need for data-driven algorithms to be secure, reliable, and interpretable; when used in decision-making, the model should be informative of the mechanisms that drive the nature of prediction.[22−24] Most powerful methods in machine learning are not interpretable, still some are more interpretable than others. Posthoc "explainable" wrapper methods have been developed to approximate the learned mechanisms under the black box. Reducing the input dimension space can also dramatically increase understanding while also improving learning.

1.3. Molecular Descriptors. Experimental and theoretical molecular descriptors have been used to capture information about a molecule related to its biological or physical properties.[25,26] Representing a molecule as a vector of descriptors, rather than as a graph or raw fingerprint, allows for a clear interpretation of the weighing schemes of contributing factors.

Zero-dimensional (0D) descriptors are easily observed features of a molecule derived from the chemical formula. They include molecular weight or number of heteroatoms and ring types.

One-dimensional (1D) descriptors encode attributes or substructures into a binary or hashed vector, known as a fingerprint. Unlike human fingerprints, however, this representation is not unique and can instead represent several molecules.

Two-dimensional (2D) descriptors use the molecular graph (topology) as a representation or as an input to a computation that yields a value. Some 2D descriptors, like the Weiner Index, also align with experimental measures.[27]

Three-dimensional (3D) descriptors capture geometric or topographical information. One such descriptor is the polar surface area (PSA) that uses the 3D molecular conformation to evaluate the surface area of polar atoms.[28,29] Since these are inherently sensitive to the conformation of the molecule, they are less common but are powerful measures containing higher information content.

Overall, using these descriptors to understand the factors that define molecular complexity can improve decision-making and understanding for goal-setting PMI.

1.4. Open SMART-PMI. In this paper, we evaluate the SMART-PMI and compound complexity tools developed by Sherer and Sheridan and adapt them for wider community use in support of green chemistry and process development acceleration.[9,18] The importance of making open-source tools to support computational green chemistry's framework cannot be overstated, as it benefits the broader community by publicizing previous advancements, encourages embracing uniform standards in the pharmaceutical industry, and facilitates the development of cost-efficient and flexible solutions.

Here, we relate machine learning principles to the retro-application of molecular complexity and make this knowledge open source alongside the source code. While we are not the first to model molecular complexity, we have initiated open-source collaboration and added interpretability to build upon the foundations of this important model. We show that the most successful algorithms, namely, random forest models, can succinctly model molecular complexity and can provide clear explainability to the user. The resulting models have different levels of complexity, with the simplest one using only four descriptors.

2. METHODS

2.1. Background. The code for SMART-PMI reported by Sheridan et al. is publicly available on their GitHub repository (https://github.com/Merck/compoundcomplexity/).[9] The crowdsourced information for compound complexity published by Sherer et al. includes 1775 nonproprietary drug-like molecules in the Supporting Information.[18] The votes on compound complexity from chemists range from 1 to 5, with 1 representing the simplest and 5 indicating a highly complex compound. The votes are summarized by the average score per molecule, denoted as mean complexity. The reported metrics for their random forest model had an 88% R^2 and RMSE of 0.27, while the standard deviation of votes agreement was 0.75. A four-term linear model was also shown as a simpler alternative, though with an 80% CV-R^2. As the random forest model is the focal point of both papers and has higher performance, we use it as the primary point of comparison. There were two primary complications in using this model. 186 of their 207 descriptors are sourced from the licensed Molecular Operating Environment (MOE) software, and the compound complexity package was written in several languages including Perl, which has waning influence in the field. The dataset and model were a pivotal step forward in understanding the makeup of molecular complexity.

2.2. Data Preparation. RDKit and Mordred are free and open-source cheminformatics libraries written in Python that enable easy calculation of many 0D−3D descriptors.[30,31] Inputs require the simplified molecular-input line-entry (SMILES) ASCII string representation, which can be transformed from a molecule's SDF file. RDKit and Mordred were used to generate a set of over 1500 initial descriptors. Throughout this work, the terms "descriptor" and "feature" are used synonymously.

Of the annotated 1775 molecules from Sheridan et al., 44 (2.5%) were unable to be sanitized by RDKit due to inconsistent valence electrons and were excluded from preprocessing and training. Of these molecules, 39/44 came from the MDL Drug Data Report (MDDR) database. This dataset was split into a 70:30 training and test set, and the model would be evaluated with 5-fold cross-validation. To reduce collinearity and improve model performance, features were pruned through several preprocessing steps. Mordred uses a specialized error object encoding; thus, any descriptor that comprised largely of the latter was removed. All of the above steps were performed on the entire dataset to capture the largest set of descriptors that would be unable to be performed on an incoming molecule. Then, to prevent leakage, where information from the hold-out dataset is used in training the model, the following preprocessing involving statistics of interest was only performed on the training set. The resulting statistics derived from the training set are later applied to transformations on the test set and future molecules. Each feature was scaled using min-max normalization. Normalization is a standard procedure for preparing ML inputs but is also crucial for comparing coefficients across models. Features with low information content (defined as variance) less than 10^{-5} were removed. Post-processing, the dataset consisted of 1196 descriptors.

2.3. Model Selection. Several architectures were tested to find the optimal algorithm for learning complexity (y) from a set of descriptors (X). In accordance with the original SMART-PMI paper, linear and ensemble methods were tested. Spatial methods and neural networks were additionally evaluated. Each model takes a different approach to learning the behavior of mean complexity, and each comes with trade-offs. While it is important to find a well-performing model, model simplicity is highly valued, as it is paramount to understanding the mechanisms driving prediction. A further discussion of the models and trade-offs can be found in the Supporting Information.

2.4. Feature Selection. To refine the model and reduce dimensionality, embedded and wrapper methods were used to select features. The methods used are complementary to each other and paint a picture of how features contribute to prediction. Toward the aim of understanding the modeling behavior and the contributing factors, we selected descriptors as chosen by high-performing models and their feature importance rankings.

Machine learning models typically optimize coefficients for each descriptor, which can be scrutinized to determine feature importance. Random forest measures a feature's mean decrease in impurity (MDI), which is an implicit method the model uses to evaluate its importance. Permutation importance calculates the decrease in model performance when a feature is randomized and is helpful as it can be computed across models.

Another method, SHAP (Shapley additive explanations), is widely employed to add interpretability to a model.[32,33] Features are analyzed as players within a game, where the features compete to influence prediction. SHAP is then able to allocate the contributions of each feature toward the predicted value for a particular instance. SHAP values can be aggregated across all instances to create average absolute scores for each feature. We can compare model coefficients and SHAP rankings to identify consensus important descriptors and eliminate features with little to no effect on modeled complexity.

2.5. Interaction Terms. Introducing interaction between terms is another way of modeling nonlinearity. Without interaction terms, independent variables could be interpreted as having a unique effect on the dependent variable, each with its own contribution. As we have a mix of physiochemical and topographical attributes, we would expect that the effect of one descriptor varies for different values of another descriptor. To maintain low computation times, we did not consider all possible interaction terms. Rather, we focus on product and quadratic terms that were generated for features filtered through selection.

2.6. Model Evaluation. It is expected that labeling molecules at the extreme ends of complexity would be the most difficult for the model, as the surveyed values of Sheridan et al.'s mean complexity follow a right-tailed normal distribution, and so there are the fewest ratings of 1, 4, 4.5, and 5 (Figure S1). Thus, negative root mean square error (denoted as −RMSE) was targeted as the primary metric when comparing models and methods, which gives sensitivity to large errors. We take the negative RMSE to standardize interpretation, where higher values indicate better fit, and the maximum theoretical value is 0. For comparing standard linear goodness-of-fit, the coefficient of determination R^2 is used. However, R^2 increases with the number of dimensions in the model, which is an important consideration in our modeling. Accordingly, adjusted R^2 was additionally considered as it corrects for the number of dimensions in the model, where R^2 increases accordingly. R^2 is given by

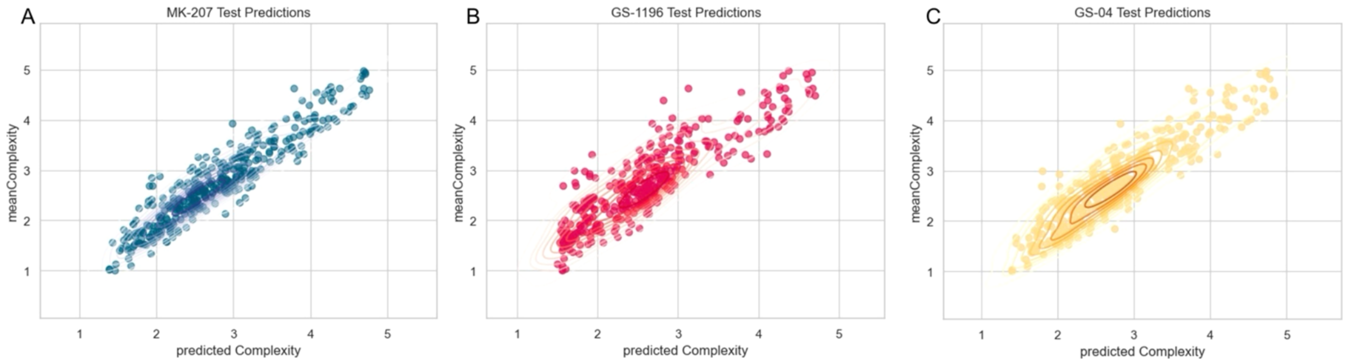


Figure 1. Test set predictions vs true mean complexity for the random forest model trained on each of the three datasets. (a) Our replication of the benchmark model and feature set by Sheridan et al. (MK-207), (b) the initial model using the full set of open-source descriptors (GS-1196), and (c) a parsimonious model using only four open-source descriptors (GS-04). The underlying contour plot represents the kernel density estimate, where the darkness of the rings increases with the density of points.

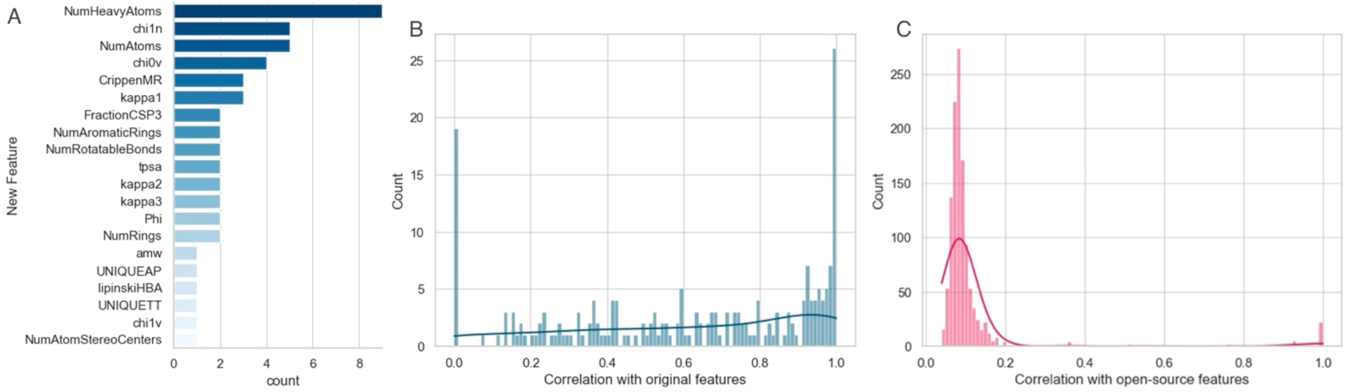


Figure 2. Coverage of feature sets. (a) Bar chart of open-source features that have over 94% correlation coefficient with an original descriptor. Several original descriptors can be summarized by an open-source feature. (b) Histogram of the original descriptors from Sheridan et al. Each bar measures the maximum correlation of an original descriptor with any open-source descriptor. (c) Histogram measuring the maximum correlation coefficients of an open-source descriptor to any original descriptor. The new set of open-source descriptors represents a larger search space that is largely independent of the original set.

$$R^2 = 1 - \frac{\text{MSE}}{\text{Var}(y) \times (N - 1)} = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y}_i)^2}$$

where $\hat{y}$ is the predicted value and adjusted R^2 is

$$\text{Adj-}R^2 = 1 - (1 - R^2) \times (n - 1)/(n - p - 1)$$

Fivefold cross-validation was further applied to each metric to generalize model performance.

3. RESULTS

In summary, we focus on the evaluation of the three models: our replication of the benchmark model and feature set by Sheridan et al. (MK-207); the initial model using the full set of open-source descriptors (GS-1196); a parsimonious model using only four open-source descriptors (GS-04) (Figure 1). The model's naming convention is described by the developer and the number of features used. Each of the three models is high performing, with an average squared prediction error that is lower than the chemists' vote variability (0.75). Still, each model comes with unique advantages and trade-offs, and each informs our understanding in different ways. The initial model is used to rank the full set of features specifically by their contribution to the predicted y variable, mean complexity. To improve accuracy and simplicity, additional interaction terms are generated for those top features and are used in evaluating a suite of models. The final parsimonious model has comparable error to the high-performing baseline and uses only four molecular descriptors.

The benchmark random forest model shared by Sheridan et al. was replicated using attributes generated from Merck's Compound Complexity repository and a temporary MOE software license. In a Python environment and hyperparameter tuning, we were only able to achieve cross-validated R^2 scores around 84%, as opposed to the 88% reported. While differences could be attributed to the former factors, hyperparameters such as the optimization function for node splitting, in conjunction with cross-validation, should be sufficient for reproducing results. Under the given environment, the accuracy of the model appears very similar, accounting for random deviation.

3.1. Feature Selection and Model Initialization. Once the new feature space was established and processed, we measured the correlation between descriptors in each of the previous and open-source spaces to evaluate similarities. Of the original 207 descriptors, 51 have a correlation coefficient of over 94% with a new open-source descriptor (Figure 2a). These include licensed descriptors deemed important in the work by Sheridan (Figure S2). Further, many of these highly correlated features map to the same open-source descriptors, where the 51 original descriptors can be represented by 20 open-source descriptors, as shown in Figure 2a. The larger set

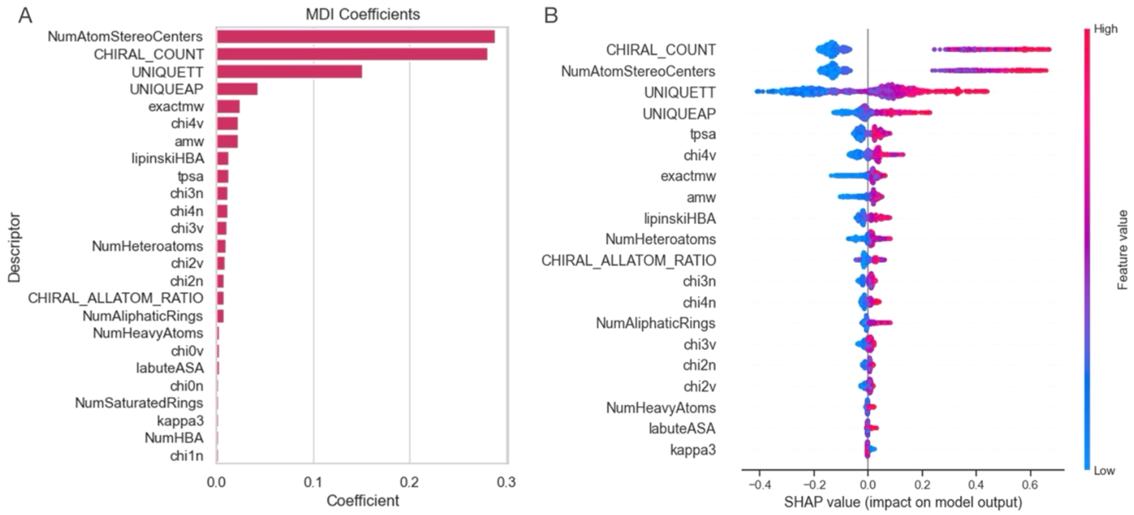


Figure 3. Ranking feature importance. (a) Mean decrease in impurity (MDI) across random forest trees. MDI is measured with the Gini index or the homogeneity of the resulting nodes and leaves created by a split at a specific feature. (b) Distribution of values contributing to SHAP feature importance values. The list is sorted by average absolute SHAP value, with the highest contributing feature at the top. The distribution of feature values is shown next to each, with colors showing the relationship between feature-instance value and the individual SHAP value.

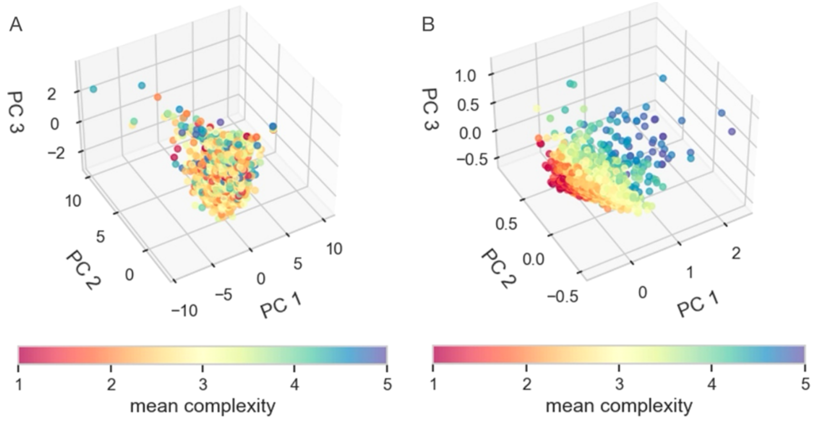


Figure 4. Principal component analysis. (a) Full training dataset for the left figure and (b) top 15 features in the right figure. The first three axes explain (a) 47% and (b) 90% of the total variance, respectively.

of open-source descriptors brings more variance and measurements, while also representing features important for the original model.

To select meaningful descriptors and initialize model tuning, we trained a random forest model tuned with grid search. For the training set of 1384 molecules and 1196 descriptors with 5-fold cross-validation, the R^2 stabilizes around 78% with an −RMSE = −0.39. Random forest significantly outperformed other models, despite high dimensionality and limited feature selection, which served as a useful baseline. SHAP values were calculated for each instance-descriptor value and aggregated to summarize each descriptor's average absolute score.

The SHAP summary plot illustrates a feature's potential to raise or lower the predicted complexity (Figure 3a). The chiral and stereocenter counts, which are similar descriptors, have a similar influence on the model. Low values (or counts) of these measures influence a low prediction, whereas very high counts increase the predicted value. The summary plot is especially helpful in allocating (the estimated) predictive power to each contributing feature's distribution. The most important descriptors calculated by random forests' implicit MDI are shown in Figure 3b. Few features meaningfully contribute to predicted complexity, and the same features stand out from both metrics of importance, indicating that the optimal model for molecular complexity would be relatively consistent. While permutation importance shows a different ranking, the high dimension of the input dataset obscures significant decreases in accuracy (Figure S3).

While it is important to note that these methods are inherently model-dependent, the results were similar across several iterations of the model initialization phase, and SHAP shows a direct relationship of a descriptor to molecular complexity. The highest ranked 13 of 17 features are in agreement (spearman ρ = 0.65). Furthermore, all descriptors selected in this stage show a correlation of above 50% to mean complexity (Figure S4a). Very few other features have a clear significant impact on the model, indicating that a parsimonious model is reasonable. These contributors, a combination of physiochemical, topological, and graph descriptors, align with the historical approaches taken by both mathematicians and domain experts. Since descriptor effects may be magnified in their interaction with another feature, the highest 15 SHAP-ranked features are selected for further modeling.

Table 1. Model Metrics for Each Algorithm Grouped by Training Set, as an Average of Fivefold Cross-Validation

		full feature set			top 15 features			+ interaction terms		
rank	model	R^2	Adj-R^2	−RMSE	R^2	Adj-R^2	−RMSE	R^2	Adj-R^2	−RMSE
1	random forest	77.8	−64.1	−0.39	84.6[a]	**84.4**	**−0.329**[a]	85.6[a]	84.2	−0.318[a]
2	multilayer perceptron	76.9	−70.8	−0.405	84.6	**84.4**	**−0.329**	84.5	83.0	−0.330
3	*k*-nearest neighbors	28.4	−4.29	−0.716	82.6	**82.5**	**−0.351**	82.9	81.2	−0.347
4	SVM	45.0	−3.06	−0.628	82.9	**82.7**	**−0.347**	71.4	68.7	−0.451
5	partial least squares	55.0	−2.32	−0.525	74.8	**74.5**	**−0.422**	59.6	55.7	−0.536

[a] Trained on non-normalized inputs.

To visualize the variance captured by the descriptors present in the dataset, we plot the first three principal components of the normalized training set and color each instance by its associated mean complexity. We use this plot to screen for linear and clustering modeling options and to determine the amount of noise contributed by extra features. Once the feature set is reduced to the top 15, we can see the selected features contain a large amount of variation associated with molecular complexity. The total explained variance improves from 47% (Figure 4a) to 90% (Figure 4b).

3.2. Model Selection.

In summary, for each dataset, we trained five algorithms, including partial least-squares, *k*-nearest neighbors, support vector machine, multilayer perceptron, and random forest. This was done for each of the three open-source datasets: the full descriptor set, the set of top descriptors, and interaction terms generated from the latter. Inputs were normalized before training for all models except for the random forest, where it is not necessary beyond comparing coefficients. The cross-validated performance metrics are listed in Table 1.

Reducing the descriptor set increased model performance across the board, though this is hardly surprising. The high R^2 and the negative adjusted R^2 in the models trained on the full feature set suggest that many predictors do not significantly contribute toward prediction. For the latter two training sets, the R^2 and Adj-R^2 are comparable, showing a reasonable ratio between the R^2 and the number of predictors. Interaction terms from the top feature set, however, decreased model performance in PLS and SVM. This might suggest that the nonlinearities are of a different nature than the product and quadratic terms evaluated here. For the reduced feature sets, the neural network and k-NN matched the random forest algorithm's performance. However, since random forest models possess more interpretability, explainability, and model robustness, we continue to use them as the model of choice.

3.3. Parsimonious Model.

Our work shows that a parsimonious ML model of complexity is possible. Using the four highest-ranked descriptors, two physiochemical 0D and two topological 2D descriptors, a random forest model can predict molecular complexity with MSE = −0.349 and R^2 = 82.7% (Table 2). This model, GS-04, has a comparable error to the licensed MSD-207 model and improves significantly on GS-1196 while reducing the number of descriptors (Figure 5a). However, MK-207 still has a lower incidence of errors larger than ±1 as compared to the GS models (Figure 5b). The physiochemical descriptors included the number of stereocenters and the number of heteroatoms (nonhydrogen or noncarbon atoms). Unique topological torsion was the foremost ranked feature in every model and seed used. Topological torsion was developed by Nilakantan et al., to be a short-range descriptor, capturing information about the number of substructures containing unique atomic types, nonhydrogen branches, and pi electron pairs.[34] The unique topological torsion then refers to the number of unique fragments within a molecule. The last descriptor, chi4n, is a structural descriptor derived from a set of connectivity indexes developed by Hall and Kier.[35] The several Chi Index variants capture increasing information about the connectivity of bonds or atoms. While chi4n was used here, other chi indices were also effective. Cross-validated performance using other chi indices in its place differed by a few decimal places.

Table 2. Performance Metrics for Select Models

model	Avg CV-R^2	Avg CV Adj-R^2	Avg CV −RMSE
GS-1196	77.8	−64.1	−0.39
MSD-207	81.3	80.1	−0.36
GS-004	82.7	82.6	−0.349

3.4. App Development.

Two applications were developed to make the final model accessible to both scientists and modelers. The architecture is summarized in Figure 6. For rapid estimates of molecular complexity and Open SMART-PMI, a user-friendly online application requires only the molecule's SDF file. The resulting outputs also include the values of each descriptor factored into the model. The application uses a one-time installation and can subsequently be run locally and therefore does not track molecule entries or results. We have also created a package and command line application (CLI) in Python for model experimentation, allowing for changes in training data, simplified model retraining, and predictions. Like the web app, the CLI can be locally installed by the user and is stateless to prioritize user privacy. We encourage users to augment the model and explore diverse datasets to continue understanding molecular complexity. All materials, including the model and associated apps, can be found at https://github.com/Gilead-IT-GCDM-PDM/Open-SMART-PMI. Programmed entirely in Python, the standard for many cheminformatics and machine learning tools, open-source accessibility is at the forefront of our approach.

In this work, we have made three primary improvements: adding explainability to the search over a large descriptor space, reducing the number of descriptors while maintaining performance, and making the model open source in the language of choice for developers.

4. DISCUSSION

While our findings contribute to refining the calculation of molecular complexity for SMART-PMI, the remaining challenges and directions merit discussion. Future progress involves addressing biases in modeling, exploring the efficacy

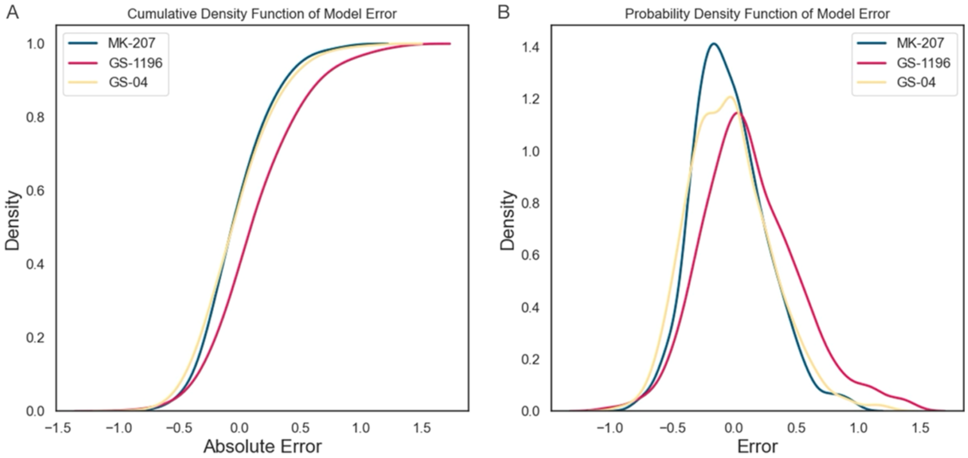


Figure 5. Comparison of model error. (a) Cumulative density functions for absolute error of each model. 80% of predictions are under 0.35 for MK-207 and 0.42 for GS-04. (b) Corresponding probability densities for the error of each model. The area of each curve integrates to 1, and so the *y*-axis represents the *probability density* of the error rate.

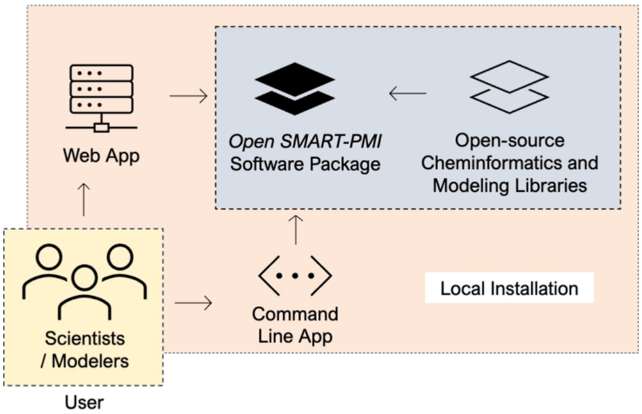


Figure 6. Architecture diagram of the browser and command line applications.

of linear models, leveraging graph networks, and refining PMI estimates for broader industry applications, particularly in multistep processes. Future research in these directions will undoubtedly contribute to the landscape of molecular informatics.

Our effort to select descriptors postmodel training was done cautiously, as the described approach may introduce bias and artificially inflate goodness-of-fit. While our evaluation encompassed both selected and rejected descriptors, mean complexity, and model building, we acknowledge the inherent bias of data-driven models toward the training set. Despite the diversity of the training set, this study is relevant mainly to pharmaceutical contexts, and more work could be done to connect molecular complexity across domains.

Surprisingly, introducing nonlinearity with interaction terms did not improve predictions and, in some cases, may have done the opposite. This observation suggests the viability of linear models, and that the complexity of molecular structures may be effectively captured in a simpler computation. Unique topological torsion and a chirality measure were similarly used in Sheridan et al.'s four-term linear model of mean complexity with an 80% CV-R^2. Given the subjectivity inherent in mean complexity as an average of chemists' assessment, it is not expected for model predictions to be 100% accurate. While they note that mean complexity is a nonlinear variable, this further suggests that a regression model may still yet be able to achieve better fit and low error, enabling a better understanding of the physical implications of molecular complexity.

For this application, where the molecule is represented as a set of molecular descriptors, the multilayer perceptron had comparable performance to the selected interpretable model and thus presented no discernible advantage. In similar use cases with complex relationships between numeric input and output variables, deep learning remains a powerful option with the potential to capture complicated data trends. However, large amounts of training data are typically necessary, which can be an obstacle for chemistry tasks. For more complex tasks, more powerful algorithms are often necessary. Graph networks, which can use a graph-structured compound as input, are also being explored for tasks such as chemical structure prediction and may translate well to understanding molecular complexity. This, coupled with an explaining layer, could also be powerful in identifying molecular features. However, while explainability approaches like Grad-CAM output annotated molecule mappings, it may be difficult to discern long-range topological patterns. More research is currently being done to explore the viability and interpretability of these models.[36,37] Overall, deep learning remains an attractive option for continuing future work.

To increase the usage and accuracy of PMI targets for widespread adoption in industry, more can be done to advance the utility and ease of tracking the PMI of a manufacturing process back to the simplest raw material building blocks. This can be especially challenging in early development where estimates must be made for the PMI of outsourced materials, such as simple starting materials. It may be possible to apply the same model for starting materials and estimate overall PMI as a linear combination of outsourced and in-house materials. Vendors could be further encouraged to track and share the process mass intensity of their materials.

5. CONCLUSIONS

Sustainability is paramount in manufacturing processes. Process chemists throughout the industry are working toward this goal by aiming to reduce the raw materials needed for processing and the waste produced as a byproduct of this process. Molecular complexity is one of the main indicators of the resource efficiency of a process, but without a clear definition, it is difficult to create good estimators for this parameter. Machine learning tools have recently emerged as a viable method to assist experts with decisions in complicated domains. While these tools have had success, more is needed in terms of explaining the basis of the decisions or predictions generated by complex models.

We have used an existing approach to evaluate molecular complexity and generated an open-source model with emphasis on robustness, simplicity, and interpretability. While our tool builds on the original compound complexity and SMART-PMI papers, we have made improvements in terms of interpretability and developed a new model as an open-source tool available to the larger community. We have focused on an open-source approach to develop a cost-effective tool for both internal and external use, as having these kinds of tools available for use to the larger community fosters rapid evaluation and selection of best solutions toward setting common standards.

■ ASSOCIATED CONTENT

Data Availability Statement

The Open SMART-PMI package, including the code, stored models, and applications, can be found online at https://github.com/Gilead-IT-GCDM-PDM/Open-SMART-PMI.

SI Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c02427.

Additional feature analysis, model comparison results, and discussion of model architectures (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

Nicole Tin − *Analytical Sciences, Gilead Sciences Inc, Foster City, California 94404-1147, United States;* orcid.org/0009-0007-2904-2895; Email: nicole.tin1@gilead.com

Remus Osan − *Analytical Sciences, Gilead Sciences Inc, Foster City, California 94404-1147, United States;* Email: remus.osan@gilead.com

Authors

Mandeep Chauhan − *Global Quality Control Systems, Gilead Sciences Inc, Foster City, California 94404-1147, United States*

Kennedy Agwamba − *Analytical Sciences, Gilead Sciences Inc, Foster City, California 94404-1147, United States;* Present Address: Department of Biology, Stanford University, Stanford, California 94305−6104, United States

Yibai Sun − *Process Development, Gilead Alberta ULC, Edmonton, Alberta T6S 1A1, Canada*

Astrid Parsons − *Process Development, Gilead Sciences Inc, Foster City, California 94404-1147, United States*

Philippa Payne − *Global External Manufacturing, Gilead Alberta ULC, Edmonton, Alberta T6S 1A1, Canada*

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c02427

Author Contributions

[#]P.P. and R.O. contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We are grateful to Chris Foti, Victor Rucker, Jeremy Zhang, Mehul Chhatrapati, and Peter Huang for supporting this project. We kindly thank Kevin Chu for the discussion regarding linear models as well as Shikhar Mohan and Wei-Ting Liu for their thoughtful guidance and feedback.

■ REFERENCES

(1) Jimenez-Gonzalez, C.; Ponder, C. S.; Broxterman, Q. B.; Manley, J. B. Using the Right Green Yardstick: Why Process Mass Intensity Is Used in the Pharmaceutical Industry To Drive More Sustainable Processes. *Org. Process Res. Dev.* **2011**, *15* (4), 912−917.

(2) Sherer, E. C.; Bagchi, A.; Kosjek, B.; Maloney, K. M.; et al. Driving Aspirational Process Mass Intensity Using Simple Structure-Based Prediction. *Org. Process Res. Dev.* **2022**, *26*, 1405−1410.

(3) Ertl, P.; Schuffenhauer, A. Estimation of Synthetic Accessibility Score of Drug-like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* **2009**, *1* (1), No. 8.

(4) Coley, C. W.; Rogers, L.; Green, W. H.; Jensen, K. F. SCScore: Synthetic Complexity Learned from a Reaction Corpus. *J. Chem. Inf. Model.* **2018**, *58* (2), 252−261.

(5) Boda, K.; Johnson, A. P. Molecular Complexity Analysis of de Novo Designed Ligands. *J. Med. Chem.* **2006**, *49* (20), 5869−5879.

(6) Takaoka, Y.; Endo, Y.; Yamanobe, S.; Kakinuma, H.; Okubo, T.; Shimazaki, Y.; Ota, T.; Sumiya, S.; Yoshikawa, K. Development of a Method for Evaluating Drug-Likeness and Ease of Synthesis Using a Data Set in Which Compounds Are Assigned Scores Based on Chemists' Intuition. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (4), 1269−1275.

(7) Caille, S.; Cui, S.; Faul, M. M.; Mennen, S. M.; Tedrow, J. S.; Walker, S. D. Molecular Complexity as a Driver for Chemical Process Innovation in the Pharmaceutical Industry. *J. Org. Chem.* **2019**, *84* (8), 4583−4603.

(8) Hann, M. M.; Leach, A. R.; Harper, G. Molecular Complexity and Its Impact on the Probability of Finding Leads for Drug Discovery. *J. Chem. Inf. Comput. Sci.* **2001**, *41* (3), 856−864.

(9) Sherer, E. C.; Bagchi, A.; Kosjek, B.; Maloney, K. M.; Peng, Z.; Robaire, S. A.; Sheridan, R. P.; Metwally, E.; Campeau, L.-C. Driving Aspirational Process Mass Intensity Using Simple Structure-Based Prediction. *Org. Process Res. Dev.* **2022**, *26* (5), 1405−1410.

(10) Roschangar, F.; Zhou, Y.; Constable, D. J. C.; Colberg, J.; Dickson, D. P.; Dunn, P. J.; Eastgate, M. D.; Gallou, F.; Hayler, J. D.; Koenig, S. G.; Kopach, M. E.; Leahy, D. K.; Mergelsberg, I.; Scholz, U.; Smith, A. G.; Henry, M.; Mulder, J.; Brandenburg, J.; Dehli, J. R.; Fandrick, D. R.; Fandrick, K. R.; Gnad-Badouin, F.; Zerban, G.; Groll, K.; Anastas, P. T.; Sheldon, R. A.; Senanayake, C. H. Inspiring Process Innovation via an Improved Green Manufacturing Metric: iGAL. *Green Chem.* **2018**, *20* (10), 2206−2211.

(11) Oprea, T. I.; Bologa, C. Molecular Complexity: You Know It When You See It. *J. Med. Chem.* **2023**, *66* (18), 12710−12714.

(12) Gutman, I.; Das, K. The First Zagreb Index 30 Years After. *Commun. Math. Comput. Chem.* **2004**, *50*, 84−92.

(13) Bertz, S. H. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* **1981**, *103* (12), 3599−3601.

(14) Bonchev, D. G.; Rouvray, D. H. *Complexity: Introduction and Fundamentals*; CRC Press, 2003.

(15) Proudfoot, J. R. A Path Based Approach to Assessing Molecular Complexity. *Bioorg. Med. Chem. Lett.* **2017**, *27* (9), 2014−2017.

(16) von Korff, M.; Sander, T. Molecular Complexity Calculated by Fractal Dimension. *Sci. Rep.* **2019**, *9* (1), No. 967.

(17) Gillet, V. J.; Myatt, G.; Zsoldos, Z.; Johnson, A. P. SPROUT, HIPPO and CAESA: Tools for de Novo Structure Generation and Estimation of Synthetic Accessibility. *Perspect. Drug Discovery Des.* **1995**, *3* (1), 34−50.

(18) Sheridan, R. P.; Zorn, N.; Sherer, E. C.; Campeau, L.-C.; Chang, C. Z.; Cumming, J.; Maddess, M. L.; Nantermet, P. G.; Sinz, C. J.; O'Shea, P. D. Modeling a Crowdsourced Definition of Molecular Complexity. *J. Chem. Inf. Model.* **2014**, *54* (6), 1604−1616.

(19) Li, J.; Eastgate, M. D. Current Complexity: A Tool for Assessing the Complexity of Organic Molecules. *Org. Biomol. Chem.* **2015**, *13* (26), 7164−7176.

(20) Bonchev, D. Information Theoretic Complexity Measures. In *Encyclopedia of Complexity and Systems Science*; Springer, 2009.

(21) Food and Drug Administration. Using Artificial Intelligence and Machine Learning in the Development of Drug and Biological Products 2023 https://www.fda.gov/media/167973/download.

(22) Vora, L. K.; Gholap, A. D.; Jetha, K.; Thakur, R. R. S.; Solanki, H. K.; Chavda, V. P. Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design. *Pharmaceutics* **2023**, *15* (7), No. 1916.

(23) Vasudevan, R. K.; Ziatdinov, M.; Vlcek, L.; Kalinin, S. V. Off-the-Shelf Deep Learning Is Not Enough, and Requires Parsimony, Bayesianity, and Causality. *npj Comput. Mater.* **2021**, *7* (1), No. 16.

(24) Rudin, C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nat. Mach. Intell.* **2019**, *1* (5), 206−215.

(25) Mauri, A.; Consonni, V.; Todeschini, R. Molecular Descriptors. In *Handbook of Computational Chemistry*; Leszczynski, J.; Kaczmarek-Kedziera, A.; Puzyn, T.; Papadopoulos, M. G.; Reis, H.; Shukla, M. K., Eds.; Springer International Publishing: Cham, 2017; pp 2065−2093.

(26) Chandrasekaran, B.; Abed, S. N.; Al-Attraqchi, O.; Kuche, K.; Tekade, R. K. Computer-Aided Prediction of Pharmacokinetic (ADMET) Properties. In *Dosage Form Design Parameters*; Academic Press, 2018; Chapter 21, pp 731−755.

(27) Rouvray, D. H.; Crafford, B. C. The Dependence of Physico-Chemical Properties on Topological Factors. *S. Afr. J. Sci.* **1976**, *72* (2), No. 47.

(28) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Polar Surface Area as a Sum of Fragment-Based Contributions and Its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714−3717.

(29) Prasanna, S.; Doerksen, R. J. Topological Polar Surface Area: A Useful Descriptor in 2D-QSAR. *Curr. Med. Chem.* **2009**, *16* (1), 21−41.

(30) Landrum, G.; Tosco, P.; Kelley, B.; Ric; Cosgrove, D.; sriniker; gedeck; Vianello, R.; Schneider, N.; Kawashima, E.; Jones, G.; N, D.; Dalke, A.; Cole, B.; Swain, M.; Turk, S.; Savelyev, A.; Vaucher, A.; Wójcikowski, M.; Take, I.; Scalfani, V. F.; Probst, D.; Ujihara, K.; Godin, G.; Walker, R.; Lehtivarjo, J.; Pahl, A.; Berenger, F.; jasondbiggs; strets123 *Rdkit/Rdkit: _09_3 (Q3 2023)2023*, Zenodo 2023 DOI: 10.5281/ZENODO.591637.

(31) Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: A Molecular Descriptor Calculator. *J. Cheminf.* **2018**, *10* (1), No. 4.

(32) Lundberg, S.; Lee, S.-I. A Unified Approach to Interpreting Model Predictions. 2017, arXiv:1705.07874v2. arXiv.org e-Print archive. https://doi.org/10.48550/arXiv.1705.07874.

(33) Molnar, C. 9.6 SHAP (SHapley Additive exPlanations) | Interpretable Machine Learning.

(34) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27* (2), 82−85.

(35) Hall, L. H.; Kier, L. B. The Molecular Connectivity Chi Indexes and Kappa Shape Indexes in Structure-Property Modeling. *Rev. Comput. Chem.* **1991**, *2*, 367−422.

(36) Wu, Z.; Wang, J.; Du, H.; Jiang, D.; Kang, Y.; Li, D.; Pan, P.; Deng, Y.; Cao, D.; Hsieh, C.-Y.; Hou, T. Chemistry-Intuitive Explanation of Graph Neural Networks for Molecular Property Prediction with Substructure Masking. *Nat. Commun.* **2023**, *14* (1), No. 2585.

(37) Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T. Could Graph Neural Networks Learn Better Molecular Representation for Drug Discovery? A Comparison Study of Descriptor-Based and Graph-Based Models. *J. Cheminf.* **2021**, *13* (1), No. 12.