ABSTRACT: Solubility is a key metric for therapeutic compounds. Conversely, insoluble compounds cloud the accuracy of assays at all stages of chemical biology and drug discovery. Herein, we disclose naive Bayesian classifier models to predict aqueous solubility. Publicly accessible aqueous solubility data were used to create two full, or nonpruned, training sets. These two sets were also combined to create a full fused set, and a training set comprised of a literature collation of solubility data was also considered as a reference. We tested different extents of data pruning on the training sets and constructed machine learning models that were evaluated with two independent, external test sets that contained compounds that were different from the training sets. The best pruned and fused model was significantly more accurate, in comparison to either the full model or the full fused model, with the prediction of these external test sets. By carefully removing data from the training set, less information can be used to create more accurate machine learning models for aqueous solubility. This knowledge and the curated training sets should prove useful to future machine learning approaches.

INTRODUCTION

More than 20 years after the landmark publication that provided the rule of 5 and asserted the importance of high solubility alongside potency and permeability, poor aqueous solubility (hereafter solubility) of drugs and their precursor hit, lead, and chemical tool compounds continue to be a major issue for chemical biology and drug discovery research. Solubility can have a dramatic impact on the absorption, distribution, metabolism, and excretion (ADME) and pharmacokinetic (PK) properties of a molecule, and it can necessitate significant efforts in formulation to enable downstream in vivo efficacy studies. Within the early stages of development, a compound’s solubility is initially gauged by its kinetic solubility and, in contrast, thermodynamic or equilibrium solubility is generally reserved for downstream assessment. These values are crucial as low solubility can cause potentially valuable compounds to be ignored or neglected, because their activity may be underestimated.

Similarly, large differences in the solubility of compounds can produce noisy and inaccurate structure–activity relationship (SAR) data, while adversely affecting in vivo performance, diminishing PK parameters, reducing efficacy, and driving toxicity. The ability to more accurately predict the solubility of druglike small molecules would, therefore, be of significant value.

Publications have detailed studies of models and methods for predicting solubility from structure alone or from other measured or estimated molecular properties. Early studies by Huuskonen and collaborators collated sets of compounds (from 24 structurally similar compounds to 884 more diverse compounds) for which solubility was determined, and molecular descriptors with artificial neural networks (ANN) or multiple linear regression (MLR) were implemented to train and validate models. The Huuskonen data have been used extensively by others to compare new methods for predicting solubility, which highlights the value of developing new, well-curated data sets to the community, far beyond the utility of any initial model. Many different machine learning methods have since been used to predict solubility, such as random forest, partial least squares, Support Vector Machines, Gaussian process Bayesian, ridge regression, and deep learning. Across these previous publications, we have seen a growth in the size of the solubility training and test sets, as well as the diversity of methods used to build predictive models.

We present a new study on predicting solubility, in which we curated the largest publicly available solubility data set—the Molecular Libraries Small Molecule Repository (MLSMR) set of 57,824 unique compounds (PubChem AID 1996). Naïve Bayesian classifier (heretofore referred to as Bayesian) models were constructed using this set, an AstraZeneca (AZ) set of 1,763 compounds (assay ID CHEMBL3301364), and a full fused set in which the MLSMR and AZ sets were combined together to generate 59,510 unique compounds (MLSMR + AZ). While contemplating training sets, it is important to mention that we previously proposed a novel data pruning...
strategy, in which moderately active compounds are deleted from the training set for a binary classifier model. The pruned Bayesian model was more accurate than the conventional full (i.e., nonpruned) model for predicting the mouse liver microsomal (MLM) stability of independent, external test, and validation sets. The pruned Bayesian model, superior to random forest and support vector machine models that utilized the same training set and descriptors, was also able to guide the evolution of an antitubercular thienopyrimidine to afford analogs with superior metabolic stability.21

In the present study, we systematically tested several different extents of pruning of the MLSMR set, the AZ set, and the fused MLSMR + AZ set to examine how different levels of pruning impact the ability to predict whether a compound has a solubility $\geq 100$ $\mu$M (a typical goal value of this property).22 Since we utilized a Bayesian classifier model with this 100 $\mu$M cutoff, our goal was to differentiate “good” or sufficiently soluble compounds, defined as having a solubility $\geq 100$ $\mu$M, from “bad” or insufficiently soluble compounds, which have a solubility <100 $\mu$M. Although some may consider 100 $\mu$M a high bar for solubility, we believe it to be a reasonable metric given our experience. Since a purpose of the model is to enable the rapid filtering of large areas of chemical space (i.e., millions of compounds), we preferred the selection of a high bar, instead of a minimum value of solubility that could be tolerated for a given program, to help us focus on the regions of chemical space that are more likely to be fruitful. This is especially important because during the course of a molecular optimization, some solubility may be sacrificed in pursuit of other compound metrics. We hypothesized that pruning an insufficient number of compounds from the training set would not produce a significant boost in accuracy, while pruning too many compounds would significantly decrease the predictive power of the model produced. Although this hypothesis is fairly obvious, finding an optimal extent of pruning was nontrivial. The resulting best model, as judged by internal and external statistics, was found to be a model we have termed the Pruned and Fused MLSMR + AZ.

RESULTS AND DISCUSSION

Our search of publicly available solubility assay data revealed that different approaches have been utilized to measure solubility.24 For example, kinetic methods are generally used for high-throughput and early discovery studies. For this method, samples are usually introduced as dimethyl sulfoxide (DMSO) stock solutions into the solubility buffer (typically pH 7.4 phosphate-buffered saline (PBS)), with light scattering data obtained within 1−4 h. In the same vein, a semi-equilibrium method also employs this protocol but with measurements being made after 24 h. It is important to note that both of these approaches assay the compound in the crystalline form, which increases the ultimately measured solubility. In contrast, the third method of assaying solubility is thermodynamic or equilibrium. In this approach, the compound, in its crystalline form, is directly added to solubility buffer with measurements taken after 24 h. It is crucial to point out that this method is a more conservative measure of solubility due to the absence of DMSO used in the kinetic and semi-equilibrium protocols.24,25 The more conservative method requires significantly greater amounts of the compound than the high-throughput methods (milligrams of the solid compound vs microliters of mM DMSO stocks). Hence, for hit-to-lead optimization programs, where relatively small amounts of compounds are being synthesized and profiled, kinetic and semi-equilibrium methods are more typically employed.

The solubility assay data curated herein were identified within PubChem BioAssay and ChEMBL.26−28 These data were reformatted by combining the PubChem data from a comma-separated-value file (csv; sorted by compound ID) and a structural data file (sdf) into a Discovery Studio spreadsheet or by converting the Excel file from ChEMBL into an sdf file using CDD (Collaborative Drug Discovery; www.crdi.org).29 The data were then curated in Discovery Studio 4.5 (BIOVIA, Inc.), and a custom Pipeline Pilot script was used to delete duplicate compounds (according to canonical SMILES) and salts or buffers (by keeping the largest fragment component). Duplicates were deleted from each full training set, as well as from the full fused set and the two external test sets to be discussed later. This preparation process in Pipeline Pilot also included calculating the dominant protonation state and tautomer of each compound at pH 7.4.

Our two original training sets were: (a) the Full MLSMR data set measured by the Sanford-Burnham Medical Research Institute (PubChem AID 1996), composed of 57 824 unique compounds, of which 31 644 compounds (54.7%) were defined as soluble (using the definition that solubility $\geq 100$ $\mu$M = soluble = good or active), and (b) the Full AZ training set (assay ID CHEMBL3301364), composed of 1763 compounds with solubility data measured by AstraZeneca, of which 618 compounds (35.1%) were defined as soluble. The MLSMR data set contains solubility data that was measured by the semi-equilibrium method, while the solubility data from the AZ set was measured via the equilibrium method. We also utilized the Huuskonen collation of solubility data as an additional training set,6,9 containing 987 soluble compounds (76.5%) out of a total of 1290 compounds. Given the Huuskonen training set’s composition of solubility data measured with more than one methodology, we hypothesized that combining the kinetic (MLSMR) and equilibrium (AZ) solubility would ultimately train a better model. Hence, the Full Fused MLSMR + AZ set has 59 510 compounds (after deleting duplicates), of which 32 228 compounds (54.2%) were defined as soluble. Given our recent efforts with the MLM stability training set pruning to remove potential disinformation31 and arrive at models with significant prospective predictive ability in a medicinal chemistry optimization,22 we examined four different extents of data pruning on the MLSMR set, four different levels of pruning on the AZ set, and eight different ranges of pruning on the fused set of AZ + MLSMR (Full Fused AZ + MLSMR). To construct Bayesian models for each training set, Pipeline Pilot 9.5 (BIOVIA, Inc.) was utilized with the nine standard (default) descriptors (i.e., molecular fractional polar surface area, molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors, ALOGP, number of aromatic rings, total number of rings, number of rotatable bonds, and the FCFP_6 substructural fingerprints) that have worked well for us in previous Bayesian models for predicting other properties.21,22,31,52 The details regarding the different ranges of compounds that were deleted from the training set to generate the different pruned models are in the Supporting Information (Table S1).

Thus, a total of 20 training sets was included in this study: the two full sets, the full fused set, sixteen different pruned sets,
Table 1. Internal Statistics and External Statistics for Select Models

| Bayesian model                          | ROC score<sup>a</sup> | sensitivity<sup>b</sup> | specificity<sup>b</sup> | concordance<sup>c</sup> | MCC<sup>i</sup> | Cohen’s kappa<sup>d</sup> | F1 score<sup>e</sup> |
|----------------------------------------|------------------------|-------------------------|-------------------------|------------------------|-----------------|---------------------------|---------------------|
| internal set (with five-fold cross validation) |                        |                         |                         |                        |                 |                           |                     |
| Full MLSMR<sup>c</sup>                 | 0.836                  | 77.8                    | 81.5                    | 79.5                   | n/a             | n/a                      | n/a                 |
| Full AZ<sup>c</sup>                    | 0.744                  | 86.7                    | 76.8                    | 80.3                   | n/a             | n/a                      | n/a                 |
| Full Fused AZ + MLSMR<sup>c</sup>     | 0.833                  | 76.6                    | 82.3                    | 79.2                   | n/a             | n/a                      | n/a                 |
| Pruned and Fused AZ + MLSMR<sup>c</sup> | 0.941                  | 93.9                    | 86.7                    | 86.9                   | n/a             | n/a                      | n/a                 |
| Huuskonen<sup>e</sup>                 | 0.926                  | 89.9                    | 94.1                    | 90.9                   | n/a             | n/a                      | n/a                 |
| External PubChem set<sup>f</sup>      |                        |                         |                         |                        |                 |                           |                     |
| Full MLSMR<sup>c</sup>                | 0.534                  | 37.1                    | 45.7                    | 44.2                   | −0.13           | −0.10                    | 0.19                |
| Full AZ<sup>c</sup>                   | 0.761                  | 34.3                    | 32.7                    | 33.0                   | −0.26           | −0.17                    | 0.15                |
| Full Fused AZ + MLSMR<sup>c</sup>     | 0.687                  | 34.3                    | 42.6                    | 41.1                   | −0.18           | −0.13                    | 0.17                |
| Pruned and Fused AZ + MLSMR<sup>c</sup> | 0.824                  | 85.7                    | 65.4                    | 69.0                   | 0.39            | 0.33                      | 0.50                |
| Huuskonen<sup>e</sup>                 | 0.630                  | 25.7                    | 59.9                    | 53.8                   | −0.11           | −0.10                    | 0.17                |
| External JSF set<sup>g</sup>          |                        |                         |                         |                        |                 |                           |                     |
| Full MLSMR<sup>c</sup>                | 0.756                  | 35.7                    | 89.5                    | 75.0                   | 0.30            | 0.28                      | 0.43                |
| Full AZ<sup>c</sup>                   | 0.714                  | 78.6                    | 60.5                    | 65.4                   | 0.35            | 0.31                      | 0.55                |
| Full Fused AZ + MLSMR<sup>c</sup>     | 0.842                  | 35.7                    | 89.5                    | 75.0                   | 0.30            | 0.28                      | 0.43                |
| Pruned and Fused AZ + MLSMR<sup>c</sup> | 0.724                  | 64.3                    | 86.8                    | 80.8                   | 0.51            | 0.51                      | 0.64                |
| Huuskonen<sup>e</sup>                 | 0.835                  | 78.6                    | 68.4                    | 71.2                   | 0.42            | 0.39                      | 0.59                |

<sup>a</sup>The Full MLSMR model was constructed from the training set composed of 57,824 unique compounds, of which 54.7% were defined as soluble using our ≥100 μM criteria. <sup>b</sup>The Full AstraZeneca (or AZ) model was built from the training set composed of 1763 compounds, of which 35.1% were soluble. <sup>c</sup>The Full Fused MLSMR + AZ model was generated by combining all of the compounds from the Full MLSMR set with the Full AstraZeneca set. It had 59,510 compounds, of which 54.2% were defined as soluble. <sup>d</sup>The Pruned and Fused MLSMR + AZ model corresponds to the training set in which we combined a pruned version of the AZ set (compounds with a solubility of 25−99 μM were deleted) with a pruned version of the MLSMR set (only the subset of compounds with a solubility <25 μM were included). This training set contains 17,460 compounds, of which 3.5% were soluble. “The Huuskonen reference model was constructed using a concatenation of previously published training and test sets.” It has 1290 compounds, of which 76.5% are soluble. <sup>e</sup>The ROC score is the area under the receiver operating characteristic curve from a five-fold cross-validation study. <sup>f</sup>Sensitivity represents the percentage of correctly identified soluble compounds (true positives). <sup>g</sup>Specificity signifies the percentage of correctly identified insoluble compounds (true negatives). <sup>h</sup>Concordance corresponds to the overall accuracy: percentage of (true positives + true negatives)/total number of compounds. <sup>i</sup>The External PubChem set has 197 unique compounds (that are not part of any training set), of which 17.8% are soluble. <sup>j</sup>The External JSF set has 52 compounds, of which 26.9% are soluble. <sup>k</sup>Mathews’s Correlation Coefficient (MCC), also referred to as the phi coefficient, is a chance-corrected statistic where MCC = 1 indicates perfect agreement, MCC = −1 indicates total disagreement, and MCC = 0 indicates that the model is no better than random. <sup>l</sup>Cohen’s Kappa is a chance-corrected statistic that uses a different method to calculate the random likelihood of making correct predictions for the external set. Kappa <0 indicates no agreement, Kappa of 0–0.2 indicates slight agreement, Kappa of 0.21–0.40 is fairly predictive, and Kappa of 0.41–0.60 indicates moderate agreement. <sup>m</sup>The F1 score is the harmonic mean of precision (i.e., the positive predictive value hit rate) and sensitivity. The best F1 score is 1, while the worst score possible is 0.

Figure 1. ROC curves for evaluation of the solubility Bayesian models with the (A) External PubChem and (B) External JSF sets.
The S2 compounds (26.9%) were soluble. Evaluations of the external statistics for these models with these two test sets are shown in Tables 1, S2, and S3. Examining model performance quantitatively through chance-corrected statistics and the calculation of a mean value for each model of its ROC and concordance statistics over the two different training sets led to the selection of an optimal model, termed the Pruned and Fused AZ + MLSMR model. This model’s training set was derived by fusing the AZ set, pruned such that compounds with a solubility of 25–99 μM were deleted, with a pruned version of the MLSMR set, such that only the subset of compounds with a solubility <25 μM were included.

The Pruned and Fused AZ + MLSMR model was then compared to its parental models as well as the model trained with the Huuskonen data set (Table 1). While the Huuskonen model’s internal statistics were slightly better than those for the Pruned and Fused AZ + MLSMR model, the external statistics for the Pruned and Fused AZ + MLSMR model were superior to all models. Histogram-based analysis demonstrated the ability of the Pruned and Fused AZ + MLSMR model to establish a better separation between the actual soluble and insoluble compounds in the External PubChem set (Figure S1) and External JSF set (Figure S2). Visually, this is most apparent in the ROC curves comparing model performance with the External PubChem set (Figure 1A). While the Pruned and Fused AZ + MLSMR model did not compare as favorably with the Huuskonen model when using the smaller External JSF set (Figure 1B) and a ROC analysis, the Pruned and Fused AZ + MLSMR model did demonstrate superior concordance. The greater performance of the Pruned and Fused AZ + MLSMR model was also reached when expanding the model comparison to include the following chance-corrected statistics (Table 1): Matthew’s Correlation Coefficient (MCC) and Cohen’s Kappa. Both metrics calculate the random probability of making correct predictions and account for unbalanced data sets. The MCC for the Pruned and Fused AZ + MLSMR model was better than random (>0) and also outperformed the Huuskonen model and both parental models with both external test sets. It was also superior to the Huuskonen model for both external sets when considering Cohen’s Kappa. Finally, when considering a measure of the model’s ability to predict only soluble compounds, F1 score—the harmonic mean of the positive predictive value (PPV) hit rate (Table S4) and sensitivity (Table 1)—the Pruned and Fused AZ + MLSMR model again outperformed these other models.

The internal and external statistics of the Pruned and Fused AZ + MLSMR model demonstrate potential predictive value superior to that of the Huuskonen model. Principal component analysis (PCA) (Figure 2) demonstrated that the Pruned and Fused AZ + MLSMR training set also occupies different physiochemical property space than the Huuskonen set. This further highlights the novelty and significance of this curated training set to the modeling community, since a well-curated training set that includes novel parts of physiochemical property space can lead to new studies with different modeling approaches that far outlive the utility of any particular initial study. In addition, the soluble compounds (large spheres in Figure 2B) and insoluble compounds (smaller spheres in Figure 2B) in the Pruned and Fused AZ + MLSMR training set generally occupy similar regions of physiochemical property space. This observation underscores the need for combining structural fingerprints with the physiochemical properties to predict solubility accurately.

**CONCLUSIONS**

This study began with the MLSMR and AZ training sets of 57,824 and 17,635 compounds, respectively, and explored fusion of the two data sets and then different levels of pruning the training set, following the precedent of our models for mouse liver microsomal stability and Vero cell cytotoxicity. The end result was a Bayesian model derived from the fusion of the MLSMR and AZ training sets where pruning of both parental data sets was conducted: compounds with a solubility of 25–99 μM were deleted in the AZ set and only the subset of MLSMR compounds with a solubility <25 μM were included. This model predicted subsets of its training set as well or better than the other models, depending on the five-fold cross-validation statistic, and, most importantly, it outperformed the other models in predicting the solubility of two independent external test sets (one from PubChem and one from our laboratory’s compound collection). Comprising a differentially localized chemical space than the previous Huuskonen data set, we anticipate that the Pruned and Fused AZ + MLSMR training set and its machine learning models will be of value to the community to predict the solubility of

---

**Figure 2.** Principal component analysis (PCA) comparing the Pruned and Fused AZ + MLSMR training set (in red) to the Huuskonen reference training set (in blue). (A) shows all compounds with the same radii, while (B) has insoluble compounds displayed with smaller radii.

http://dx.doi.org/10.1021/acsomega.0c01251
ACS Omega 2020, 5, 16662–16677
chemical tool and drug discovery of small molecules of therapeutic relevance.

■ METHODS

Please refer to the main text for the construction of the Bayesian models and to the Supporting Information for additional details regarding their construction and evaluation.

Principal Component Analyses. The physiochemical property space sampled by each data set was compared to the other data sets by concatenating their relevant sdf files together and performing a principal component analysis (PCA) in Discovery Studio 4.5. (BIOVIA, Inc.), using the default protocol.25,31 Briefly, eight different interpretable molecular descriptors (that characterize the properties of each compound as a whole entity) were calculated and utilized to describe the physiochemical properties of each compound: number of hydrogen bond donors, number of hydrogen bond acceptors, AlogP, molecular weight, number of rings, number of aromatic rings, number of rotatable bonds, and molecular fractional polar surface area. These are eight of the nine descriptors that were utilized to construct our Bayesian models. The ninth descriptor (which was used for the Bayesian models but not for PCA) was FCFP_6, which describes the topology of different substructural pieces of each compound.

■ ASSOCIATED CONTENT

* Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/10.1021/acsomega.0c01251.

Full MLSMR.sdf, Full AstraZeneca.sdf, Full Fused MLSMR + AZ.sdf, Pruned and Fused MLSMR + AZ.sdf, Huuskonen.sdf, External PubChem.sdf (Supporting files); Histograms depicting how the (A) Pruned and Fused MLSMR + AZ Bayesian performed when scoring the External PubChem set, as compared to the (B) Full MLSMR model and the (C) Full Fused MLSMR + AZ model (Figure S1); Histograms illustrating how the (A) Pruned and Fused MLSMR + AZ Bayesian performed when scoring the External JSF set, as compared to the (B) Full MLSMR model and the (C) Full AZ model (Figure S2); Internal statistics from five-fold cross validation comparing the sixteen different pruned solubility Bayesian models and the three full models (Table S1); External statistics from evaluating the sixteen different pruned solubility Bayesian models and the three full models on the External PubChem set (Table S2); External statistics from evaluating the sixteen different pruned solubility Bayesian models and the three full models on the External JSF set (Table S3); Additional evaluation of select solubility Bayesian models: hit rates, filtering rates, and enrichment factors for independent, external sets (Table S4) (PDF) Supporting files, containing relevant data sets: Full MLSMR.sdf, Full AstraZeneca.sdf, Pruned and Fused MLSMR + AZ.sdf, Full Fused MLSMR + AZ.sdf, Huuskonen.sdf, and External PubChem.sdf (ZIP)

■ AUTHOR INFORMATION

Corresponding Author

Joel S. Freundlich — Department of Pharmacology, Physiology, and Neuroscience and Division of Infectious Disease, Department of Medicine and the Ray V. Lawrence Center for the Study of Emerging and Re-emerging Pathogens, Rutgers University—New Jersey Medical School, Newark, New Jersey 07103, United States; orcid.org/0000-0002-3411-3455; Phone: +1 973-972-7165; Email: freundjs@rutgers.edu

Authors

Alexander L. Perryman — Department of Pharmacology, Physiology, and Neuroscience, Rutgers University—New Jersey Medical School, Newark, New Jersey 07103, United States

Daigo Inoyama — Department of Pharmacology, Physiology, and Neuroscience, Rutgers University—New Jersey Medical School, Newark, New Jersey 07103, United States

Jimmy S. Patel — Department of Pharmacology, Physiology, and Neuroscience, Rutgers University—New Jersey Medical School, Newark, New Jersey 07103, United States

Sean Ekins — Collaborations in Chemistry, Inc., Fuquay-Varina, North Carolina 27526, United States; orcid.org/0000-0002-5691-5790

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.0c01251

Author Contributions

A.L.P. and D.I. contributed equally to this work.

Notes

The authors declare the following competing financial interest(s): Dr. Sean Ekins is the founder and CEO of Collaborations Pharmaceuticals.

■ ACKNOWLEDGMENTS

J.S.F. acknowledges support from award number U19AI109713 NIH/NIAID for the “Center to develop therapeutic countermeasures to high-threat bacterial agents,” from the National Institutes of Health: Centers of Excellence for Translational Research (CETR). J.S.F. acknowledges BIOVIA for kindly providing Discovery Studio and Pipeline Pilot.

■ REFERENCES

1. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv. Drug Delivery Rev. 1997, 23, 3–25.

2. Saal, C.; Pietereit, A. C. Optimizing solubility: kinetic versus thermodynamic solubility temptations and risks. Eur. J. Pharm. Sci. 2012, 47, 589–595.

3. Di, L.; Kerns, E. H. Biological assay challenges from compound solubility: strategies for bioassay optimization. Drug Discovery Today 2006, 11, 446–451.

4. Huuskonen, J.; Livingstone, D. J.; Manallack, D. T. Prediction of drug solubility from molecular structure using a drug-like training set. SAR QSAR Environ. Res. 2008, 19, 191–212.

5. Livingstone, D. J.; Ford, M. G.; Huuskonen, J. J.; Salt, D. W. Simultaneous prediction of aqueous solubility and octanol/water partition coefficient based on descriptors derived from molecular structure. J. Comput.-Aided Mol. Des. 2001, 15, 741–752.

6. Huuskonen, J.; Rantanen, J.; Livingstone, D. Prediction of aqueous solubility for a diverse set of organic compounds based on atom-type electrotropological state indices. Eur. J. Med. Chem. 2000, 35, 1081–1088.

7. Huuskonen, J. Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. J. Chem. Inf. Comput. Sci. 2000, 40, 773–777.

8. Huuskonen, J.; Salo, M.; Taskinen, J. Aqueous solubility prediction of drugs based on molecular topology and neural network modeling. J. Chem. Inf. Comput. Sci. 1998, 38, 450–456.
(9) Huuskonen, J.; Salo, M.; Taskinen, J. Neural network modeling for estimation of the aqueous solubility of structurally related drugs. *J. Pharmocol. Sci.* 1997, 86, 450–454.

(10) Yan, A.; Gasteiger, J.; Krug, M.; Anzali, S. Linear and nonlinear functions on modeling of aqueous solubility of organic compounds by two structure representation methods. *J. Comput.-Aided Mol. Des.* 2004, 18, 75–87.

(11) Wegner, J. K.; Zell, A. Prediction of aqueous solubility and partition coefficient optimized by a genetic algorithm based descriptor selection method. *J. Chem. Inf. Comput. Sci.* 2003, 43, 1077–1084.

(12) Gao, H.; Shamunsgasundaram, V.; Lee, P. Estimation of aqueous solubility of organic compounds with QSAR approach. *Pharm. Res.* 2002, 19, 497–503.

(13) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of aqueous solubility of organic compounds by the general solubility equation (GSE). *J. Chem. Inf. Comput. Sci.* 2001, 41, 1208–1217.

(14) Liu, R.; So, S. S. Development of quantitative structure-property relationship models for early ADME evaluation in drug discovery. 1. Aqueous solubility. *J. Chem. Inf. Comput. Sci.* 2001, 41, 1633–1639.

(15) Korotkov, A.; Tkachenko, V.; Russo, D. P.; Ekins, S. Comparison of Deep Learning with Multiple Machine Learning Methods and Metrics Using Diverse Drug Discovery Data Sets. *Mol. Pharmaceutics* 2017, 14, 4462–4475.

(16) McDonagh, J. L.; Nath, N.; De Ferrari, L.; van Mourik, T.; Mitchell, J. B. Uniting cheminformatics and chemical theory to predict the intrinsic aqueous solubility of crystalline druglike molecules. *J. Chem. Inf. Model.* 2014, 54, 844–856.

(17) Lusci, A.; Pollastri, G.; Baldi, P. Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. *J. Chem. Inf. Model.* 2013, 53, 1563–1575.

(18) Obrezanova, O.; Gola, J. M.; Champness, E. J.; Segall, M. D. Automatic QSAR modeling of ADME properties: blood-brain barrier penetration and aqueous solubility. *J. Comput.-Aided Mol. Des.* 2008, 22, 431–440.

(19) Schröter, T. S.; Schwaighofer, A.; Mika, S.; Ter Laak, A.; Suelze, D.; Ganzer, U.; Heinrich, N.; Muller, K. R. Estimating the domain of applicability for machine learning QSAR models: a study on aqueous solubility of drug discovery molecules. *J. Comput.-Aided Mol. Des.* 2007, 21, 651–664.

(20) Palmer, D. S.; O’Boyle, N. M.; Glen, R. C.; Mitchell, J. B. Random forest models to predict aqueous solubility. *J. Chem. Inf. Model.* 2007, 47, 150–158.

(21) Perryman, A. L.; Stratton, T. P.; Ekins, S.; Freundlich, J. S. Predicting Mouse Liver Microsomal Stability with “Pruned” Machine Learning Models and Public Data. *Pharm. Res.* 2016, 33, 433–449.

(22) Stratton, T. P.; Perryman, A. L.; Vilcheze, C.; Russo, R.; Li, S. G.; Patel, J. S.; Singleton, E.; Ekins, S.; Connell, N.; Jacobs, W. R., Jr.; Freundlich, J. S. Addressing the Metabolic Stability of Antituberculars through Machine Learning. *ACS Med. Chem. Lett.* 2017, 8, 1099–1104.

(23) Inoyama, D.; Patet, S. D.; Russo, R.; Kandasamy, S.; Kumar, P.; Singleton, E.; Occi, J.; Tuckman, M.; Zimmerman, M. D.; Ho, H. P.; Perryman, A. L.; Dartois, V.; Connell, N.; Freundlich, J. S. Novel Pyrimidines as Antituberculars. *Antimicrob. Agents Chemother.* 2018, 62, e02063–e02017.

(24) Di, L.; Fish, P. V.; Mano, T. Bridging solubility between drug discovery and development. *Drug Discovery Today* 2012, 17, 486–495.

(25) Hancock, B. C.; Parks, M. What is the true solubility advantage for amorphous pharmaceuticals? *Pharm. Res.* 2000, 17, 397–404.

(26) Wang, Y.; Bryant, S. H.; Cheng, T.; Wang, J.; Gindulyte, A.; Shoemaker, B. A.; Thiessen, P. A.; He, S.; Zhang, J. PubChem BioAssay: 2017 update. *Nucleic Acids Res.* 2017, 45, D955–D963.

(27) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A.; Wang, J.; Yu, B.; Zhang, J.; Bryant, S. H. PubChem Substance and Compound databases. *Nucleic Acids Res.* 2016, 44, D1202–D1213.

(28) Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F.; A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* 2014, 42, D1083–D1090.

(29) Hohman, M.; Gregory, K.; Chibale, K.; Smith, P. J.; Ekins, S.; Bunin, B. Novel web-based tools combining chemistry informatics, biology and social networks for drug discovery. *Drug Discovery Today* 2009, 14, 261–270.

(30) Landis, J. R.; Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics* 1977, 33, 159–174.

(31) Perryman, A. L.; Patel, J. S.; Russo, R.; Singleton, E.; Connell, N.; Ekins, S.; Freundlich, J. S. Naive Bayesian Models for Vero Cell Cytotoxicity. *Pharm. Res.* 2018, 35, No. 170.

(32) Ekins, S.; Reynolds, R. C.; Kim, H.; Koo, M. S.; Ekonomidis, M.; Talauke, M.; Patet, S. D.; Woolhiser, L. K.; Lenaerts, A. J.; Bunin, B. A.; Connell, N.; Freundlich, J. S. Bayesian models leveraging bioactivity and cytotoxicity information for drug discovery. *Chem. Biol.* 2013, 20, 370–378.

(33) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* 1975, 405, 442–451.

(34) Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* 1960, 20, 37–46.

(35) Dice, L. R. Measures of the Amount of Ecologic Association between Species. *Ecology* 1945, 26, 297–302.