Chemical Research in Toxicology

pubs.acs.org/crt

Article

# Transparency in Modeling through Careful Application of OECD's QSAR/QSPR Principles via a Curated Water Solubility Data Set

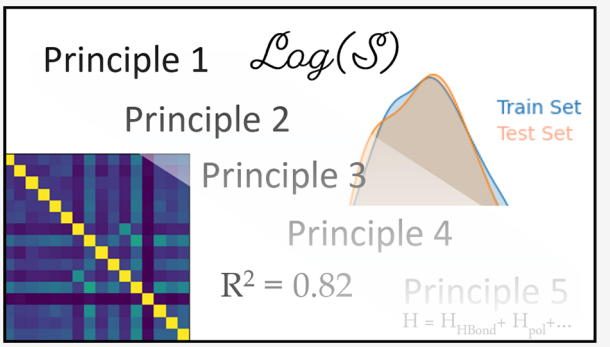
Charles N. Lowe,*,§ Nathaniel Charest,*,§ Christian Ramsland, Daniel T. Chang, Todd M. Martin, and Antony J. Williams

Cite This: *Chem. Res. Toxicol.* 2023, 36, 465−478

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

**ABSTRACT:** The need for careful assembly, training, and validation of quantitative structure−activity/property models (QSAR/QSPR) is more significant than ever as data sets become larger and sophisticated machine learning tools become increasingly ubiquitous and accessible to the scientific community. Regulatory agencies such as the United States Environmental Protection Agency must carefully scrutinize each aspect of a resulting QSAR/QSPR model to determine its potential use in environmental exposure and hazard assessment. Herein, we revisit the goals of the Organisation for Economic Cooperation and Development (OECD) in our application and discuss the validation principles for structure−activity models. We apply these principles to a model for predicting water solubility of organic compounds derived using random forest regression, a common machine learning approach in the QSA/PR literature. Using public sources, we carefully assembled and curated a data set consisting of 10,200 unique chemical structures with associated water solubility measurements. This data set was then used as a focal narrative to methodically consider the OECD's QSA/PR principles and how they can be applied to random forests. Despite some expert, mechanistically informed supervision of descriptor selection to enhance model interpretability, we achieved a model of water solubility with comparable performance to previously published models (5-fold cross validated performance 0.81 $R^2$ and 0.98 RMSE). We hope this work will catalyze a necessary conversation around the importance of cautiously modernizing and explicitly leveraging OECD principles while pursuing state-of-the-art machine learning approaches to derive QSA/PR models suitable for regulatory consideration.






## INTRODUCTION

The need for careful assembly, training, and validation of quantitative structure−activity/property models (QSAR/PR) is more significant than ever as data sets become larger and sophisticated machine learning approaches become increasingly ubiquitous and accessible to the scientific community. The United States Environmental Protection Agency (US EPA) Center for Computational Toxicology and Exposure (CCTE) has an active research program in cheminformatics and structure-based modeling, employing a wide range of machine learning methods to model physicochemical properties as well as toxicity, bioactivity, transport, and exposure end points.[1−3]

The ability of a compound to dissolve in water is a key attribute in assessing its chemical behavior in terrestrial and aquatic systems. Biological, medical, environmental, and many industrial processes tend to be inextricably linked to aqueous solvation due to water's ubiquitous presence on earth and within living systems. Thus, water solubility, defined as the maximum amount of a solute that can be dissolved in a volume of water, is of critical interest to chemists looking to assess, design or regulate around many of the chemical processes that facilitate the modern world. In environmental fate assessment, water solubility is an important consideration in determining how a chemical will partition in air, water, or soil. Highly soluble compounds tend to rapidly migrate into groundwater, whereas less soluble compounds are more likely to be associated with higher bioaccumulation potential.

For pharmaceutical development, accurate water solubilities are often needed to optimize candidate drugs to possess suitable transport properties within tissues of interest, with hydrophilic (high water solubility) species possessing differing localization and metabolic properties than their hydrophobic (low water solubility) counterparts. Reliable water solubility values play an important role in reducing the costs and enquiry space surrounding expensive and potentially challenging (i.e., difficult-to-test substances) empirical investigations.[4]

**Received:** November 30, 2022
**Published:** March 6, 2023




https://doi.org/10.1021/acs.chemrestox.2c00379
*Chem. Res. Toxicol.* 2023, 36, 465−478

As a thermodynamic property, the value of water solubility depends on environmental conditions such as pressure and temperature.[5] It additionally depends on structural characteristics such as exposed van der Waals surface area, quantity of hydrogen-bond acceptors and donors, and acidity. These structural characteristics contribute to a mechanistically well-defined property, i.e., solubility of a compound in water, making it a promising target to model.

Multiple works exist exploring the development of quantitative structure−property relationships (QSPR) models for water solubility and have been reviewed within the literature.[6,7] Generally, works within this sphere focus on drawing out the highest performance via choice of molecular descriptor, regression algorithm, data sets, or other details involved in the QSPR development process.[8−10] The intent of this work is not to develop and present a novel model with the assertion it is superior to any of the dozens presented within the literature. This work is intended to abstract the presentation and development process and leverage existing guidelines to improve practices in transparency and reporting when publishing models to reflect the importance of these values to model adoption by third parties, particularly in light of the complexities introduced by sophisticated machine learning algorithms such as those used both in this work and in the papers cited.

There are several aspects as to why, even considering the modern revolutions of modeling algorithms, achieving regulatory acceptance for quantitative structure−property relationships (QSPR), QSPR including QSPRs for water solubility, remains elusive. The first is simple: the quality of data is too poor to provide a sufficiently strong chemical signal for any algorithm to learn. This aspect, at least for water solubility, has been explicitly addressed in the recent literature[11] with the creation of AqSolDB. In this study, we similarly leverage a large data set of curated points to correct for poor data quality to the best of contemporary ability. The second aspect is within the purview of modelers themselves, which involves a need to rigorously deconstruct and analyze the strengths and weaknesses of their own models to step away from the elements of machine learning that obfuscate its imperfections as a "black box".

A prescription for how to address this second challenge exists prototypically in the OECD's principles of model validation[12] and in the follow-up and more expansive "Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models".[13] The OECD principles were drafted and agreed upon by all OECD member countries with the expectation that the principles would provide a basis for evaluating (Q)SAR models and their predictions within chemical safety assessments. As a conceptual and general framework, the principles represented a major advance toward furthering appropriate reporting and regulatory consideration of QSARs and are stated as follows: "To facilitate the consideration of a (Q)SAR model for regulatory purposes, it should be associated with the following information:

1. a defined end point;
2. an unambiguous algorithm;
3. a defined domain of applicability;
4. appropriate measures of goodness-of-fit, robustness and predictivity;
5. a mechanistic interpretation, if possible".

Each OECD principle, in turn, subscribes to the reporting of types of information that are considered useful for the regulatory review and acceptance of QSARs. For example, Principle 2's intent is to ensure transparency in reporting the model algorithm used, while also allowing one to reproduce the calculation, if needed, since reproducibility is an important part of acceptance of QSAR models. Within the 2004 report from the Expert Group on the OECD principles itself,[12] a recommendation was made to update the guidance on the application of the principles in future work, especially with respect to "new considerations" and "how the principles can be applied for different types of (Q)SAR models."

Building on these stated principles, we attempt herein to explicitly deconstruct and further delineate their application to one of the most common machine learning approaches to QSAR/QSPR models: the random forest.[14] The motivation for this study arises from a need to reconsider the interpretation of some of the principles so that they can be robustly and routinely applied to algorithms that have been developed in the interceding years. This manuscript has therefore been structured according to the OECD principles, with each section intended to consider the interface between the intent of the principle and modern modeling approaches. Additionally, this work attempts to dispel the shroud of the "black box" that is often invoked as a means of distrusting or dismissing the interpretability of more modern modeling algorithms.

Some work has been done in the space of improving the process by which QSAR/QSPR models are developed, including Dearden, Dronin, and Kaiser's work remarking on common errors in these models and improving practices.[15,16] We build on this to emphasize the challenges involved with moving to nonlinear models and, rather than a point-by-point review of common errors, trying to holistically present a reporting and analysis workflow that is compliant with OECD's goals.

Herein, we train and deconstruct our modeling approach using random forests for a newly curated set of water solubility data. Our goal is to practically demonstrate how one can robustly characterize a model developed using random forest approach with respect to the OECD's validation principles.

## RESULTS AND DISCUSSION

**Principle 0: Characterizing the Data.** Implicit in the discussion of any data-first model (i.e., a model constructed using existing data as opposed to first principle calculations like density functional theory), is the commonly understood GIGO principle (i.e., garbage in, garbage out). In practice, this represents the need for due diligence and transparent reporting as well as analysis of the data set that will be used to build the machine learning model. While the original OECD principles did not call out a specific principle to capture the importance of data aggregation and curation prior to modeling, we contend that an additional principle to capture this requirement, "Principle 0", warrants further consideration.

Competing with the concern for data quality in building machine learning models is the need for sufficient data representation across the end point parameter space, in this case, the water solubility space. The challenge is managing the trade-off between predefined data quality thresholds to minimize noise and uncertainties, maximizing sufficient data representation to inform the model yet ensuring transparency and reproducibility. Clearly data quality is an important consideration for modeling, yet practical considerations require


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

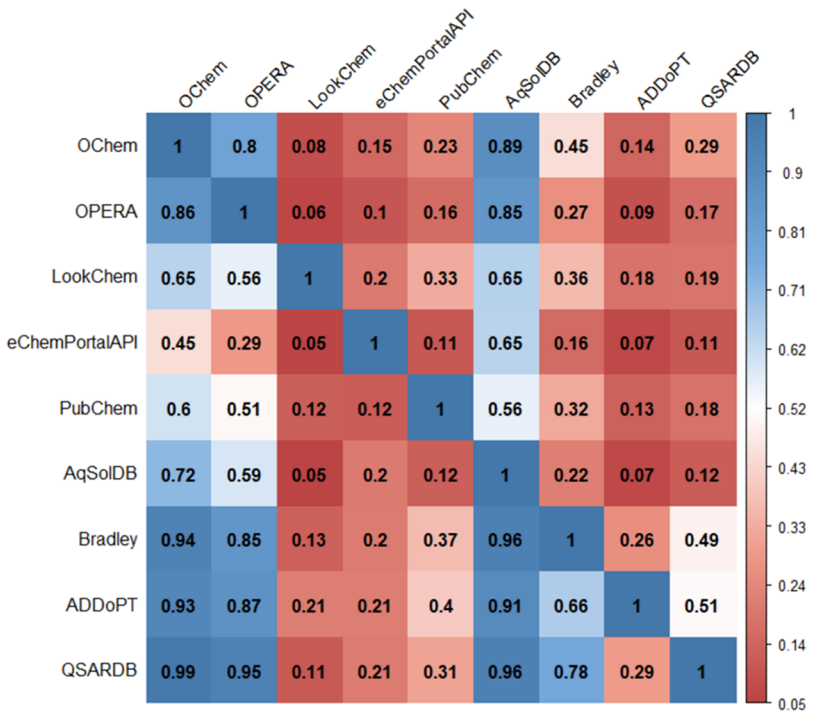



**Figure 1.** A redundancy matrix showing the overlap, $X_{ij}$, of chemical compounds in the $i^{th}$ data set with the $j^{th}$ data set, where $i$ and $j$ represent the row and column, respectively.

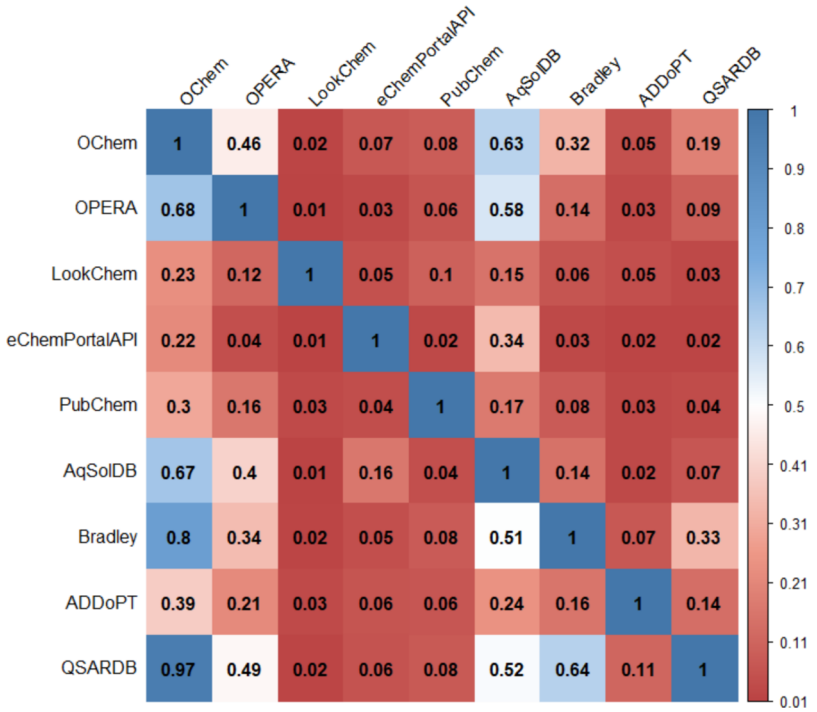



**Figure 2.** A redundancy matrix showing the overlap, $X_{ij}$, of chemical compound and water solubility measurement pairs in the $i^{th}$ data set with the $j^{th}$ data set, where $i$ and $j$ represent the row and column, respectively. Two water solubility measurements are considered to overlap if, for the same chemical, each is within 0.01 log10(mol/L) of the other.

some tolerance of imperfect quality measures. Most modeling in the chemical property or activity realm relies on the use of secondary data. This is usually compiled from multiple data sources, where measurement details may be missing, and it is often difficult to quantify the level of imprecision associated with data measures. Despite this, data quality measures and filters can be imposed, including establishing minimum data reporting thresholds and procedures for dealing with data inconsistencies.

Sorkun et al.[11] provided a convenient template for characterizing quality of water solubility measurements that included the consideration of ensuring chemical identifiers



https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

map to the same chemical structure (i.e., chemical name and CAS registration number). Briefly, this was done via a cyclic conversion between molfile and InChI key to ensure consistency. This process was not necessary in our curation efforts. In this study, as our method, the problem of structural and identifier consistency was obviated through using the US EPA internal chemical registration system ChemReg that underpins DSSTox.

Our data set considered the Sorkun's AqSolDB data set as well as a number of other sources that, after curation, afforded a much larger data set available for training our algorithm. The nine data sources and a brief description of each are as follows:

1. eChemPortal[17] is a database of physicochemical properties and toxicity measurements provided by the OECD.
2. Advanced Digital Design of Pharmaceutical Therapeutics[18] is a collaboration between pharmaceutical companies and academia to establish digital design approaches more usable for drug discovery.
3. AqSolDB[11] is a data collection resulting from work described in a publication by Sorkun et al.
4. The Bradley data set[19] is a collection of measured solubilities from the Open Notebook Science Challenge.
5. The online chemical modeling environment[20] is a physical property database and modeling platform with data sets provided by the users.
6. LookChem[21] is a global chemical trading platform which includes physical property values for advertised chemicals. A caveat with this source is that each entry lacks a citation, thus it was difficult to rectify if physicochemical properties were really measured or predicted.
7. QSAR DataBank[22] is a repository of QSA/PR models and associated data following the FAIR (findable, accessible, interoperable, reusable) principles.
8. PubChem[23] is an open chemistry database developed by the National Institutes of Health.
9. The OPEn structure−activity/property relationship app (OPERA)[1] is a collection of models and associated data developed by Mansouri et al.

There was considerable overlap among the chemicals contained within these data sets, although not necessarily among the water solubility measurements associated with these overlapping chemicals themselves. To illustrate the overlap of these data sets, redundancy matrices are plotted in Figures 1 and 2. In Figure 1, fractional values are provided showing the overlap, $X_{ij}$, of chemical compounds in the $i^{th}$ data set with the $j^{th}$ data set, where $i$ and $j$ represent the row and column, respectively. This shows that a number of data sets intend have high overlap in terms of similar chemical compounds. For example, 60% of the PubChem data set overlaps with the OChem data set. In Figure 2, the matrix captures the overlap of each water solubility measurement, expressed as log10(mol/L) rounded to two decimal places, and the chemical compound as a pair. Two water solubility measurements are considered overlapping if, for the same chemical, each is within 0.01 log10(mol/L) of the other. For example, if benzoic acid has a measurement of −1.61 log10(mol/L) in data set A and −1.39 log10(mol/L) in data set B, those chemical/measurement pairs would not overlap; however, in the case where data set B's measurement was −1.60, the pairs would be considered overlapping. Note that both redundancy matrices are not symmetric due to the varying sizes of the data sets. The sample median of the solubility values for each chemical was used for modeling water solubility owing to the low overlap of chemicals with the same water solubility measurement across the data sets.

For the collection of data, only records represented as mass solute per volume of water or in molar units were retained, excluding those with mass or volume percentage units. The strategy for parsing the property strings varied depending on the source, but all nonmolar unit records were standardized to units of grams per liter. Experimental conditions and all compound identifying information were preserved to ensure that measurements were taken at ambient conditions (*vide infra*) and structures were faithfully mapped.

A priority of many regulatory agencies is to ensure public transparency of the data and models underpinning those regulatory decisions. The US EPA's efforts have historically disseminated a number of software tools (standalone and web-based) as well as databases to serve this need. The US EPA's CompTox Chemicals Dashboard, https://comptox.epa.gov/dashboard/,[24] serves as a hub for chemical exposure and toxicity information. The database that contains the chemical substance (and associated structural data when available) served by the Dashboard is the Distributed Structure-Searchable Toxicity (DSSTox) database.[25] Each chemical contained in DSSTox is mapped to a unique identifier, referred to as a DSSTox Substance Identifier (DTXSID) which, in turn, can be mapped to a structure identifier (DTXCID) where appropriate, consistent with the principles outlined by other researchers such as of Fourches et al.[26]

Each record was mapped to a DTXSID using any available chemical identifier(s) (e.g., chemical name and CAS-RN), as well as structural information encoded as SMILES. Of the 75,319 records complied from among the 9 data sources, 41,662 were mapped to DTXSIDs. Records that did not map to DTXSIDs were a combination of those not currently contained in DSSTox, biologicals (e.g., polypeptides), and chemicals lacking discrete structures. Property values were required to be sensible quantitative measurements, ranging between 0 g/L and 1000 g/L. Unit conversion was performed to standardize solubility values to units of molarity, mol/L, using molecular mass values from DSSTox. If units of measure were deemed ambiguous (i.e., matching the chemical's solubility in another solvent), the record was excluded. If water temperature information was available, then records outside of the range of room temperature, 20−30 °C, were excluded from further consideration. Records with reported pressures outside the range of 740−780 mmHg were also excluded. For the vast majority of records, >75%, where these details were not provided, an assumption was made that intrinsic solubility had been measured within these normal parameter ranges.

Records needed to be mapped to a single organic molecular structure for inclusion into the modeling data set. Unknown, variable composition, complex reaction products or of biological materials (UVCBs), as well as inorganic and salt-containing compounds were excluded from the modeling data set. While the organic portion of a salt can be modeled by QSAR, the salt compounds in this data set were found to vary by >0.5 log(mol/L) on average when compared to measurements of the parent chemical. Given this would have introduced considerable variability if collapsed to the parent measurement, these were excluded from the modellng data set. Ultimately, a total of 10,717 unique chemicals were collected among 39,853 records. Thus, just under 2000 additional


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

records were removed for the above reasons. The original counts and curated counts of records among the nine sources are provided in Table 1.

**Table 1. Comparison of the Original Number of Records for Each Listed Source versus the Final Number of Curated Records That Were Included in the Modeling Work Described Herein**

| Source Abbreviation | Short Description | Original No. Records | Curated No. Records |
|---|---|---|---|
| ADDoPT | Advanced Digital Design of Pharmaceutical Therapeutics | 1484 | 761 |
| AqSolDB | Aqueous Solubility Database | 9959 | 7408 |
| Bradley | Data set curated by Bradley et al. | 3948 | 2493 |
| eChemPortalAPI | OECD chemical substance database | 8040 | 3752 |
| LookChem | Chemical trading platform | 6035 | 532 |
| OChem | Online chemical modeling environment | 28,683 | 16,864 |
| OPERA | PhysProp data set curated for OPERA | 5267 | 5084 |
| PubChem | NLM's chemical database | 10,800 | 2031 |
| QsarDB | FAIR repository of (Q)SAR/QSPR models | 1103 | 928 |
| Total Records: | | 75,319 | 39,853 |

**OECD Principle 1: Defining the End point.** This principle is meant to precisely bound and clarify the outputs of a QSPR model. An extension of this principle, in particular for modelers, would be to provide technical justification for why a specific property was a suitable candidate for QSAR/QSPR modeling. Prototypical arguments would address the mathematical suitability of the end point, the theoretical chemistry mechanisms relating structure to the property where known, the signal-to-noise characteristics of the data, as well as any supportive reasoning which would provide key context of use for the end-user. The requirements of OECD Principle 1 being proposed here involve substantiating the suitability of the end point before modeling is attempted, thus providing a theoretical grounding for why the approach is expected to be successful.

In this work, the log-transformed water solubility measured in mol/L was modeled. The exact environmental conditions for each end point value were not always known due to the incomplete reporting of the data. As described earlier, our curation process rejected measurements taken outside the temperature range of 20−30 °C; however, in over 75% of the cases, temperature was not reported, requiring the acceptance of those records regardless of the true measurement temperature. Additionally, the pH of measurement was rarely reported. Lack of control of the experimental factors going into these measurements will inevitably contribute to the overall irreducible error (those outside of the modeler's control) of the machine learning model.

Beyond the quality of the end point data itself, OECD Principle 1 offers scope to articulate the theoretical foundations of the end point's suitability for modeling. This will become more important as the range of potential QSAR/PR model targets continues to evolve, with not all end points being made equal when it comes to data-driven learning of the chemical relationships between structure and property.

We contend that water solubility is a good candidate for machine learning modeling. Water solubility is a simple thermodynamic property derived from the equilibrium state in which the energetics of the solute within its own phases is balanced by the energetics of the solute being solvated by water. The Hamiltonian associated with this equilibrium can be abstractly separated into contributions from electrostatics, the enthalpies and entropies of hydrogen bonding, van der Waals surface area, and ionic solvation. This separability of the governing physics results in a hierarchical and separable set of underlying factors which ultimately generate the distribution of water solubilities. Further, because it is this Hamiltonian that governs the end point, species with high similarity will yield similar Hamiltonians. Hence, for this end point, a firm theoretical grounding supports the premise that similar structures will tend to yield similar water solubilities. These two qualities, hierarchical separation of explanatory factors and coherence of the end point between similar structures, are articulated by Bengio, Courville, and Vincent[27] as being prior qualities that make the relationship easier for representation learning algorithms to learn. In our next section, OECD Principle 2: Defining the Algorithm, these characteristics of the water solubility end point render it suitable for the algorithms we have investigated in this study.

**OECD Principle 2: Defining the Algorithm.** The definition of the algorithm is one of the two principles that most directly suffers from the shroud of the black box. Under the paradigm that a model need not be interpretable, there is typically minimal rationale given for why a specific algorithm was chosen short of oblique references to performance statistics that themselves often lack their own degree of transparency.

A more substantive approach to defining the algorithm is to offer supportive arguments as to why the method is considered suitable to modeling the end point. This has parallels to arguments over which spectroscopic method might be more suitable for ascertaining qualities of an analyte beyond the fact that the latter case concerns itself with extracting information for a physical sample and the former concerns itself with extracting information from data.

The random forest is a popular algorithm introduced by Breiman[14] for classification tasks. It has been extended to apply to regression tasks and, thus, is used in many QSAR/PR models. Briefly, the random forest generates an ensemble of decision tree regressors that are individually trained on ensembles of data bootstrapped from the training bulk. Because each tree does not see every single sample, overfitting is mitigated by the averaging process and more abstracted, general explanatory trends within the data will tend to dominate over local chemistry trends. This provides a framework for a general QSAR/PR model rather than a local one.

Decision tree regressors operate by iteratively producing splits in their training data based on logical criteria applied to their input descriptors. The decision trees as implemented by Breiman perform these splits by randomly considering some number of descriptors, typically approximately the total number of descriptors divided by three for regression forests, at each split, and then choosing the logical separations that result in the minimization of variance of the end point for the samples within the two child nodes. This quality of the algorithm in combination with the aforementioned coherence of end point for similar compounds is an argument for why this


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

particular approach may be more suitable for this end point than another. The process of minimizing variance based on the structurally descriptive inputs should result in structurally similar compounds hierarchically grouping together, which we know from *a priori* considerations should work as a means of grouping together similar water solubility end point values.

Important to this argument, however, is the choice of which structural features are fed to the model in the first place. The choice of inputs to the model is a decisive point in the modeling process at which the modeler can affect the ability to extract meaning from their results. From a very early point in QSAR's relationship with machine learning,[28] an approach was adopted in which large numbers and types of descriptors with largely unexamined relationships to one another were fed into a model, thus providing the model with as much structural description as conceivably possible. Such an approach may produce empirically high-performance statistics but comes at a cost to model interpretability.

As part of OECD Principle 2 for machine learning, it is important to articulate why a specific choice of descriptors was made, ideally in a way that is not linked to incremental improvement in a performance statistic. In this study, QSAR-ready SMILES were used as the basis to obtain both 1D and 2D molecular descriptors using the PaDEL-descriptor software.[29] PaDEL descriptors are a set of molecular descriptors that capture electrotopological properties, atomic and bonding counts, as well as group contribution models for end points like log*P* which are used extensively in the QSAR/QSPR modeling community. PaDEL descriptors have been widely used in QSAR and machine learning models.[1,30] Taking a semiheuristic approach in the selection of descriptors, our model used a combination of algorithmically defined and expert selected descriptors, with the two sets of 16 descriptors listed in Table 2. More complete definitions and references for these descriptors can be found in the Table S3. Although many PaDEL descriptors do not lend themselves to easy interpretation, most can be related to valid chemical concepts. Briefly, the autocorrelation category of descriptors assesses internal correlations within a series or interval relative to a given property; in this case, we filtered descriptor selections based on the specified property and its relation to water solubility mechanisms.

These descriptors were identified in the PaDEL library using the caret package[31] implementation of recursive feature elimination (RFE) in the R language[32] followed by expert review and selection. Briefly, RFE is a function that performs a backward selection of descriptors based on a descriptor importance ranking. In this work, the variable importance associated with a random forest was used. A plot of the root-mean-squared error (RMSE) associated with the inclusion of each descriptor, is shown in Figure S1. This plot shows that RMSE is reduced with the inclusion of an additional descriptor up to the 16th descriptor where no difference in RMSE is noted. This led to the initial 16 descriptors shown in the algorithmically selected descriptors column of Table 2. Considerations involving the theory of water solvation led us to introduce descriptors such as hydrogen-bond acceptors that we knew to be intimately connected to aqueous behavior as well as systematically remove algorithmically determined descriptors that could not be rationalized with the same theory.

The selection of descriptors was robust, in that each selection parametrized sufficiently distinct elements of structure such that no particular explanatory factor of the property was overstated or confounded with another. It should be noted that this process involved manually removing automatically selected descriptors, such as several articulations of structural autocorrelation that simply restated the same information in nonlinearly dependent ways. This "restatement" phenomenon is a known problem in QSAR and is typically handled by correlation filtering; however, due to the nonlinear capabilities of machine learning, a more general method of consideration is necessary. To generalize beyond measure linear correlations, we utilized normalized mutual information to capture nonlinear relationships, termed "entanglements", between variables used in the model.

Figure 3 displays the normalized information (see eqs 1.a−1.c) between these descriptors as heat maps, indicating several of the automatically selected descriptors encoded much more similar information than the expert-selected descriptors. This is a quantification of our assertion, showing that each descriptor contributes its own information to knowledge of the

**Table 2. Optimal Descriptors As Determined by Recursive Feature Elimination and by Expert Judgement[a]**

| Algorithmically Selected Descriptors | Expert Selected Descriptors | Abbreviated Descriptor Definition |
|---|---|---|
| XLog*P* | XLog*P* | Log(octanol/water partition coefficient) |
| GATS1s | GATS1s | Autocorrelation weighted by Ionization-state |
| GATS1m | GATS1m | Autocorrelation weighted by mass |
| SIC0 | SIC0 | Structural information content index |
| GATS1e | GATS1e | Autocorrelation weighted by electronegativities |
| piPC5 | piPC5 | Conventional bond order ID number of order 5 |
| piPC7 | piPC7 | Conventional bond order ID number of order 7 |
| ZMIC3 | ZMIC3 | Z-modified information content index |
| GATS1p | GATS1p | Autocorrelation weighted by polarizabilities |
| nAcid | nAcid | Number of acidic groups |
| MATS1e | *replaced* | Autocorrelation weighted by electronegativities |
| MATS1s | *replaced* | Autocorrelation weighted by Ionization-state |
| AATSC1e | *replaced* | Autocorrelation weighted by electronegativities |
| ATS3m | *replaced* | Autocorrelation weighted by mass |
| ATSC1e | *replaced* | Autocorrelation weighted by electronegativities |
| ATSC2e | *replaced* | Autocorrelation weighted by electronegativities |
| | nHBAcc | Number of hydrogen-bond acceptors |
| | nHBDon | Number of hydrogen-bond donors |
| | GATS1i | Autocorrelation ionization potential |
| | GATS2e | Autocorrelation weighted by electronegativities |
| | GATS1v | Autocorrelation weighted by van der Waals volumes |
| | GATS1c | Autocorrelation weighted by charges |

[a]Note that the algorithmic selection process informed the experts' judgement.


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

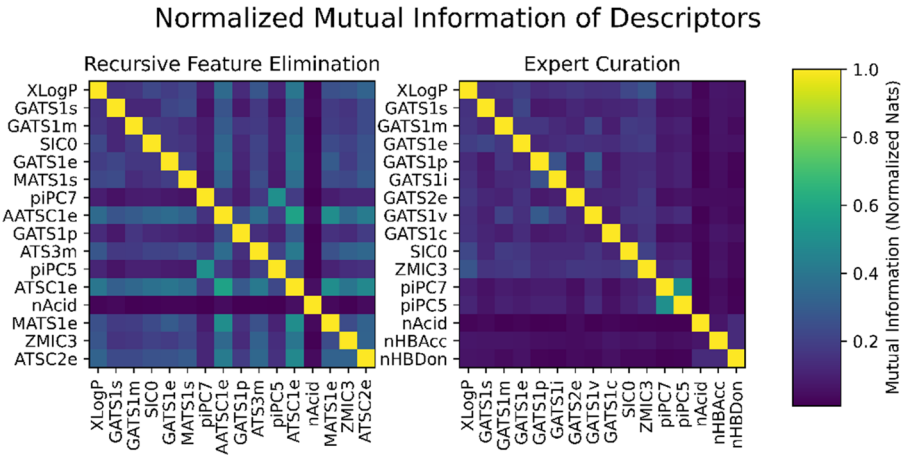



**Figure 3.** Plots of the normalized mutual information for two sets of PaDEL descriptors. The first (left) selected automatically through recursive feature elimination, and the second (right) replacing 6 of the descriptors with expert-selected features; lighter intensity squares indicate higher correlation or greater degree of information overlap between descriptors.

end point distribution. We achieved this via mutual information calculations[33] that can more robustly detect informatic entanglements than the typical standard of linear correlation.[34]

Hence, expert-intervention guided by automatic recursive feature elimination to give increased weight to less colinear and more mechanistically relevant descriptors can lead to increased descriptor-end point clarity and model interpretability. We contend that this can be achieved without sacrificing model performance, as evidenced by the equivalent performance of both approaches. Such a semisupervised approach can help prevent a model from falling into a nonlinear variant of the concerns that historically have related to the inclusion of highly colinear descriptors in a model. It bears mention that the most important descriptor, XlogP, is itself an indirect estimate of water solubility derived from an additive model.[35] XlogP quantifies the relative solubility between an octanol phase and an aqueous phase, whereas "water solubility" quantifies the relative solubility between an aqueous phase and solute bulk phase. Exploiting the thermodynamics relating these partition coefficients invokes a form of transfer learning. The simplistic additive model that generates the descriptor transfers that information into the machine learning algorithm, which is then able to refine that information using the additional descriptors. Relatively successful models of water solubility have been purely additive, and incorporation of these simple models' successes into more sophisticated algorithms represents one of the powers of machine learning, i.e., to transfer strengths of certain models while mitigating their weaknesses through the inclusion of other qualifying pieces of information.

Figure 4 provides a measure of the relative contribution of each descriptor to the end point based on normalized information content. Normalized information content ($MI_N$) is computed according to eqs 1.a−1.c:

$$MI_N(D, R) = \frac{MI(D_i, R_j)}{\text{mean}[H(D_i), H(R_j)]} \quad (1.a)$$

$$MI(D, R) = P(D_i, R_j)\ln\left[\frac{P(D_i, R_j)}{P(D_i)P(R_j)}\right] \quad (1.b)$$

$$H(D) = -P(D_i)\ln[P(D_i)] \quad (1.c)$$

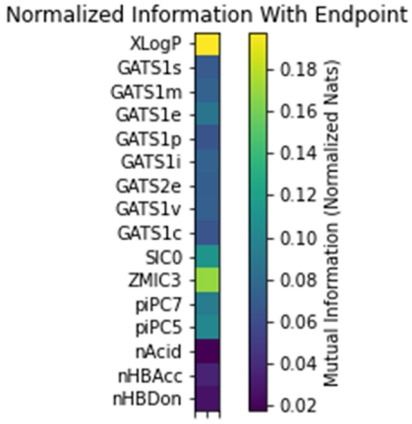



**Figure 4.** Calculation of the degree of informatic entanglement between descriptors and end points shows the major contributors to the model.

Sums over repeated indices are implicit.

In the above equations, $D$ and $R$ are the discretized sets of variables being compared, where $X_i$ indicates the $i^{th}$ bin of the discretized set for variable $X$. $P(X_i)$ is the probability a value occurred in the $i^{th}$ bin. Bins were created by histogram, where the number of bins was approximated using the Friedman−Diaconis rule to approximate a bin count that accurately captured the distribution of the data. $H(X_i)$ captures the entropy of the data, integrating the native uncertainty associated with a distribution into the degree of entanglement. These mutual information metrics are generalized measures of how related two observations are, moving beyond the linear relationships captured by "correlation" and thus acting as more appropriate measures of "entanglement" when considering the capabilities of nonparametric and nonlinear models such as those used in modern machine learning.

It is conspicuous that the acid group count, number of hydrogen-bond donors, and number of hydrogen-bond acceptors make relatively little contribution. One explanation of this is that the range of these discretized descriptors is quite small compared to the continuous range of water solubilities, and thus, there are limits to how much information those descriptors can encode. Despite this, they are included because of their explicit theoretical relevance to the end point and the


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

Table 3. Performance Metrics for 3 Models, Including Statistics for 5-Fold Cross Validation of the Training Data, External Prediction on a Test Data Set, and Prediction on the Excluded Portion of the Dataset

| | Training Data | | | Test Data | | | Excluded/Added Functional Group Test Data | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Data set | $R^2$ 5CV | RMSE 5CV | Size | $R^2$ | RMSE | Size | $R^2$ | RMSE | Size | $R^2$ | RMSE |
| Entire Data set | 0.82 | 0.96 | 8037 | 0.97 | 0.41 | 2680 | 0.82 | 0.97 | | | |
| bond:C(=O)O_carboxylic_ Ester_alkyl (*Excluded*) | 0.82 | 0.96 | 7236 | 0.97 | 0.4 | 2415 | 0.82 | 0.98 | 1066 | 0.78 | 1.11 |
| bond:C(=O)N_carboxamide_ (NHR) (*Excluded*) | 0.82 | 0.96 | 7329 | 0.97 | 0.41 | 2454 | 0.82 | 0.97 | 934 | 0.74 | 1.14 |
| Entire Data set (Metallics *test set*) | 0.82 | 0.96 | 8037 | 0.97 | 0.41 | 2680 | 0.82 | 0.97 | 116 | 0.31 | 2.57 |

understanding that the other descriptors are not encoding similar information. Because of their known mechanistic significance to the hydrogen-bond network, they are kept with the understanding that their low contribution to normalized information is likely an artifact due to the limitations of the normalized information quantity when comparing discrete and continuous distributions with the Friedman−Diaconis binning process we implemented.

By referencing the mechanisms of the machine learning regression algorithm and reconciling them with the resultant descriptors of the structure-end point model, we can formulate an argument for why our specific choice of algorithm is appropriate for the task at hand with some level of reasoned justification.

The final *a priori* consideration before training the model is the applicability domain of the model.

**3. OECD Principle 3: Defining the Applicability Domain.** The third principle is an articulation of the region of inputs in which the model predictions are expected to be reliable. Gramatica et al.[36] state that the applicability domain is a property of the training data for QSAR models, which is intuitively sensible as one cannot expect a model to understand the behavior of chemistries it has not observed. For instance, a water solubility model trained exclusively on isomers of alkanes will likely determine that branching is a significant feature in predicting the end point, where branching is captured by Randić's zeroth chi index.[37] Generalizing this observation to branched structures that are not simple alkanes, however, will likely result in significantly reduced or compromised signals due to the lack of correspondence between branching and simple van der Waals surface area, and thus applying this model outside its applicability domain of alkanes is ill-advised.

Most machine learning algorithms, by design, are intended to extract generalized factors explaining properties of interest. The random forest achieves this by bootstrapping its individual trees, the support vector regressor by interpolating its support vectors, the gradient boosted trees by limiting the depth of their constituent estimators, and so-forth. A commonly used algorithm that is an exception to this is the k-nearest neighbor (kNN) regressor, which utilizes a simple similarity statement to take averages for neighbors of like compounds. Mansouri et al.[1] produced state-of-the-art kNN models for QSPR end points with OPERA.

OPERA's approach to providing an applicability domain is algorithmic, using a similarity calculation like the algorithm's similarity measure to determine if the neighborhood assigned to a query compound is objectively similar. Because applicability domain remains a somewhat unsolved problem within modeling, a full exploration of improvements to this principle requires its own future work. For the water solubility model at hand, we can remark on several empirical observations regarding our method that inform our applicability domain.

Table 3 shows the results of applying our algorithm to several different training and test sets of compounds. In each case, 5-fold cross validation was applied to the training data set, and the model was applied prospectively to an external test set. As mentioned in the Principle 0: Characterizing the Data section, we considered only organic chemicals in building our model. Hence, we removed all compounds containing metallic atoms (specifically Hg, Fe, Co, Si, Te, Cu, As, Zn, Ni, Pb, Sn, Ge, Sb, Se, Cd, and Ti) due a variety of factors, including multiple valence states and bond attributes resulting in lack of sufficient representation relative to the more conventional organic atoms. In the first row of Table 3, performance statistics for the random forest model are reported for the entire 10,717 water solubility data set, split into training and test data sets, yielding a respectable $R^2$ of 0.82 on the external test data set. Structures encoded as SMILEs and values of all 16 model descriptors for each of the chemicals in the training and test data sets are provided in the Tables S1 and S2. For the next two examples, we defined local domains of functional chemistry by use of public ToxPrint features.[38] In the first example, a carboxylic ester moiety (attached to an alkyl group) is present in 1066 chemicals, which were excluded from the training and test data sets; the subsequent model was used to predict water solubility values for the excluded chemicals, with very little degradation in performance ($R^2$ = 0.78). We performed a similar experiment for the ToxPrint domain of carboxamides, yielding a slightly worse performance ($R^2$ = 0.74). In the last example in Table 3, we assessed the local performance of the model using the entire data set for predicting water solubility of the subset of excluded metallics only. We saw a significant degradation of model performance, to $R^2$ = 0.31, when predicting within the metallics subdomain. Hence, we can consider the metallic atom-containing compounds lie firmly outside the applicability domain.

However, we can additionally show that even in cases where a specific organic functional group [carboxylic esters (defined by the ToxPrint bond:C(=O)O_carboxylic_Ester_alkyl) and carboxamides (defined by the Toxprint bond:C(=O)-N_carboxamide_(NHR)] is singled out and removed, our model generalized sufficiently to do a good job of predicting within those local domains. This is because the underlying chemistries likely to govern the major water solubility properties of these unique moieties, hydrogen-bond acceptor and donors as well as polarizabilities and charge distributions, are captured by the chosen descriptors in Table 1. Specifically, the relevant Geary autocorrelations and explicit counts of hydrogen-bond acceptors and donors transfer information learned from similar, but not identical, organic functional groups.


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

This evidence suggests that our applicability domain should exclude metallic or metalloid atoms. Our inputs, or molecular representations, were chosen to handle molecules whose water solubility properties are best explained by their ability to interact with the hydrogen-bonding network of the surrounding water or based on motifs of neighboring atomic properties captured by the Geary autocorrelations. Metallics likely would involve considerations of valency or other molecular properties that are not sufficiently represented within the data set to reliably argue their proper internalization of the model even if we were to attempt to learn from them.

We quantitatively implement a similarity calculation to characterize the suitability of applying a standard mathematical applicability domain method to the model. Figure 5 shows the distribution of the OPERA Local Index[1] calculated for the training and test sets of the data. This shows a bimodal distribution, indicating that while there are concentrations of compounds with highly similar analogues (smaller peak to the right) there are many that possess relatively low similarity (left side of figure). Hence, whereas a concentration of highly similar points exists, the major mode of the distribution suggests most points do not have direct analogues within the training set based on the five nearest neighbors.

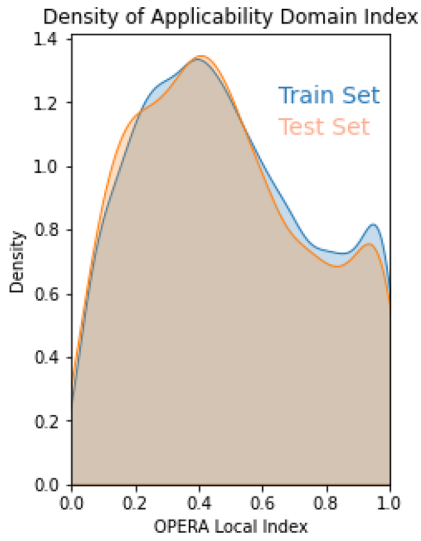



**Figure 5.** Distribution of OPERA Local Indices, a similarity measure that can be adopted as a quantitative applicability domain criterion, shows a diversity of well represented and poorly represented points through the data.

The performance curve in Figure 6 demonstrates that a thresholding value of the local index could be used to improve the statistics of the model. This form of reporting the applicability domain gives a more comprehensive sense of the trade-off between coverage and performance for the published architecture and thus promotes transparency in allowing users to understand the continuous nature between the selected quantitative applicability domain and its effect on the model.

**OECD Principle 4: Internal and External Validation.** Internal validation is one of the most well-explored areas of model validation in the literature. As such, many tools already exist for performing robust internal validation on machine learning models in the general sense, with the details left to QSPR experts involving robust discussion on how these tools differ between more general machine learning tasks and those specific to the structure−property domain specifically.

The key concepts to be examined within this section, and the next, are the precise meanings of "internal" and "external" in the chemical sense. The conventional wisdom is that the internal data set is that which is used for training and the external data set is excluded at the beginning of the process to provide a true, blind measure of how well the model generalizes to unseen data.

A global chemical model is typically expected to convey a model that has learned an abstracted chemistry from the training data that governs the behaviors of all chemicals within a broad applicability domain. A proper global model that has appropriately leveraged machine learning to generalize and, therefore, avoided being overly leveraged by local domains is expected to predict well not only within the areas of highest density of its trained chemical space but will also be able to

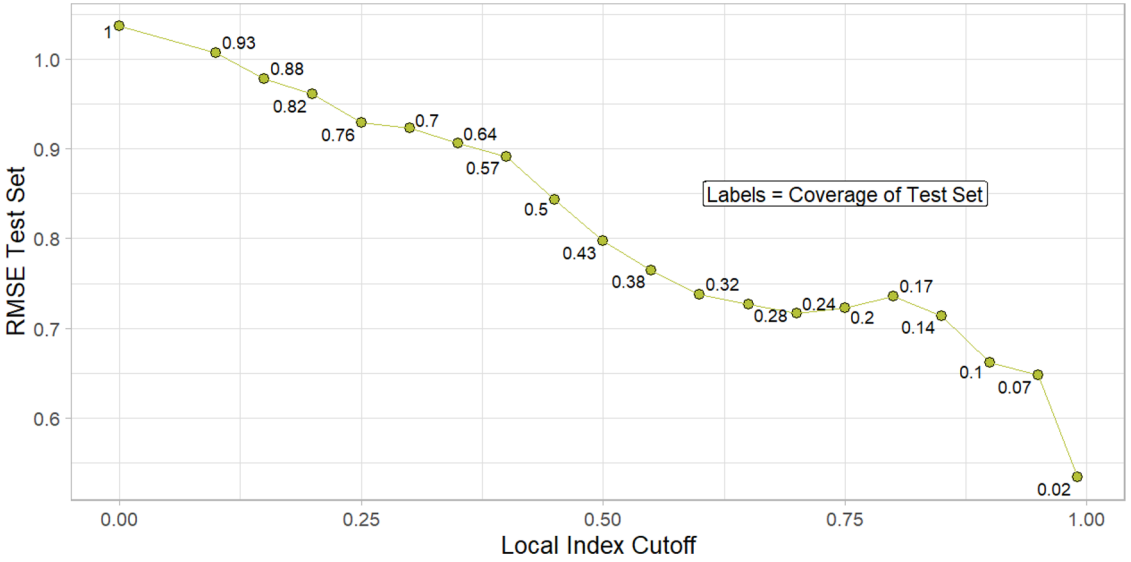



**Figure 6.** A plot of applicability domain cutoffs using the local index and its effect on performance via RMSE. The fractional coverage of the test data for each local index cutoff is provided as labels on the plot.


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

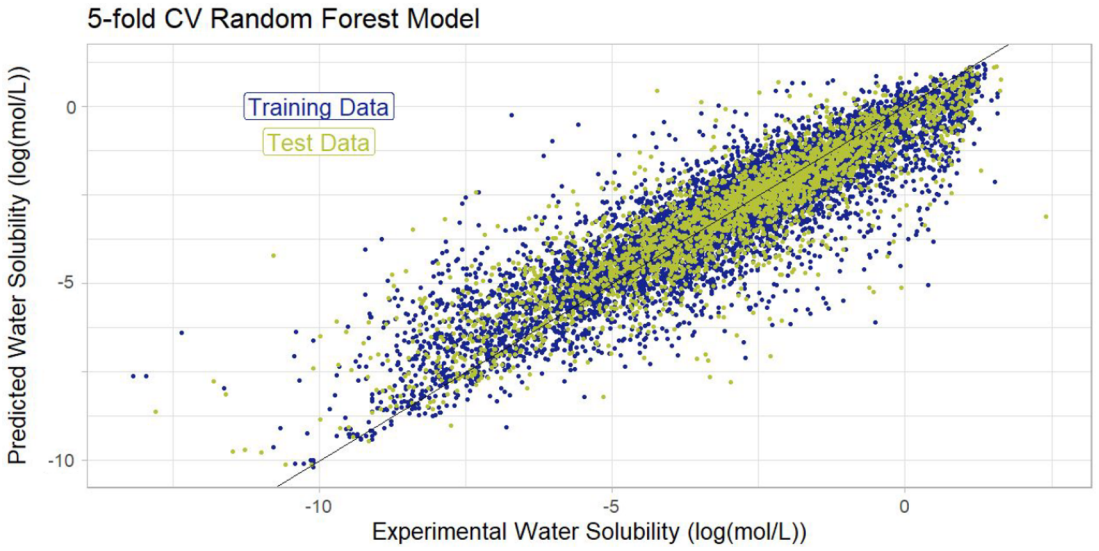



**Figure 7.** A plot of predicted water solubilities against their experimental water solubility values.

interpolate to regions that are more sparsely sampled by transferring the general chemistries within the applicability domain.

Thus, comes the subtlety of the QSAR/QSPR formulation of internal and external validation. A randomly drawn external set cannot be necessarily declared truly external; whereas those specific points may not be within the training space, the random selection process may cause the "external" data set to overrepresent local chemistries and thus overestimate the global capabilities of the model. Research has shown that the process for splitting test and training sets[39] can have an impact on the test statistics while not informing external predictivity. Further research is needed to rigorously address the influence of constructing the external set to be chemically external. However, the results of rational design of the training set do suggest that this phenomenon interferes with many of the performance statistics cited as the end-all measures of judging a model's predictivity.[39]

The Table 3 experiment on the fourth row showed that the model failed to predict on metal-containing compounds, which informs the applicability domain. Based on this, one can surmise that this model is not suited to chemicals outside the domain of small organic molecules containing CHNOPS (carbon, hydrogen, nitrogen, oxygen, phosphorus, and sulfur). In addition, the removal of particular functional moieties represented by ToxPrints in Table 3, i.e., in effect creating external test sets, demonstrate the subtlety in considering "external" and "internal" validation when considering structure−property relationships specifically.

Random forests possess two relatively rare properties that factor into interpretation of the data. The first is that random forests are ensembles of "partially blind" simple estimators, where "partially blind" refers to the fact that each tree is only exposed to a fraction of the training set during learning. These simple estimators are decision trees. The second is that random forests, in some sense, "memorize by design". What is meant by this is that the constituent learner decision trees are not pruned, resulting in trees that "memorize" exactly the training set samples to which they had access. These two properties result in random forests requiring more nuanced discussion around their validation to ensure transparency.

It is conventional wisdom that models that memorize their data−that is, achieve training scores substantially higher than their external scores−have internalized noise and erroneous patterns that overall lower their predictivity. Random forests, however, will generally produce training performances close to perfect prediction due to the majority of their decision trees having exact representations of the training set. Thus, unlike many algorithms, the random forest requires internal cross-validation, such as that used to generate the training predictions visualized in Figure 6, as well as out-of-bag estimates, which are summary statistics generated by considering the decision trees in the ensemble that were not exposed to a training point due to bootstrapping. Agreement between cross-validation statistics and out-of-bag statistics serves as internal validation that the forest has low variance across the chemical space comprising the training set. This has parallels to how one looks for similar performance between training prediction and testing prediction for models such as when using support vector machines or least-squares regressions.

Table 3 and Figure 7 summarize tests performed for the current model. A refinement of OECD Principle 4 would rely on the consideration of the goals of most machine learning practitioners—generalization and predictivity—and how they interface with the theoretical underpinnings of chemical space. It is the purview of modelers to report validation tests that accurately approximate the different meanings of the word external to help inform potentially naïve users on the limitations of the published model. It bears mentioning that modeling a chemical phenomenon is simply not sourced in the same theoretical underpinnings as a classic benchmark task such as the iris data set classification[40] or prediction of California housing prices. While data tasks are often highly analogous, the underlying structure of the explanatory factors, or descriptors, heavily influences the behavior of the model. Any one particular or small subset of validation statistics are generally too limited to properly characterize whether a model has adequately ascertained underlying chemical principles to anchor the model in plausibility, or a rote pattern in the data that overestimates the prominence of a motif due to overrepresentation of some local chemistry within the data.[41]


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

**5. OECD Principle 5: Proposal of Mechanism.** OECD Principle 5 is likely the greatest casualty of the black box paradigm of machine learning, as data-fueled innovations like molecular descriptor sets, automatic feature reduction, and increasingly complex algorithms have fed into the notion that many modern and well performing models are simply beyond scientific comprehension.

The authors, echoing the spirit of Principle 5, adopt the perspective that a model that cannot be tied to plausible mechanistic underpinnings of the end point will fail to overcome persistent historical barriers to adoption and usability, particularly by chemists and regulatory scientists, where plausibility arguments carry increased weight. Providing such a mechanistic connection offers a powerful complement and enhancement to purely statistical measures in these contexts. This claim can be mitigated, to be certain, not only by the mechanistic clarity of the end point, but also by the size and chemical coverage of the end point data set being modeled. Modeling of a physicochemical (as opposed to biological) end point, such as water solubility, which can be considered to be largely governed by intrinsic properties of the chemical if environmental conditions are controlled, is a substantially different and potentially tractable challenge than modeling of biological end points. Not only are the available data sets for modeling the latter typically much smaller, with corresponding limited coverage of chemical space, but also biological end points, such as various types of toxicity, are potentially reached through multiple biological pathways and mechanisms (a different type of black box) that can respond to underlying chemical mechanistic drivers in potentially different ways. Prediction of enzyme target-mediated, whole animal adverse outcomes provide a classic example of this added complexity. Hence, the requirements of Principle 5 should be calibrated to reflect the nature of the end point being modeled, and the size (and quality) of the available end point data set. Where biological mechanisms are largely obscured and intractable, greater weight should be placed on plausible chemical reaction mechanisms and knowledge of how such chemicals act within biological systems. Principle 5 therefore remains no less important in the machine learning era than it was in the era of linear regressions and additive models. It warrants mention that if proper attention has been paid to argumentation in each of the prior four principles, then Principle 5 has likely written itself.

Our water solubility model is a random forest model that has been trained on a curated set of over ten-thousand chemicals and their associated data points. Consideration of the theoretical energetics that govern the equilibrium concentrations of a solute in an aqueous phase informed the selection of model inputs (see Table 1). Each of these descriptors is firmly grounded in the energetics of water solubility, and their appearance in the automated process' list of important descriptor reinforces this intuition. We know from the importance associated with XLog*P* that the major force of the determination arises from the information contained in the XLog*P*'s progenitor additive model. This is refined by the inclusion of other theoretically meaningful aspects of structure that the additive model may erase or fail to capture, such as local motifs captured by lag one autocorrelations in the Geary coefficients (i.e., patterns in the neighboring atom which encode functional groups such as alcohols, ketones, etc.), the presence or absence of certain topologies captured by the path counts, and explicit energetics from hydrogen bonding due to solute donor and acceptor moieties.

The random forest's mechanism is not completely opaque, and thus, it is a good candidate for our modeling effort. A random forest is a consensus estimate derived from many decision trees performing internal separations of the training data to develop "neighborhoods" of water solubility values based on the input descriptors. Because we have asserted the theoretical meaning of each of the possible features that can generate these neighborhoods, we can infer that the clustering of these structures results in neighborhoods with theoretical justification for possessing similar water solubilities. This, in turn, is mechanistically sourced in the fact that small perturbations in the energetics of one solvent−solute interaction should result in a relatively small perturbation of the equilibrium concentrations. We see for a (Principle 1) suitable end point with a (Principle 2) well-defined and intelligently motivated algorithm that is being applied within a (Principle 3) well-characterized domain of explanatory factors that it is an achievable goal to synthesize the theoretical model of the structure−activity relationship with the algorithm and selection of inputs to produce a coherent explanation of how the machine learning model produces the performances observed in internal and external validation. The five OECD principles are linked in that if each is practiced appropriately, the shroud of the "black box" should at least be mitigated enough that all are achievable, even accounting for the modern complexity of available algorithms.

## CONCLUSIONS

We have presented a water solubility model trained on the largest curated set of water solubility data to date and considered the performance of a machine learning algorithm, i.e., random forest, upon this data set. We recognize the prevalence of water solubility models in the scientific literature as illustrated by a Google Scholar search of "water solubility model" OR "aqueous solubility model", showing an average of roughly 10 published models since 2017. To distinguish the present effort, we accompany the release of this model with a deconstruction of the model framed to modernize the five OECD principles of QSAR/QSPR model validation such that we can rigorously apply them to contemporary challenges introduced by the increasing complexity of performative machine learning algorithms. The value in this process is greatly diminished if the ultimate model is not made publicly available. To this end, this model and the ability to generate predictions from it will be made available in a future release of US EPA's CompTox Chemicals Dashboard (where both open-source models such as OPERA and commercially available model predictions are already available) or a similarly available US EPA webtool.

We started by introducing Principle 0 in which we provided a detailed report of the process by which our model data set was compiled and curated with respect to our compilation process. This introduces into the OECD validation framework a reporting protocol that gives due emphasis to the importance of the data quality in assessing the final model.

Next, we addressed the rigors of identifying, by combination of theoretical chemistry and the theories of machine learning, whether our declared end point was suitable for machine learning. This extended Principle 1 and appreciates that machine learning, while powerful, is subject to dependencies on the underlying structure of the data that render the


https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

algorithms reliable at extracting truly meaningful trends versus simply memorizing high local patterns in the data. We additionally defined and discussed our exact algorithm—the random forest regressor—and reconciled that decisions with the known aspects of what explanatory patterns our model hoped to capture from that data.

Addressing Principle 3 and the need for an applicability domain, we discussed the selection of inputs and nature of the training set as the determining factor in our model's applicability being restricted to small organic molecules containing no metallic or metalloid atoms. We demonstrate this by showing that our model is capable of transferring information on water solubility to unseen, simple organic functional groups but breaks down upon being asked to characterize the water solubility of organometallics. Modernization of Principle 3 requires reflection of the purpose of most complex machine learning algorithms, their capability to abstract and generalize trends in data, while understanding certain chemistries remain beyond transfer learning and thus exclude certain compounds from even a sophisticated model's applicability domain.

The rigorous intents behind defining an internal and external set of data for validation was examined in OECD Principle 4, with a careful distinction drawn between the idea of a data point being external from the training set versus a data point being external to the highest density regions of chemical space. Validation approaches that account for these chemical considerations were included in the analysis of our model, and discussion centered on why external validation is a uniquely posed challenge within the QSA/PR domain when compared to seemingly isomorphic problems in other fields.

Finally, OECD Principle 5 was revisited and addressed in terms of proper appreciation of the previous four principles. With thorough validation as practiced by Principles 1−4, it is unlikely a modeler is so deprived of insight into their proposed model that they cannot address some sort of mechanism by which their carefully chosen algorithm suitably describes known theory about their end point. OECD Principle 5 is presented as a final check that the system being modeled is sufficiently understood that a model of the phenomenon is advisible. With the proliferation of machine learning models and common software packages such as R and scikit-learn making the development of statistical models easier than ever, it is worth revisiting as a QSAR/QSPR community whether we can successfully transcribe the spirit of the original validation principles put forth by leading experts into a modern paradigm. The machine learning algorithms of the new decade are increasing complex and powerful, but with this increasing complexity comes greater shrouding of the underlying chemical mechanistic drivers of the modeled end point. We have attempted to show that it is possible to take advantage of features of the model algorithms, in this case random forests, to empirically guide models toward more mechanistically relevant descriptors that can ultimately enhance model plausibility and interpretability, which increases the probability for uptake and use of the models. Harnessing the power of these newer modeling approaches is an invigorating reason for theorists and modelers to reconcile these techniques with verified intuitions and theory.

As larger data sets become available, more sophisticated machine learning will be needed to make sense of them. The practice of the OECD principles as laid out in this work will hopefully provide a framework for future researchers to continue publishing transparent, trustworthy models that can be linked back to the theoretical underpinnings of model targets. This ensures that the adoption of these algorithms within the scientific community continues to produce robust predictions that are defensibly suited for inclusion in cost-reduction of commercial research cycles, academic expansion of chemical knowledge, and regulatory action by governing bodies.

## ■ ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.chemrestox.2c00379.

Figure S1 (PDF)
Tables S1−S4 (XLSX)
QMRF for the water solubility model discussed in this work (PDF)

## ■ AUTHOR INFORMATION


### Corresponding Authors

**Charles N. Lowe** − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0001-9151-6157; Email: lowe.charles@epa.gov

**Nathaniel Charest** − *ORAU Student Services Contractor to Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States*; Email: charest.nathaniel@epa.gov

### Authors

**Christian Ramsland** − *ORAU Student Services Contractor to Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States*

**Daniel T. Chang** − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States*

**Todd M. Martin** − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0001-5844-8754

**Antony J. Williams** − *Center for Computational Toxicology and Exposure, Office of Research and Development, United States Environmental Protection Agency, Research Triangle Park, North Carolina 27711, United States;* orcid.org/0000-0002-2668-4821

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.chemrestox.2c00379


### Author Contributions

[§]These individuals contributed equally to this work. CRediT: **Charles N. Lowe** conceptualization, data curation, formal analysis, methodology, software, writing-original draft, writing-review & editing; **Nathaniel Charest** data curation, formal analysis, methodology, software, writing-original draft, writing-review & editing; **Christian Ramsland** data curation, writing-review & editing; **Daniel T Chang** formal analysis, supervision,



https://doi.org/10.1021/acs.chemrestox.2c00379
Chem. Res. Toxicol. 2023, 36, 465−478

writing-review & editing; **Todd Matthew Martin** conceptualization, data curation, methodology, writing-review & editing; **Antony John Williams** conceptualization, data curation, methodology, supervision, writing-review & editing.

### Notes

Disclaimer: The views expressed in this manuscript are solely those of the authors and do not represent the policies of the US EPA. Mention of trade names of commercial products should not be interpreted as an endorsement by the US EPA. This work has been internally reviewed at the US EPA and has been approved for publication.

The authors declare no competing financial interest.

## ■ ACKNOWLEDGMENTS

We appreciate our colleagues Grace Patlewicz, Ann Richard, and Barbara Wetmore for their feedback and comments on the manuscript. We also acknowledge the tremendous contributions of Gabriel Sinclair, without whom this work would not have been possible. The information in this document has been funded wholly or in part by the U.S. Environmental Protection Agency.

## ■ REFERENCES

(1) Mansouri, K.; Grulke, C. M.; Judson, R. S.; Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *Journal of cheminformatics* **2018**, *10*, 1−19.

(2) Martin, T. M.; Harten, P.; Venkatapathy, R.; Das, S.; Young, D. M. A hierarchical clustering methodology for the estimation of toxicity. *Toxicology Mechanisms and Methods* **2008**, *18*, 251−266.

(3) Thomas, R. S; Bahadori, T.; Buckley, T. J.; Cowden, J.; Deisenroth, C.; Dionisio, K. L.; Frithsen, J. B.; Grulke, C. M.; Gwinn, M. R.; Harrill, J. A.; et al. The next generation blueprint of computational toxicology at the US Environmental Protection Agency. *Toxicol. Sci.* **2019**, *169*, 317−332.

(4) Letinski, D. J.; Redman, A. D.; Birch, H.; Mayer, P. Interlaboratory comparison of water solubility methods applied to difficult-to-test substances. *BMC Chemistry* **2021**, *15*, 52.

(5) Dougherty, R. C. Temperature and pressure dependence of hydrogen bond strength: A perturbation molecular orbital approach. *J. Chem. Phys.* **1998**, *109*, 7372−7378.

(6) Delaney, J. S. Predicting aqueous solubility from structure. *Drug discovery today* **2005**, *10*, 289−295.

(7) Huuskonen, J. Estimation of Aqueous Solubility in Drug Design. *Combinatorial Chemistry & High Throughput Screening* **2001**, *4*, 311−316.

(8) Meftahi, N.; Walker, M. L.; Smith, B. J. Predicting aqueous solubility by QSPR modeling. *Journal of Molecular Graphics and Modelling* **2021**, *106*, 107901.

(9) Raevsky, O. A.; Grigorev, V. Y.; Polianczyk, D. E.; Raevskaja, O. E.; Dearden, J. C. Six global and local QSPR models of aqueous solubility at pH= 7.4 based on structural similarity and physicochemical descriptors. *SAR and QSAR in Environmental Research* **2017**, *28*, 661−676.

(10) Lovric, M.; Pavlovic, K.; Zuvela, P.; Spataru, A.; Lucic, B.; Kern, R.; Wong, M. W. Machine learning in prediction of intrinsic aqueous solubility of drug-like compounds: Generalization, complexity, or predictive ability? *Journal of Chemometrics* **2021**, *35*, e3349.

(11) Sorkun, M. C.; Khetan, A.; Er, S. AqSolDB, a curated reference set of aqueous solubility and 2D descriptors for a diverse set of compounds. *Scientific data* **2019**, *6*, 1−8.

(12) OECD, Organisation for Economic Co-operation and Development. *OECD principles for the Varidation, for Regulatory Purpose, of (Q)SAR Models*, 2004.

(13) OECD, Organisation for Economic Co-operation and Development. *Guidance Document on the Validation of (Quantitative) Structure−Activity Relationship [(Q)SAR] Models, Series on Testing and Assessment*, 2007.

(14) Breiman, L. Random forests. *Machine learning* **2001**, *45*, 5−32.

(15) Dearden, J.C.; Cronin, M.T.D.; Kaiser, K.L.E. How not to develop a quantitative structure−activity or structure−property relationship (QSAR/QSPR). *SAR and QSAR in Environmental Research* **2009**, *20*, 241−266.

(16) Walker, J. D.; Jaworska, J.; Comber, M. H.I.; Schultz, T. W.; Dearden, J. C. Guidelines for developing and using quantitative structure-activity relationships. *Environmental Toxicology and Chemistry: An International Journal* **2003**, *22*, 1653−1665.

(17) ECHA, European Chemicals Agency. Registration Dossier. *Registration Dossier*, 2022. https://echa.europa.eu/registration-dossier/ (accessed 12/21/2021).

(18) ADDoPT, Advanced Digital Design of Pharmaceutical Therapeutics. *Advanced Digital Design of Pharmaceutical Therapeutics*, 2022. https://www.addopt.org/ (accessed 12/21/2021).

(19) Bradley, J.-C.; Abraham, M. H; Acree, W. E; Lang, A. S.; Beck, S. N; Bulger, D. A; Clark, E. A; Condron, L. N; Costa, S. T; Curtin, E. M; et al. Determination of Abraham model solute descriptors for the monomeric and dimeric forms of trans-cinnamic acid using measured solubilities from the Open Notebook Science Challenge. *Chemistry Central Journal* **2015**, *9*, 1−6.

(20) Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; et al. Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *Journal of computer-aided molecular design* **2011**, *25*, 533−554.

(21) *LookChem*, 2022. https://www.lookchem.com/ (accessed 12/21/2021).

(22) Ruusmann, V; Sild, S; Maran, U QSAR DataBank repository: open and linked qualitative and quantitative structure–activity relationship models. *Journal of Cheminformatics* **2015**, *7*, 1−11.

(23) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A; Thiessen, P. A; Yu, B.; et al. PubChem in 2021: new data content and improved web interfaces. *Nucleic acids research* **2021**, *49*, 1388−1395.

(24) Williams, A. J.; Grulke, C. M.; Edwards, J.; McEachran, A. D.; Mansouri, K.; Baker, N. C.; Patlewicz, G.; Shah, I.; Wambaugh, J. F.; Judson, R. S.; et al. The CompTox Chemistry Dashboard: a community data resource for environmental chemistry. *Journal of cheminformatics* **2017**, *9*, 1−27.

(25) Grulke, C. M.; Williams, A. J.; Thillanadarajah, I.; Richard, A. M. EPA's DSSTox database: history of development of a curated chemistry resource supporting computational toxicology research. *Computational Toxicology* **2019**, *12*, 100096.

(26) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *J. Chem. Inf. Model.* **2010**, *50*, 1189−1204.

(27) Unsupervised feature learning and deep learning: A review and new perspectives.Bengio, Y.; Courville, A. C.; Vincent, P. *arXiv (Machine Learning)*, 1206.5538, ver. 1, 2012. https://arxiv.org/abs/1206.5538 (accessed 11/01/2022).

(28) Contrera, J. F; Matthews, E. J; Benz, D. R. Predicting the carcinogenic potential of pharmaceuticals in rodents using molecular structural similarity and E-state indices. *Regul. Toxicol. Pharmacol.* **2003**, *38*, 243−259.

(29) Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of computational chemistry* **2011**, *32*, 1466−1474.

(30) Lowe, C. N.; Isaacs, K. K.; McEachran, A.; Grulke, C. M.; Sobus, J. R.; Ulrich, E. M.; Richard, A.; Chao, A.; Wambaugh, J.; Williams, A. J. Predicting compound amenability with liquid chromatography-mass spectrometry to improve non-targeted analysis. *Anal. Bioanal. Chem.* **2021**, *413*, 7495−7508.

(31) Kuhn, M. Building predictive models in R using the caret package. *Journal of statistical software* **2008**, *28*, 1−26.

(32) *R: A language and environment for statistical computing*, 2013.

(33) Shannon, C. E. A Mathematical Theory of Communication. *Bell System Technnical Journal* **1948**, *27*, 623−656.
(34) Vinh, N. X.; Epps, J. Information theoretic measures for clusterings comparison: variants, properties, normalization and correction for chance. *J. Mach Learn Res.* **2010**, *11*, 2837−2854.
(35) Meylan, W. M.; Howard, P. H. *Estimating log P with atom/fragments and water solubility with log* P. *Perspectives in drug discovery and design* **2000**, *19*, 67−84.
(36) Gramatica, P. Principles of QSAR modeling: comments and suggestions from personal experience. *International Journal of Quantitative Structure-Property Relationships (IJQSPR)* **2020**, *5*, 61−97.
(37) Randic, M. Characterization of molecular branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609−6615.
(38) Yang, C.; Tarkhov, A.; Marusczyk, J.; Bienfait, B.; Gasteiger, J.; Kleinoeder, T.; Magdziarz, T.; Sacher, O.; Schwab, C. H.; Schwoebel, J.; et al. New publicly available chemical query language, CSRML, to support chemotype representations for application to data mining and modeling. *J. Chem. Inf. Model.* **2015**, *55*, 510−528.
(39) Martin, T. M.; Harten, P.; Young, D. M.; Muratov, E. N.; Golbraikh, A.; Zhu, H.; Tropsha, A. Does rational selection of training and test sets improve the outcome of QSAR modeling? *J. Chem. Inf. Model.* **2012**, *52*, 2570−2578.
(40) Fisher, R. A. The use of multiple measurements in taxonomic problems. *Annual Eugenics* **1936**, *7*, 179−188.
(41) Tricarico, G. A.; Hofmans, J.; Lenselink, E. B.; Ramos, M. L.; Dreanic, M.-P.; Stouten, P. F. W. Construction of balanced, chemically dissimilar training, validation and test sets for machine learning on molecular datasets. *ChemRxiv*, ver. 1, 2022. https://chemrxiv.org/engage/chemrxiv/article-details/6253d85d88636ca19c0de92d (accessed 11/01/2022).




https://doi.org/10.1021/acs.chemrestox.2c00379
*Chem. Res. Toxicol.* 2023, 36, 465−478