A comparison of molecular representations for lipophilicity quantitative structure–property relationships with results from the SAMPL6 logP Prediction Challenge

Article *in* Journal of Computer-Aided Molecular Design · May 2020

DOI: 10.1007/s10822-020-00279-0

CITATIONS
6

READS
326

3 authors:

Raymond Lui
The University of Sydney
5 PUBLICATIONS **22** CITATIONS
SEE PROFILE

Davy Guan
The University of Sydney
8 PUBLICATIONS **40** CITATIONS
SEE PROFILE

Slade Matthews
The University of Sydney
56 PUBLICATIONS **1,350** CITATIONS
SEE PROFILE

Journal of Computer-Aided Molecular Design
https://doi.org/10.1007/s10822-020-00279-0

A comparison of molecular representations for lipophilicity quantitative structure–property relationships with results from the SAMPL6 logP Prediction Challenge

Raymond Lui[1] · Davy Guan[1] · Slade Matthews[1]

Received: 16 October 2019 / Accepted: 8 January 2020

Abstract

Effective representation of a molecule is required to develop useful quantitative structure–property relationships (QSPR) for accurate prediction of chemical properties. The octanol–water partition coefficient logP, a measure of lipophilicity, is an important property for pharmacological and toxicological endpoints used in the pharmaceutical and regulatory spheres. We compare physicochemical descriptors, structural keys, and circular fingerprints in their ability to effectively represent a chemical space and characterise molecular features to correlate with lipophilicity. Exploratory landscape continuity analyses revealed that whole-molecule physicochemical descriptors could map together compounds that were similar in both molecular features and logP, indicating higher potential for use in logP QSPRs compared to the substructural approach of structural keys and circular fingerprints. Indeed, logP QSPR models parameterised by physicochemical descriptors consistently performed with the lowest error. Our best performing model was a stochastic gradient descent-optimised multilinear regression with 1438 descriptors, returning an internal benchmark RMSE of 1.03 log units. This corroborates the well-established notion that lipophilicity is an additive, whole-molecule property. We externally tested the model by participating in the 2019 SAMPL6 logP Prediction Challenge and blindly predicting for 11 protein kinase inhibitor fragment-like molecules. Our model returned an RMSE of 0.49 log units, placing eighth overall and third in the empirical methods category (submission ID 'hdpuj'). Permutation feature importance analyses revealed that physicochemical descriptors could characterise predictive molecular features highly relevant to the kinase inhibitor fragment-like molecules.

Keywords QSPR · logP · Physicochemical properties · Machine learning · SAMPL6

Introduction

The accurate prediction of the octanol–water partition coefficient, logP, is necessary to elucidate the pharmacological and toxicological activity of chemicals, such as cell membrane traversal and hydrophobic target binding. Mathematical correlations between chemical structure and logP are well-established [1–4] and advances in computational modelling approaches have provided new algorithmic bases for lipophilicity quantitative structure–property relationships (QSPRs) [5, 6]. As these approaches become increasingly complex, it is important to understand the foundations of encoding and presenting a molecule to computational models to ensure effective QSPRs are developed for generalisable chemical property prediction.

A machine learning QSPR approach to modelling logP in a diverse chemical space involves three primary stages [6]. First, each chemical is characterised in silico by a molecular representation—a set of identifying structural and molecular features. A supervised machine learning algorithm is then employed to mathematically correlate each feature in the representation to the known logP value of the respective chemical. Finally, the result is a QSPR model of structure–lipophilicity relationships which can be correlated with the molecular representations of other chemicals with unknown logP to generate an output prediction.

The in silico presentation of a molecule to a computational model is achieved through a range of theoretical dimensions (Fig. 1) [5, 7]. At the zeroth dimension,

✉ Slade Matthews
slade.matthews@sydney.edu.au

[1] Pharmacoinformatics Laboratory, Discipline of Pharmacology, School of Medical Sciences, Faculty of Medicine and Health, The University of Sydney, Sydney, NSW 2006, Australia

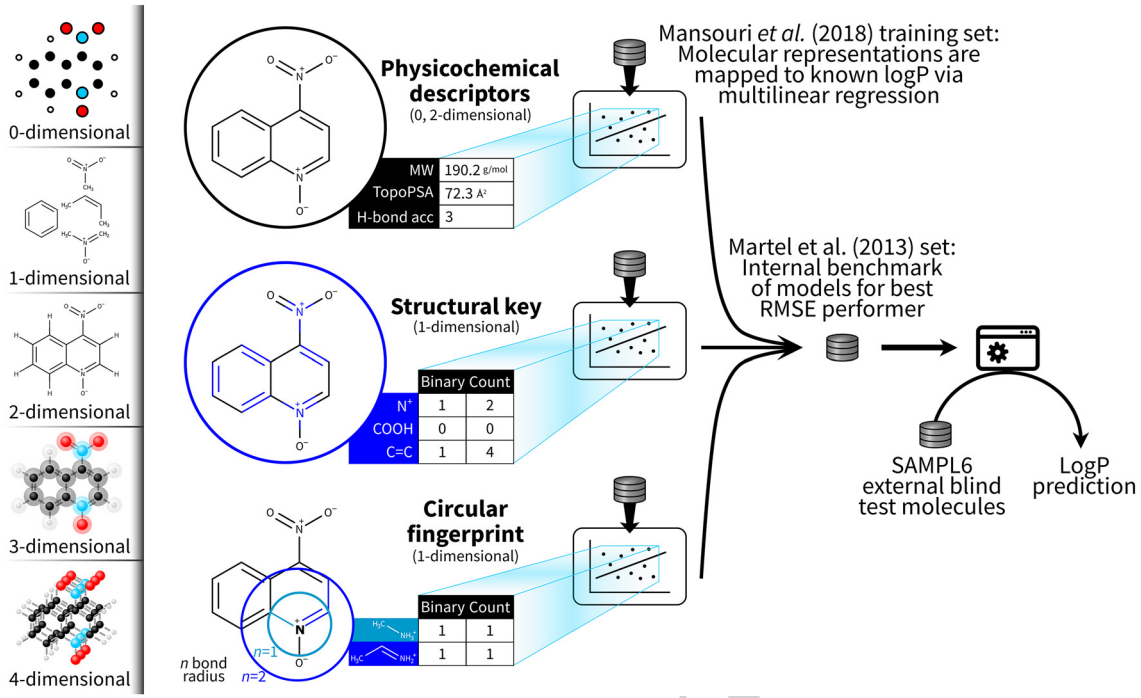


Fig. 1 Overview of this quantitative structure–property relationship modelling study. We begin at the theoretical dimensions of a molecule which form the bases of the molecular representations presented to an algorithm. The resulting model which best fits the represented chemical space was selected as the canonical logP model to participate in the 2019 SAMPL6 logP Prediction Challenge

molecules are depicted by their constituent atoms regardless of bond connectivity. A one-dimensional representation extracts molecular substructures, ranging from smaller functional groups to larger substituent fragments, that together comprise the molecule. In the two-dimensional plane, molecules are depicted through an integrated chemical graph with nodes and edges delineating atomic and bonding connectivity on which topological features can be calculated. Molecules are characterised in three-dimensional space through the further consideration of steric, volumetric, and electronic features. These features can also be interpreted through an additional fourth dimension such as a temporal scale or different molecular conformations.

Three molecular representations were selected for this study based on their abilities to encode distinct molecular scales: physicochemical descriptors, structural keys, and circular fingerprints (Fig. 1). Physicochemical descriptors characterise molecules on a global molecular scale—that is, they consider the whole molecule in their representation. This is achieved by calculating constitutional chemical properties in the zeroth dimension, or atom-specific indices as two-dimensional topological functions of all other atoms in the molecule [8–10]. Conversely, circular fingerprints depict molecules on a local molecular scale—that is, they encode unique one-dimensional substructures from specific regions of a molecule bounded by local circular 'neighbourhoods' [11]. Structural keys fall in the centre of this molecular scale, encoding the presence of specific, predefined chemical moieties but from a whole-molecule macroperspective.

Each of these molecular representation approaches has been extensively used in existing logP models [3, 4, 12–16]. The question we aim to answer through this particular study is *how should molecules be encoded and presented to a logP QSPR model to ensure optimal predictivity?* Specifically, physicochemical descriptors, structural keys, and circular fingerprints are compared, through a standardised benchmark, in their ability to represent a chemical space and characterise molecular features to correlate with lipophilicity.

In early 2019, the Statistical Assessment of the Modeling of Proteins and Ligands (SAMPL) and Drug Design Data Resource (D3R) programs hosted a blind logP Prediction Challenge with 11 protein kinase inhibitor fragment-like molecules. We took this opportunity to investigate how well the best molecular representation could model the octanol–water partition coefficient in a real-world drug-like scenario.

Methods

Our study consisted of two stages: (1) identification of the best performing molecular representation for predicting logP by assessing chemical space landscapes and internally benchmarking logP QSPR models, then (2) external blind

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

testing of the best performing model by participating in the SAMPL6 logP Prediction Challenge (Fig. 1).

Molecular representation

Physicochemical descriptors, structural keys, and circular fingerprints were generated as vector data structures for each molecule:

1. A set of 1444 physicochemical descriptors was calculated using PaDEL-descriptor v2.21 [17]. Examples of descriptors included: whole-molecule counts of functional groups such as halogens, chains, rings, etc.; counts and summations of atom-type electrotopological state indices which quantify the electronegativity of an atom relative to the other atoms of the molecule [8–10]; and graph-based spatial autocorrelation indices as functions of electronegativity, ionisation energy, molecular mass, etc. [18]. Six descriptors which involved the explicit calculation of logP were excluded to avoid information leakage when predicting logP.
2. A 1354-bit structural key was computed using PaDEL-descriptor v2.21 [17] by combining MACCS (Molecular ACCess System), SMARTS (SMiles ARbitrary Target Specification) substructure, and PubChem fingerprints. Keys contain bits representing predefined elements (e.g. "≥ 1 Na" or "≥ 4C") or substructures (e.g. "O–C–C–N") within a molecule. Each bit is marked with a "1" or "0" denoting the presence or absence, respectively, of a substructural feature. Structural keys were also calculated in count form where each bit is an integer value denoting the quantity of that substructural feature within the molecule.
3. A 1024-bit circular fingerprint with a maximum radius of 2 bond lengths was computed using the RDKit-based Fingerprint Calculator package (https://github.com/isidroc/FingerprintCalculator; accessed March 2019) for Python v2.7.15; this protocol is an approximate reimplementation of an extended connectivity fingerprint with maximum diameter of 4 bond lengths (ECFP4) as devised by Rogers and Hahn [11]. In brief, circular fingerprints are computed by assigning an atom with a unique numerical identifier consisting of atomic features, then assigning identifiers for neighbouring atoms within the defined maximum bond radius, essentially 'tracing' out substructural pathways along the bonds that radiate outward from the initial atom. This is repeated for every atom in a molecule then all identifiers are hashed (as per the RDKit hashing function; https://www.rdkit.org) into a 1024-bit vector. Duplicate identifiers are retained for count fingerprints and removed for binary fingerprints.

Each representation was reproduced with varying molecular feature densities to investigate the balance between model underfitting (too little molecular information to make useful correlations with logP) and overfitting (too much molecular information about the chemical space of the training set, resulting in the model not generalising well with other chemical species). Lower density subsets of each molecular representation were generated by selecting 50% and 25% of the original number of descriptors or bits. Molecular feature selection for the physicochemical descriptors and structural keys was performed using a random forest algorithm (Scikit-Learn v0.20.3 [19] for Python v3.6.8) to rank the logP predictive power of each feature based on its relative ability to generate a decision split; the top 50% and 25% ranked molecular features were selected for each representation. Circular fingerprints did not require feature selection since lower density fingerprints could be computed by hashing the unique identifiers into lower bit vectors (i.e. 512 and 256 bits).

Composite representations were generated by combining different molecular representations at each feature density level together into a single vector. This was performed to investigate whether there are benefits in representing broader molecular scales for the development of structure–lipophilicity relationships.

It is noted that molecular representations only up to two dimensions were considered for this study due to the relatively fast speed at which they can be calculated, allowing for a more readily deployable model. Prior to representation calculation, all compounds were formatted in SMILES nomenclature then canonicalised by removing any metal atoms, salts, and solvents, neutralising remaining charged fragments, and recalculating the two-dimensional coordinates of new or changed atoms. This protocol was implemented in ChemAxon Standardizer v18.13.0 (ChemAxon Ltd., 2018; https://www.chemaxon.com).

Exploratory analysis of represented chemical spaces with regard to logP

Qualitative and quantitative approaches were undertaken to gauge how well each molecular representation generates a chemical space amenable to further modelling with logP. A principal component analysis (PCA) combined linearly correlated molecular features within each representation into two new principal component pseudo-features, allowing the molecular features to be more easily visualised in a two-dimensional x–y chemical space. A logP z-axis was added to form a three-dimensional structure–lipophilicity landscape, allowing us to identify where in the represented chemical space to expect a molecule of low or high logP and how these logP gradients change over the chemical landscape; in

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

much the same way valleys and hills undulate over a geomorphological landscape. In this study, we defined a suitable structure–lipophilicity landscape as one with minimal undulations and a smooth transitioning logP gradient, meaning the molecular representation possesses the appropriate features to cluster molecules by their degree of lipophilicity.

To quantify the 'smoothness' of the structure–lipophilicity landscape, we drew upon the work of Peltason and Bajorath [20] and Guha and Van Drie [21] involving the calculation of indices that assess changes in bioactivity with molecular features. However, where they prioritise the identification of activity cliffs (distinct changes in bioactivity with minor structural changes) as is beneficial for applications such as QSAR in drug discovery, we prioritise the identification of smoother, more continuous landscapes that allow for coherent logP estimation through effective molecular representation. To this end, we devised an index relating the degree of landscape continuity with the Dice similarity [22, 23] and logP difference (scaled between 0 and 1) between a pair of compounds, AB, for all possible pairwise combinations of compounds, n, in a given dataset:

$$\textit{Structure–lipophilicity landscape continuity index} = \frac{1}{n}\sum_{AB=1}^{n}\left\{\textit{Dice similarity}_{AB} \times \left(1 - \nu \log P'_{AB}\right)\right\}$$

where,

$$\textit{Dice similarity between molecules A and B for continuous features} = \frac{2\sum_{i=1}^{n} x_{iA}x_{iB}}{\sum_{i=1}^{n}\left(x_{iA}\right)^2 + \sum_{i=1}^{n}\left(x_{iB}\right)^2}$$

where $x_{iA/B} = i\text{th descriptor in molecule } \frac{A}{B}$

$$\textit{Dice similarity between molecules A and B for binary features} = \frac{2c}{a+b}$$

where a/b = number of on bits in molecule A/B. c = number of on bits in molecules A and B.

$$\Delta\log P' = \frac{\Delta\log P - (\Delta\log P)_{min}}{(\Delta\log P)_{max} - (\Delta\log P)_{min}}$$

Continuity index values range from zero to one since Dice similarity and the scaled logP difference (ΔlogP′) are also bounded between zero and one. Here, a continuous and effective structure–lipophilicity landscape (i.e. a higher continuity index) is defined as one with, on average, more clusters of molecules with high chemical similarity and comparable logP values.

Multilinear regression algorithms to generate quantitative structure–lipophilicity relationships

A multiple linear regression approach generated structure–lipophilicity relationships for training molecules by correlating their computationally-derived molecular representations to their respective, experimentally-derived logP values. The linear and additive nature of lipophilicity has been previously documented [1, 2] and used in the development of established logP predictive models that use substructural-based (e.g. XLOGP [3, 13, 24]) and whole molecule-based representations (e.g. MLOGP [4]). Furthermore, the relative ease of calculation of linear regressions complements the aforementioned fast speeds at which zero- to two-dimensional representations are computed, allowing for efficient development and rapid deployment of logP QSPR models.

Five replicate multilinear regressions were developed for each molecular representation, each with coefficients optimised in different ways to fit the projected chemical space: (1) Ordinary least squares regression fits a model by minimising the sum of the squared residuals between the predicted and experimental logP values; (2) Ridge regression fits a model by minimising the sums of squared residuals and the squared coefficients, thereby penalising the weights and driving down larger coefficients to reduce overfitting; (3) Lasso regression reduces overfitting by minimising the sums of squared residuals and the absolute values of the coefficients, similarly penalising the coefficients but instead driving smaller coefficients to zero; (4) ElasticNet regression combines ridge and lasso regularisation at varying ratios; (5) Stochastic gradient descent (SGD) regression fits a model by randomly initialising and iteratively updating the coefficients, facilitating efficient modelling of dimensionally-large chemical spaces.

The training set used to parameterise the QSPR models consisted of 14,050 chemicals structurally curated by

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

Mansouri et al. [25], with experimental and extrapolated logP values compiled within the PHYSPROP database (SRC Inc., North Syracuse, NY, USA) as part of the US EPA EPI Suite. Continuous/non-binary molecular representations were scaled to address differences in units and ranges. Hyperparameters of each multilinear regression method were optimised using threefold cross validation on the training set. All modelling was performed using the Scikit-Learn v0.20.3 machine learning library [19] for Python v3.6.8.

Internal benchmarking of logP models developed using different molecular representations

All developed models were used to predict on a held-out internal validation dataset to (a) benchmark and compare the quality of the logP QSPR models developed on each molecular representation, and (b) select a final flagship logP QSPR model for participation in the SAMPL6 logP Prediction Challenge.

The internal benchmarking set consisted of 707 chemicals with reference logP values experimentally determined by Martel et al. [26] using reversed-phase ultra-high-performance liquid chromatography with a stationary phase of C_{18} chains and mobile phase of methanol/aqueous buffer. The empirical logP values were rigorously validated by Martel et al. [26] based on consensus with four different logP prediction software (ALOGP2, KowWIN, ACD/logP, and ABlogP), forming the basis for a reliable validation dataset that has been used for standardised benchmarking across different logP estimation methods [27–30].

For models with a random state hyperparameter (e.g. ridge/SGD regression requiring random data shuffling, or lasso/ElasticNet regression requiring random coefficient updating), five different random states were initialised to produce five replicate predictions which were averaged to form the final logP prediction. Root mean square error (RMSE) was selected to quantify model performance since larger errors between predicted and experimental values are appropriately penalised given the logarithmic scaling of logP.

External blind testing through the 2019 SAMPL6 logP Prediction Challenge

The best performing logP model, as defined by the lowest overall RMSE in the internal benchmark, was used to estimate the logP of 11 protein kinase inhibitor fragment-like molecules prepared for the SAMPL6 logP Prediction Challenge. Experimental logP values were experimentally determined by Isik et al. [31] using potentiometric titrations in a biphasic 1–octanol–water system. Our submission ID was 'hdpuj'.

Overall model performance relative to other challenge entrants was ranked according to RMSE, mean absolute error (MAE), mean error (ME), square of Pearson correlation coefficient (R^2), and slope of the predicted vs experimental logP line of best fit (m). 95% confidence intervals for all statistics were calculated by bootstrapping with 10,000 replications. Complete challenge results are available at https://github.com/samplchallenges/SAMPL6/tree/master/physical_properties/logP/analysis.

Once experimental logP values were released following the submission deadline, a permutation feature importance analysis was computed to identify specific physicochemical descriptors deemed necessary for our model to predict the logP of the 11 molecules. The analysis involves quantifying the increase or decrease in mean squared error between the predicted and experimental logP when a descriptor column is randomly permuted so that any potentially useful molecular information is replaced with noise. A higher permutation feature importance weight (i.e. a greater increase in error) indicates that the permuted descriptor contains essential information necessary for an accurate logP prediction. Permutation importance was computed using the eli5 package for Python v3.6.8 (https://github.com/TeamHG-Memex/eli5; accessed October 2019), with the final change in error for each descriptor averaged from 100 permutations per descriptor column.

All datasets, molecular representations, modelling code, and raw results from this study are available at https://github.com/luiraym/RAYLOGP.

Results and discussion

Chemical data preparation

The finalised training sets consisted of 14,045 molecules for physicochemical descriptors (descriptors for 5 molecules could not be calculated), 14,050 molecules for structural keys, and 13,710 molecules for circular fingerprints (340 molecules could not be processed). Empirical logP values in the training set ranged from − 5.08 to 11.29, with mean 2.07 (Fig. 2). The internal benchmarking sets consisted of 707 molecules for all three molecular representations, with empirical logP values ranging from 0.30

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

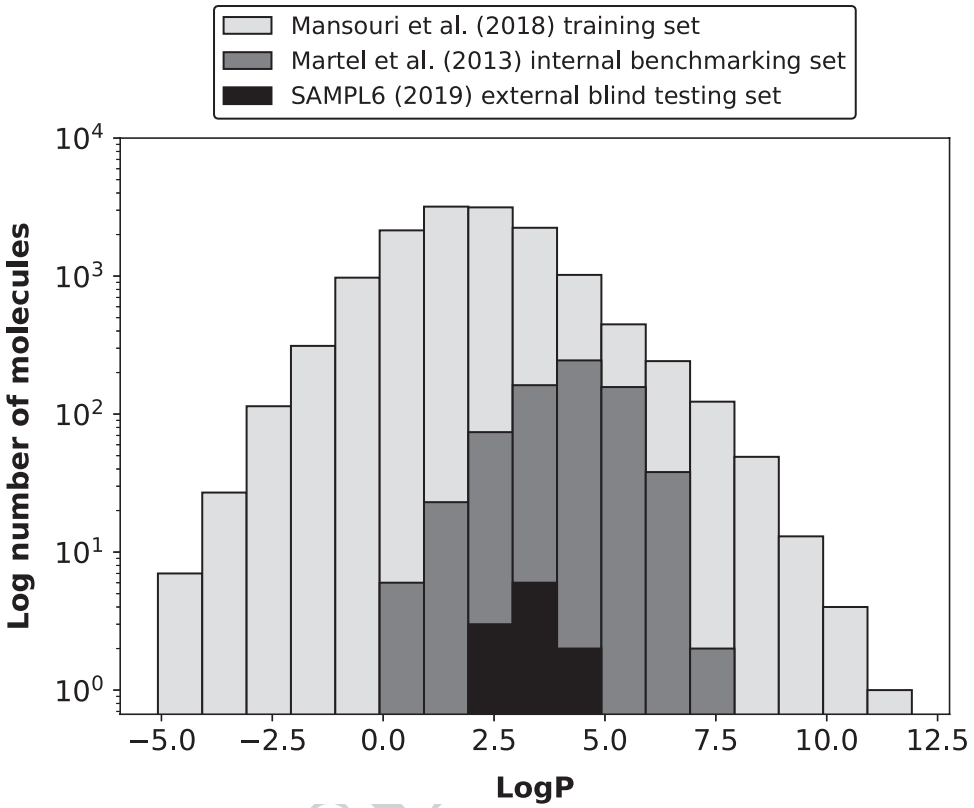


Fig. 2 Histogram presenting logP distributions within each chemical dataset used in this study

to 6.96; mean 4.19 (Fig. 2). The SAMPL6 external blind testing set consisted of 11 molecules with empirical logP values revealed after the challenge, ranging from 1.94 to 4.09; mean 3.08 (Fig. 2).

The final feature densities (full/50%/25%, respectively) for each molecular representation were as follows: 1438/719/360 physicochemical descriptors, 1354/677/339 binary or count bits for the structural keys; and 1024/512/256 binary or counts bits for the circular fingerprints.

Chemical spaces mapped by physicochemical descriptors identify compounds that are both structurally and lipophilically similar

The 707-chemical internal benchmarking set [26] was used as the chemical space in which we performed pre-modelling exploratory analysis of the molecular representations. Physicochemical descriptors produced the most ideal landscape, with a clear gradient from low to high logP revealing the ability of a molecule's physicochemical properties to adequately estimate its logP (Fig. 3a). Conversely, the landscape as mapped by circular fingerprints is much more homogenous in distribution with no defined separation between low and high logP values (Fig. 3c). The structural key produced a landscape with a clear section of chemical space of low logP similar to that of the physicochemical descriptors, but a more diffuse space of high logP reminiscent of the circular fingerprints (Fig. 3b). We note that the jagged peaks of our visualised landscapes are not indicative of activity cliffs but rather result from the partial imputation between the inherently scattered datapoints [32]; we are more interested in the overall heterogeneity or homogeneity of the distribution from low to high logP as represented by colour transition from light to dark.

The results of the qualitative landscape analysis were corroborated by the structure–lipophilicity continuity indices calculated for each molecular representation. Physicochemical descriptors returned a relatively high index of 0.76, indicating that, on average, pairs of molecules with similar lipophilicity (i.e. low $\Delta\log P'$) were also chemically similar when comparing their physicochemical properties (i.e. high Dice similarity). Therefore, physicochemical descriptors were identified as an early candidate for effective structure–lipophilicity relationship generation since the molecular features appeared to correlate better with logP. Contrastingly, binary and count circular fingerprints returned a relatively low

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

index of 0.18 and 0.02 respectively, and thus we expected fingerprints to generate less effective structure–lipophilicity relationships since the substructural features failed to adequately identify aspects of chemical similarity for comparably lipophilic molecules. As with the landscapes, the continuity index for structural keys was ranked in between the descriptors and fingerprints (0.48 for binary, 0.48 for count).

A multilinear regression parameterised by physicochemical descriptors produced the lowest internal benchmarking RMSE

Our logP QSPR models consistently performed best when parameterised by physicochemical descriptors, producing an overall average RMSE across three feature densities of 1.23 log units on the internal benchmarking set, compared to 1.35 for structural keys and 1.95 for circular fingerprints (Fig. 4). Linearly correlating logP with whole-molecule properties generated the most useful structure–lipophilicity relationships, reflecting the molecule-encompassing nature of logP itself. Nevertheless, several successful and well-established logP estimation methods have been developed with a substructural approach [33], particularly atomic contribution methodologies as applied in XLOGP [3, 13, 24]. Physicochemical descriptors and structural keys include *whole-molecule* counts of atoms and small fragments, which is likely to be responsible for their relatively similar high performance. On the other hand, the circular fingerprints computed in this study consisted of substructures calculated within *local neighbourhoods* two bond lengths in radius. Identifying unique fragments only in a specific region of a molecule might fail to capture the whole-molecule additivity required for logP estimation, hence the significantly higher RMSE. Therefore, circular fingerprints may predict logP better if the neighbourhood is configured to a smaller bond length radius, i.e. zero or one, where it begins to emulate atomic identification in the context of the whole molecule.

The logP model parameterised by a composite of all three molecular representations predicted the internal benchmarking set with an average RMSE of 1.29 log units (Fig. 4). A second composite representation was developed using physicochemical descriptors and structural keys only, returning an average RMSE of 1.26 log units (Fig. 4). This suggests that the addition of physicochemical properties reinforces the logP predictivity of substructural representations, potentially paving way for a flexible representation in the future combining different scales of molecular features. Recent advances in multitask deep learning can facilitate the simultaneous modelling of several molecular representations using multi-input neural networks to incorporate different aspects of a chemical for more generalisable QSAR/QSPR predictions [34].

logP models developed on count structural keys and circular fingerprints across five multilinear regressions produced lower average RMSEs of 1.30 and 1.88, respectively, compared to their binary counterparts which produced average RMSEs of 1.40 and 2.02, respectively (Fig. 4). The continuous nature of count variables complements the additivity found in logP and thus is a better basis for structure–lipophilicity relationships; i.e. when a molecule is represented as having multiple lipophilic moieties then the logP estimation will scale higher with the count.

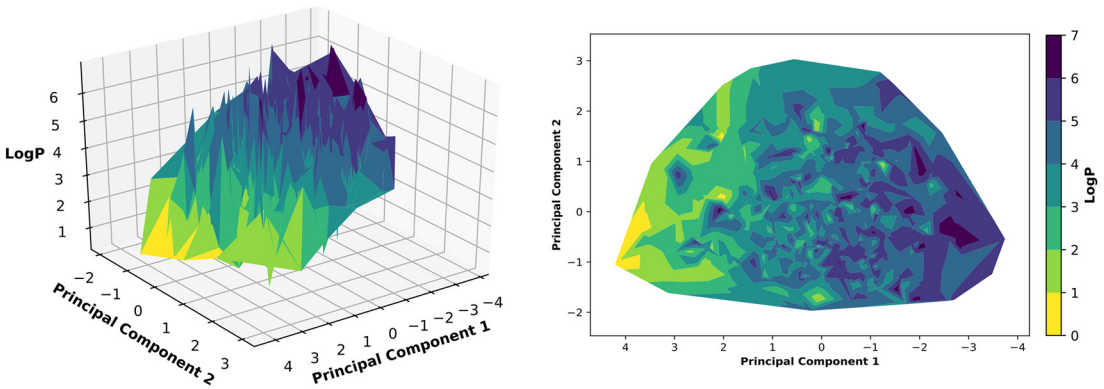
The average RMSE of logP models trained on physicochemical descriptors increased as the number of descriptors decreased (Fig. 4), indicating reliance on most of the physicochemical properties for lipophilicity QSPRs. Circular fingerprint-based models similarly increased in error with decreasing bits which may be attributed to higher frequencies of bit collisions as the substructural data is hashed into lower fingerprint lengths, resulting in a loss of molecular information for QSPR generation. Conversely, structural keys produced lower average RMSE with decreasing molecular feature density (Fig. 4), which may be due to the removal of correlated substructural bits common between the three keys (MACCS, SMARTS, and PubChem) that were combined.

We conclude that there is no 'one-size-fits-all' molecular feature density for predicting logP. It is important to include enough features to provide molecular information for QSPRs to exploit but not such an excessive amount that redundant and noisy QSPRs are generated, leading to low generalisability on chemical species outside the training chemical space. Robust molecular features that can be consistently computed for input chemicals should only be included to avoid reproducibility issues. Moreover, different modelling algorithms can lead to varying optimal feature densities where regularisation methods, such as ridge and lasso in the present study, manipulate feature coefficients/weights as part of the model training process. In the wider machine learning field, algorithms can also be ensembled in deeper ways expanding the possible combinations of molecular features available for QSPR modelling. Ultimately, we recommend feature density to be tailored to individual modelling endeavours based on computational resources, algorithmic approaches, and specific training chemical space.

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

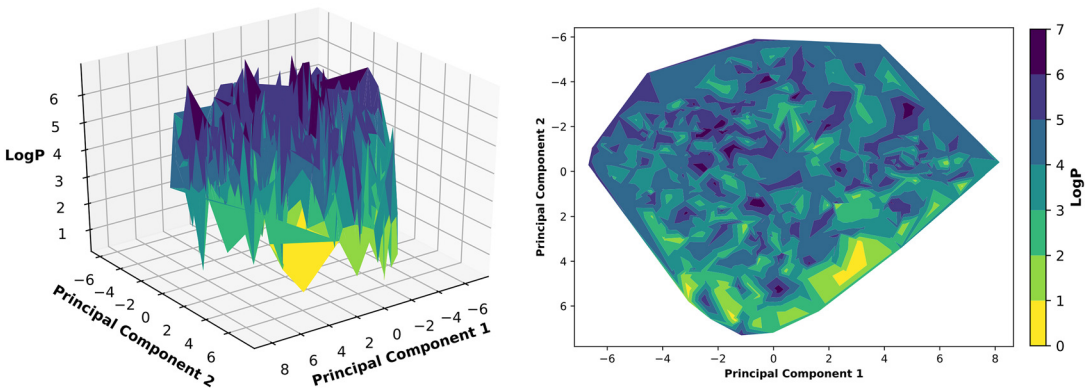
(A) Physicochemical descriptors

Continuity index = 0.760



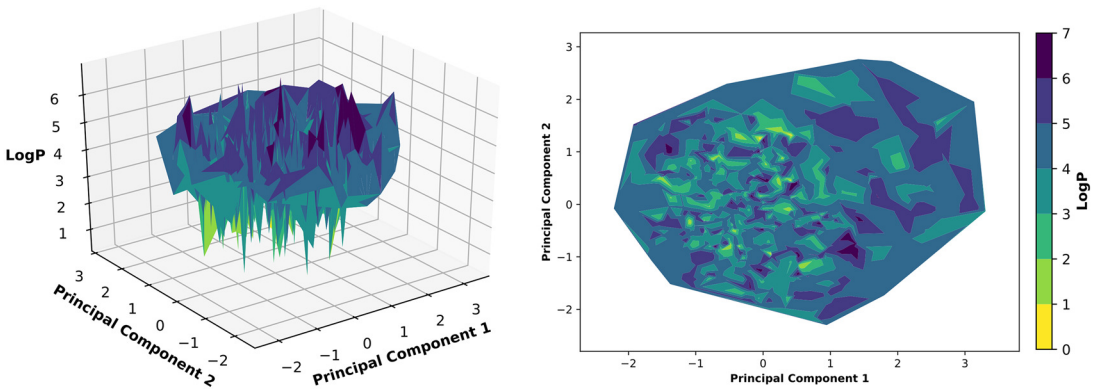
(B) Structural key

Continuity index = 0.477



(C) Circular fingerprint

Continuity index = 0.181



Journal : Large 10822	Article No : 279	Pages : 12	MS Code : 279	Dispatch : 10-1-2020

◂**Fig. 3** Three- and two-dimensional views of structure–lipophilicity landscapes as mapped using **a** physicochemical descriptors, **b** a binary structural key, and **c** a binary circular fingerprint. Each molecular representation was transformed into two dimensions using principal components analysis. The chemical space being mapped is the 707-chemical Martel et al. [26] internal benchmarking set

Our best performing logP model, following comprehensive internal benchmarking of molecular representations, was revealed to be a stochastic gradient descent-optimised multilinear regression with 1438 physicochemical descriptors, returning an internal benchmark RMSE of 1.03 on the Martel et al. [26] dataset. SGD enabled the model to efficiently fit the extremely large dataspace (14,045 training molecules with 1438 descriptors each) by randomly initialising feature coefficients and optimising them until minimum RMSE was achieved. Ultimately, the fast calculation of zero- and two-dimensional physicochemical descriptors combined with the computationally inexpensive SGD multilinear regression results in a rapidly deployable model for practical logP estimation. This model was codenamed 'RAYLOGP' and used as our flagship logP estimator for the SAMPL6 logP Prediction Challenge.

Physicochemical descriptors could identify molecular properties relevant to the prediction of SAMPL6 external test molecules

RAYLOGP predicted logP for the 11 protein kinase inhibitor fragment-like molecules of the SAMPL6 logP Prediction Challenge with an average RMSE of 0.49 (95% CI 0.37–0.61), average MAE of 0.44 (0.32–0.57), and average ME of − 0.29 [(− 0.52) − (− 0.05)]. logP predictions were positively correlated with experimental values with an R^2 of 0.74 (0.41–0.94) and line of best fit slope, m, of 1.02 (0.86–1.37) (Fig. 5).

Physicochemical descriptors are a suitable molecular representation for the SAMPL6 kinase inhibitor fragment-like chemical space. Permutation feature importance analysis found the maximum, minimum, sum, and count of electrotopological state indices for secondary amine groups to be amongst the top 1% influential descriptors for predicting logP. This is particularly relevant for all substituted 4-aminoquinazoline molecules (SM02, SM04, SM07, SM09, SM12, and SM13) and SM16 as they contain a secondary amine (Fig. 5). The effect of random permutation ablating these descriptors substantially increased prediction error (Table 1). Other important physicochemical descriptors that are relevant to the chemical space of the challenge (Table 1) included counts of 6-membered ring structures containing heteroatoms (present in all molecules except SM14 and SM15), odd-order self-returning walk counts which can characterise the connectivity of odd-number membered rings (such as those in SM14 and SM15), and the electrotopological states for carbons that are bound to three aromatic carbons (present in the fused rings of all molecules except SM16).

RAYLOGP was able to correctly rank the last three molecules (SM12, SM04, and SM02) of the 4-aminoquinazoline series (Fig. 5). Geary autocorrelation descriptors as a function of electronegativity and partial charges [18] and electrotopological state descriptors for hydrogen bond donors [10] were identified as determinant features for predicting logP (Table 1). This is relevant when capturing the substituent effect of halogens present in these three molecules. We also note that our final model was selected based on performance on the internal benchmarking set, which had a mean logP of 4.19 skewed toward the lipophilic end (Fig. 2). Therefore, we could more accurately predict the logP of the last three molecules of the 4-aminoquinazoline series as they approached the logP range in which the model was optimised. However, this notably limited predictive performance of lower logP molecules, meaning rank order was not well captured for the first three 4-aminoquinazoline molecules (SM13, SM09, SM07) (Fig. 5). In particular, our model considerably overpredicts the ether-containing SM13 and SM9 molecules.

A simple physicochemical property-based multilinear approach remains a viable option for contemporary lipophilicity prediction

RAYLOGP placed eighth out of 91 submissions overall, and third out of 17 submissions categorised as using an empirical method. Amongst the myriad of state-of-the-art logP estimation methods, we were surprised at how well a simple multilinear regression parameterised by traditional zero- and two-dimensional descriptors performed, thereby corroborating the well-established notion that lipophilicity is an additive, whole-molecule chemical property.

High variance may be a limiting factor for our model due to the large number of physicochemical descriptors included. This could lead to over-characterisation of the molecules in the training set, resulting in poor generalisability of the QSPR model on molecules that share few commonalities with the training chemical space (i.e. model overfitting). A common solution is molecular feature selection for a more concise model, however, earlier experimentation revealed lower numbers of physicochemical descriptors produced higher RMSE (Fig. 4) on our training set. A proposed direction to address this is the coupling of molecular structure optimisation in the gas phase using semi-empirical PM7 methodologies with the

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

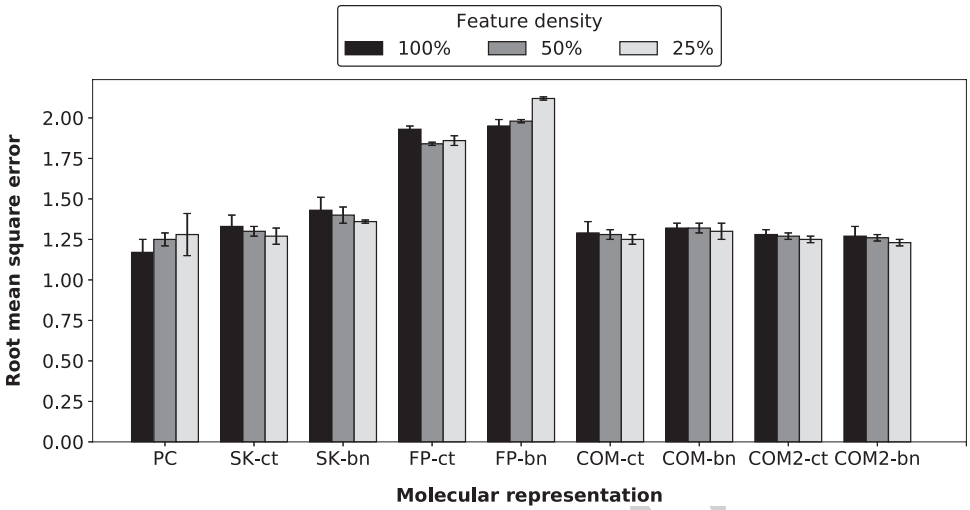


Fig. 4 Average RMSE across five multilinear regressions for physicochemical descriptors (PC), count/binary structural keys (SK-ct/SK-bn), count/binary circular fingerprints (FP-ct/FP-bn), count/binary composites of all three representations (COM-ct/COM-bn), and count/binary composites of physicochemical descriptors and structural keys only (COM2-ct/COM2-bn). Varying feature densities (full, 50%, 25%) of each molecular representation are presented. Errors bars represent standard deviation between the five regression replicates

Mordred two-/three-dimensional descriptor calculation package [35], a more modern Python-integrated alternative to the PaDEL package used in the present study. To this end, potential bias arising from dimensionality reduction is accounted for by the generation of higher quality quantum mechanically-optimised descriptors.

Moreover, a non-linear modelling approach could unveil new structure–lipophilicity correlations that would otherwise not be possible with traditional linear methods used here. This has been demonstrated with great success by model submissions 'gmoq5' and 'sq07q' which employed a gradient-boosted tree-based approach predicting at 0.39 and 0.47 RMSE respectively, placing first and second above our model in the empirical category. More generally, contemporary machine learning algorithms such as random forests, support vector machines, and gradient boosted frameworks have rapidly become the default in QSAR/QSPR model development [5, 6, 14–16, 36–38]. A novel approach that we plan to use for future modelling endeavours involves automated genetic algorithm pipelines to tune and optimise hyperparameters of both linear and non-linear machine learning algorithms to better map molecular representations to a chemical space [39]. This facilitates more efficient and comprehensive QSPR model development, while also freeing up time to curate chemical data and calculate higher quality molecular representations.

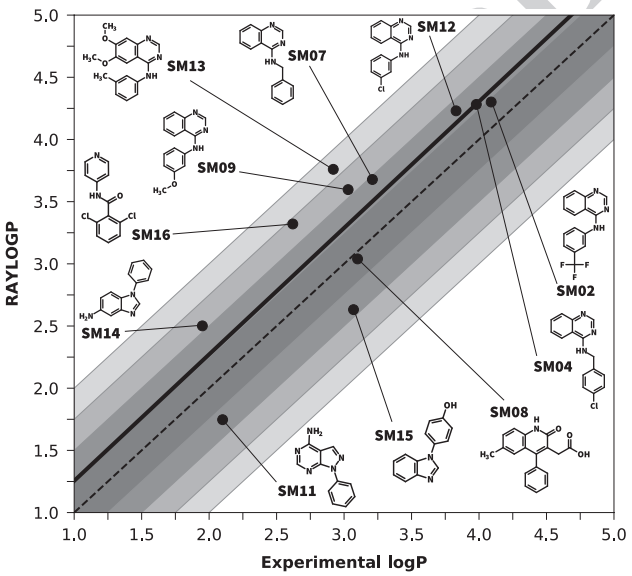


Fig. 5 Correlation plot of RAYLOGP predictions and experimental logP (determined by Isik et al. [31]) of the 11 protein kinase inhibitor fragment-like molecules in the SAMPL6 logP Prediction Challenge. Solid black line represents line of best fit. Gradient region surrounding the dashed perfect correlation line represents error up to one log unit, with each of the four shades denoting 0.25 log unit intervals

Conclusions

Computational logP model performance is highly influenced by the way a molecule is represented. Physicochemical descriptors, structural keys, and circular fingerprints were compared to test their ability to model a chemical space for correlation with lipophilicity. Compounds that were both structurally and lipophilically similar were more cohesively mapped together by physicochemical descriptors. Linear logP models parameterised by physicochemical descriptors predicted with the lowest RMSE. Count structural keys and circular fingerprints better suit the additive nature of lipophilicity compared to their binary counterparts. The optimal density of molecular features is dependent on the modelling approach and bias/variance trade-off. A classic multilinear regression based on zero-/two-dimensional descriptors performed well in the 2019 SAMPL6 logP Prediction Challenge, nonetheless there is room for improvement with

Journal : Large 10822 | Article No : 279 | Pages : 12 | MS Code : 279 | Dispatch : 10-1-2020

Table 1 Top 1% physicochemical descriptors contributing to RAYLOGP predictions for 11 protein kinase inhibitor fragment-like molecules in the SAMPL6 logP Prediction Challenge

Rank	Permutation importance weight (logP MSE)	Physicochemical descriptor code	Physicochemical descriptor documentation
1	0.0367443	maxssNH	Maximum atom-type electrotopological state for secondary amines (–NH–)
2	0.0367436	minssNH	Minimum atom-type electrotopological state for secondary amines (–NH–)
3	0.0235856	SRW9	Self-returning walk count of order 9
4	0.0234578	GATS2e	Geary autocorrelation (lag 2) weighted by Sanderson electronegativities
5	0.0210460	GATS2c	Geary autocorrelation (lag 2) weighted by partial charges
6	0.0169979	minHBd	Minimum atom-type electrotopological state for hydrogen bond donors
7	0.0141076	n6HeteroRing	Number of six-membered rings containing heteroatoms (N, O, P, S, or halogens)
8	0.0117158	nT6HeteroRing	Number of 6-membered rings (including fused rings) containing heteroatoms (N, O, P, S, or halogens)
9	0.0115259	GATS2s	Geary autocorrelation (lag 2) weighted by intrinsic state
10	0.0108340	SssNH	Sum of atom-type electrotopological states for secondary amines (–NH–)
11	0.0095664	GATS1i	Geary autocorrelation (lag 1) weighted by first ionisation potential
12	0.0092998	maxaaaC	Maximum atom-type electrotopological state for a carbon bound to three aromatic carbons (aaaC)
13	0.0091869	SRW7	Self-returning walk count of order 7
14	0.0090311	nHssNH	Count of atom-type electrotopological states for secondary amines (–NH–)
15	0.0090220	maxHBd	Maximum atom-type electrotopological state for hydrogen bond donors

Importance weights of descriptors were calculated as the increase in mean squared error (MSE) when that descriptor was randomly permuted and used to predict logP

greater accessibility to quantum mechanical-based structure optimisation and automated machine learning approaches.

Acknowledgements We thank the National Institutes of Health (Grant No. R01-GM124270) for their support in funding the SAMPL6 Challenges and associated experimental work.

References

1. Fujita T, Iwasa J, Hansch C (1964) A new substituent constant, π, derived from partition coefficients. J Am Chem Soc 86(23):5175–5180
2. Iwasa J, Fujita T, Hansch C (1965) Substituent constants for aliphatic functions obtained from partition coefficients. J Med Chem 8(2):150–153
3. Wang R, Fu Y, Lai L (1997) A new atom-additive method for calculating partition coefficients. J Chem Inf Comput Sci 37(3):615–621
4. Moriguchi I et al (1992) Simple method of calculating octanol/water partition coefficient. Chem Pharm Bull 40(1):127–130
5. Lo Y-C et al (2018) Machine learning in chemoinformatics and drug discovery. Drug Discov Today 23(8):1538–1546
6. Mitchell JBO (2014) Machine learning methods in chemoinformatics. WIREs Comput Mol Sci 4(5):468–481
7. Polanski J, Gasteiger J (2017) Computer representation of chemical compounds. In: Leszczynski J et al (eds) Handbook of computational chemistry. Springer International Publishing, Cham, pp 1997–2039
8. Hall LH, Mohney B, Kier LB (1991) The electrotopological state: an atom index for QSAR. Quant Struct Act Relat 10(1):43–51
9. Kier LB, Hall LH (1990) An electrotopological-state index for atoms in molecules. Pharm Res 7(8):801–807
10. Hall LH, Kier LB (1995) Electrotopological state indices for atom types: a novel combination of electronic, topological, and valence state information. J Chem Inf Comput Sci 35(6):1039–1045
11. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50(5):742–754
12. Wang J-B et al (2015) In silico evaluation of logD7,4 and comparison with other prediction methods. J Chemom 29(7):389–398
13. Wang R, Gao Y, Lai L (2000) Calculating partition coefficient by atom-additive method. Perspect Drug Discov Des 19(1):47–66
14. Chen H-F (2009) In silico log P prediction for a large data set with support vector machines, radial basis neural networks and multiple linear regression. Chem Biol Drug Des 74(2):142–147
15. Lowe EW et al (2011) Comparative analysis of machine learning techniques for the prediction of logP. In: 2011 IEEE symposium on computational intelligence in bioinformatics and computational biology (CIBCB), IEEE, Paris
16. Zang Q et al (2017) In silico prediction of physicochemical properties of environmental chemicals using molecular fingerprints and machine learning. J Chem Inf Model. 57(1):36–49
17. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7):1466–1474
18. Todeschini, R, V Consonni (2009) Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references, vol 41. Wiley, Weinheim
19. Pedregosa F et al (2011) Scikit-learn: machine learning in python. J Mach Learn Res 12:2825–2830
20. Peltason L (2007) J Bajorath, SAR index: quantifying the nature of structure–activity relationships. J Med Chem 50(23):5571–5578

21. Guha R, Van Drie JH (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 48(3):646–658
22. Dice LR (1945) Measures of the amount of ecologic association between species. Ecology 26(3):297–302
23. Bajusz D (2015) A Rácz, K Héberger, Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? J Cheminform 7(1):20
24. Cheng T et al (2007) Computation of octanol−water partition coefficients by guiding an additive model with knowledge. J Chem Inf Model 47(6):2140–2148
25. Mansouri K et al (2018) OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform 10(1):10
26. Martel S et al (2013) Large, chemically diverse dataset of logP measurements for benchmarking studies. Eur J Pharm Sci 48(1–2):21–29
27. Daina A (2014) O Michielin, V Zoete, iLOGP: a simple, robust, and efficient description of *n*-octanol/water partition coefficient for drug design using the GB/SA approach. J Chem Inf Model 54(12):3284–3301
28. Fraaije JGEM et al (2016) Coarse-grained models for automated fragmentation and parametrization of molecular databases. J Chem Inf Model 56(12):2361–2377
29. Gedeck P (2017) S Skolnik, S Rodde, Developing collaborative QSAR models without sharing structures. J Chem Inf Model 57(8):1847–1858
30. Plante J (2018) S Werner, JPlogP: an improved logP predictor trained using predicted data. J Cheminform 10(1):61
31. Işık M et al (2019) Octanol-water partition coefficient measurements for the SAMPL6 Blind Prediction Challenge. J Comput Aided Mol Des. https://doi.org/10.1007/s10822-019-00271-3
32. Peltason L (2010) P Iyer, J Bajorath, Rationalizing three-dimensional activity landscapes and the influence of molecular representations on landscape topology and the formation of activity cliffs. J Chem Inf Model 50(6):1021–1033
33. Mannhold R, van de Waterbeemd H (2001) Substructure and whole molecule approaches for calculating log P J Comput Aided Mol Des 15(4), 337–354.
34. Zakharov AV et al (2019) Novel consensus architecture to improve performance of large-scale multitask deep learning QSAR models. J Chem Inf Model 59(11):4613–4624
35. Moriwaki H et al (2018) Mordred: a molecular descriptor calculator. J Cheminform 10(1):4
36. Cherkasov A et al (2014) QSAR modeling: where have you been? Where are you going to? J Med Chem 57(12):4977–5010
37. Wu Z et al (2018) MoleculeNet: a benchmark for molecular machine learning. Chem Sci 9(2):513–530
38. Tiño P et al (2004) Nonlinear prediction of quantitative structure−activity relationships. J Chem Inf Comput Sci 44(5):1647–1653
39. Olson RS, Moore JH (2019) TPOT: a tree-based pipeline optimization tool for automating machine learning. In: Hutter F, Kotthoff L, Vanschoren J (eds) Automated machine learning: methods, systems, challenges. Springer, Cham, pp 151–160

Journal : **Large 10822**	Article No : **279**	Pages : **12**	MS Code : **279**	Dispatch : **10-1-2020**